A Treatment of Stereochemistry in Computer Aided Organic Synthesis

Anthony Peter Fendick Cook

Submitted in accordance with the requirements for the degree of

Doctor of Philosophy

The University of Leeds

School of Chemistry

January, 2015

The candidate confirms that the work submitted is his/her own, except where work which has formed part of jointly-authored publications has been included. The contribution of the candidate and the other authors to this work has been explicitly indicated below. The candidate confirms that appropriate credit has been given within the thesis where reference has been made to the work of others.

Part of a review paper is reproduced in chapter 1 of this thesis. The reference is:

Computer-aided synthesis design: 40 years on

Anthony Cook[*], A. Peter Johnson, James Law, Mahdi Mirzazadeh, Orr Ravitz and Aniko Simon

Wiley Interdisciplinary Reviews: Computational Molecular Science Volume 2, Issue 1, pages 79-107, January/February 2012.

Article first published online: 11 MAY 2011   DOI: 10.1002/wcms.61

The paper is the candidates own work except for parts of the 'Automated Generation of Reaction Rules' section and the entirety of the 'Future Directions' section and the last 4 paragraphs of the 'Conclusions' section which were contributed by the co-authors. The co-authored portions are *omitted* from this thesis. The publisher has granted permission for the partial reproduction of the paper in this thesis.

Acknowledgements

I would like to thank my supervisors, Professor A.P. Johnson and Professor S. Marsden for their help, encouragement and critical guidance during the course of this research project. I would also like to thank my sponsors, SymBioSys Inc, for financial support, and Dr Aniko Simon and James Law for their assistance with practical aspects of the work.

I would also like to thank my wife Karina for her love, tolerance and support and her assistance in preparing this thesis. A special mention must go to my daughters Celyn and Scarlet for their patience while dad wrote a book.

This thesis is dedicated to the memory of my parents, Anne and Neville and my grandmother Joan who introduced me to the joys of chemistry when I was just 8 years old.

Abstract

This thesis describes the author’s contributions to a new stereochemical processing module constructed for the ARChem retrosynthesis program. The purpose of the module is to add the ability to perform enantioselective and diastereoselective retrosynthetic disconnections and generate appropriate precursor molecules. The module uses evidence based rules generated from a large database of literature reactions.

Chapter 1 provides an introduction and critical review of the published body of work for computer aided synthesis design. The role of computer perception of key structural features (rings, functions groups *etc*.) and the construction and use of reaction transforms for generating precursors is discussed. Emphasis is also given to the application of strategies in retrosynthetic analysis. The availability of large reaction databases has enabled a new generation of retrosynthesis design programs to be developed that use automatically generated transforms assembled from published reactions. A brief description of the transform generation method employed by ARChem is given.

Chapter 2 describes the algorithms devised by the author for handling the computer recognition and representation of the stereochemical features found in molecule and reaction scheme diagrams. The approach is generalised and uses flexible recognition patterns to transform information found in chemical diagrams into concise stereo descriptors for computer processing. An algorithm for efficiently comparing and classifying pairs of stereo descriptors is described. This algorithm is central for solving the stereochemical constraints in a variety of substructure matching problems addressed in chapter 3. The concise representation of reactions and transform rules as hyperstructure graphs is described.

Chapter 3 is concerned with the efficient and reliable detection of stereochemical symmetry in both molecules, reactions and rules. A novel symmetry perception algorithm, based on a constraints satisfaction problem (CSP) solver, is described. The use of a CSP solver to implement an isomorph-free matching algorithm for stereochemical substructure matching is detailed. The prime function of this algorithm is to seek out unique *retron* locations in target molecules and then to generate precursor molecules without duplications due to symmetry. Novel algorithms for classifying asymmetric, pseudo-asymmetric and symmetric stereocentres; *meso*, *centro*, and C_2 symmetric molecules; and the stereotopicity of trigonal (sp^2) centres are described.

Chapter 4 introduces and formalises the annotated structural language used to create both retrosynthetic rules and the patterns used for functional group recognition. A novel functional group recognition package is described along with its use to detect important electronic features such as electron-withdrawing or donating groups and leaving groups. The functional groups and electronic features are used as constraints in retron rules to improve transform relevance.

Chapter 5 details the approach taken to design detailed stereoselective and substrate controlled transforms from organised hierarchies of rules. The rules employ a rich set of constraints annotations that concisely describe the keying retrons. The application of the transforms for collating evidence based scoring parameters from published reaction examples is described. A survey of available reaction databases and the techniques for mining stereoselective reactions is demonstrated. A data mining tool was developed for finding the best reputable stereoselective reaction types for coding as transforms.

For various reasons it was not possible during the research period to fully integrate this work with the ARChem program. Instead, Chapter 6 introduces a novel one-step retrosynthesis module to test the developed transforms. The retrosynthesis algorithms use the organisation of the transform rule hierarchy to efficiently locate the best retron matches using all applicable stereoselective transforms. This module was tested using a small set of selected target molecules and the generated routes were ranked using a series of measured parameters including: stereocentre clearance and bond cleavage; example reputation; estimated stereoselectivity with reliability; and evidence of tolerated functional groups. In addition a method for detecting regioselectivity issues is presented.

This work presents a number of algorithms using common set and graph theory operations and notations. Appendix A lists the set theory symbols and meanings. Appendix B summarises and defines the common graph theory terminology used throughout this thesis.

Table of Contents

Chapter 1

Introduction

Computer Aided Synthesis Design

"*The intent … is not to replace art in organic synthesis but to show where real art lies*" J. Hendrickson.[1]

Introduction

It is now more than 45 years since the pioneering ideas concerning the *logic* of synthesis design were published in an essay by E. J. Corey.[2] Until then little thought had been paid to the development of reusable protocols that could be applied when designing a synthetic plan. Indeed the incentive to apply a logical analysis to the synthesis design problem was due to the coalescence of a number of important factors during the 1960s. First there had been a sustained period from the late 1940s onwards of successful total syntheses of complex natural products. Although an impressive accomplishment, these had not given a true insight into a *general* reusable approach to synthesis design. Each of these milestone syntheses (such as vitamin A,[3] cortisone,[4] morphine,[5] and chlorophyll [6]) were developed from a deep specialised understanding

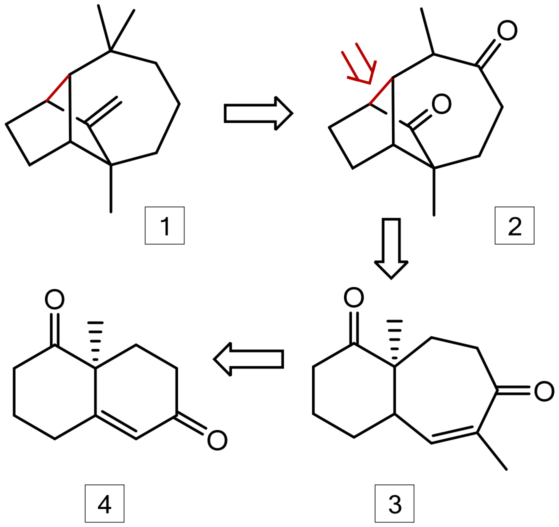


Figure 1 Corey's retrosynthetic plan for the synthesis of Longifolene.

of the chemistry of specific classes of compounds focusing on the forward synthesis from possible starting materials, an approach that limited the prospects of general design reuse. Secondly there had been the steady maturing of the disciplines of computer science and artificial intelligence coupled with the concurrent development of capable computer hardware and peripherals. These factors gave rise to the practical notion that a logical basis to synthetic analysis could be researched and demonstrated with a computer program.

Corey states that he had first applied a purely logical antithetic approach to the synthesis of the tricyclic sesquiterpene Longifolene **1** in 1957 (Figure 1).[7] Application of this novel synthesis design strategy lead to a number of promising pathways ultimately culminating in a successful total synthesis completed four years later.[8] The successful solution identified the need for an appropriate functionalised precursor **2**, from where the disconnection of a key strategic ring forming bond (arrowed) linked back to the bicyclic intermediate **3**. From here it was recognised that **3** could be formed via a ring expansion reaction from the readily available Wieland-Miesher ketone **4**. The novel and logical approach adopted was to work *backwards* from the target molecule, instead of attempting to progress forwards from presumed key intermediates or starting materials as had been the favoured strategy of the period. This approach was termed retrosynthetic analysis.

Corey's seminal 1967 essay[2] set out a number of axioms pertaining to synthetic analysis as a whole: the various elements involved in a synthetic solution are inseparable; a very large number of routes can usually be generated; the possible routes derive from the recognition of key structural units or "retrons"; there are criteria by which competing synthetic routes can be ranked; the individual steps of the synthesis are chosen from known chemical reactions, especially those with understood mechanism and scope. The paper also introduced the ideas of *systematic* simplification, where the repeated bond disconnections applied in a retrosynthetic direction generate progressively smaller precursors, thus working the problem *backwards* until recognised starting materials emerge. Indeed the vocabulary of *synthon*,[a] *transform*, *disconnection* and *retrosynthesis* introduced in this essay is now a standard part of undergraduate teaching of synthesis design.[9, 10, 11] In his essay, Corey also offered the first concrete algorithm representing a strategy for conducting a logical synthetic analysis of a target molecule:[2]

[a] In contrast to the original intention, the term "synthon" is now used by the chemistry community to mean "synthetic equivalent" and Corey has introduced the word "retron" which takes the original meaning of "synthon".

1. Repetition to generate alternative schemes
   2. Repetition for each intermediate (parallel synthetic sequences may exist)
      3. Systematic recognition of retrons
      4. Generation of equivalents and modified retrons
      5. Addition of control retrons
      6. Systematic disconnection of retrons
      7. Formulation of the possible synthetic transformations which reform the starting structure from the derived intermediate(s)
   8. Generation of intermediates until a starting point is reached
   9. Removal of inconsistencies
   10. Identification of unresolved problems
11. Assignment of merit

This protocol reveals some essential components of logical synthesis design. Steps 1 and 2 dictate that the problem should be solved by working backwards from the desired target molecule and that it should be iterative, with each generated intermediate molecule taking the place of the target molecule for the next step in the analysis. Alternative sequences should be generated (for comparison of merit) and convergence should be exploited (for efficiency). Step 3 alludes to the need for a computer to be able to internally represent a molecule and to perceive within this representation important chemical features such as: the presence, location and relationships of functional groups; functional group characteristics such as electron-donating or electron-withdrawing abilities; the presence of synthetically important rings; the relationships between groups of rings; ring characteristics such as size and aromatic character; recognition of stereogenic centres and their relationships to each other; and molecular symmetry. Steps 4 and 5 represent the concept of applying tactical sub-goals, either to exchange an existing functional group with a more appropriate one, or to introduce or remove a temporary moiety that provides activation or protection appropriate for a proposed goal reaction. Step 6 represents reaction selection as dictated by the perceived relationships between identified retrons, while step 7 identifies the concept of applying an operator equivalent to a *reaction* in the forward direction or a *transform* in the retrosynthetic direction. The selected transform lists the bond modifications that are needed to change a target molecule to a precursor molecule. Step 8 represents the ultimate goal of the plan, for example to terminate a retrosynthetic route when a readily available starting material is generated. Steps 9 and 10 recognise that the application of the transform may create an undesirable intermediate that is either unstable or even impossible and that this particular synthesis pathway should be abandoned. Finally step 11 allows us to compare all generated synthesis plans and rank them accordingly.

The first software prototype was the short lived but pivotal OCSS (Organic Chemical Simulation of Synthesis) program developed by Corey and Wipke.[12] This prototype pioneered a number of key developments: the use of interactive graphics to aid the input of target molecules and to review suggested intermediates; the representation of the analysis as a synthesis "tree" having the target molecule placed at the root node with AND/OR branches representing alternative or convergent pathways connecting to intermediate precursor molecules; the development of a perception module with algorithms to "see" the important features of the molecule such as functional groups, rings and molecular symmetry; the use of binary sets and set-theory to efficiently process the perceived structural features;[13] and the introduction of a strategy and control module to select appropriate synthetic goals. However the reaction transforms were hard coded as program subroutines ultimately limiting the scope and flexibility of OCSS.

A basic blueprint for logical synthesis design had been defined and prototyped by Corey and Wipke. Subsequently a number of competing research programs were initiated in the 1970s to explore this novel retrosynthetic concept and to discover and formalise powerful new strategies and heuristics that could aid chemists engaged in synthesis design. Some of Corey's axioms would be challenged by researchers in subsequent years,[14, 15, 16] especially the notion that the steps should be selected from known reactions or that the synthetic plan should be exclusively worked backwards.

Reviews

A number of reviews of this important field have been published. Todd[17] gives a good recent overview of the whole field including retrosynthetic and synthetic planning programs, reaction prediction programs (CAMEO[18]) and starting material selection programs (CHIRON,[19, 20] SST[21]). However the newest reaction prediction systems (ROBIA,[22, 23] NELB [24]) are not covered. Ihlenfelt and Gasteiger[25] offer a critical review of the early pioneering systems and highlight the difficulties of developing and maintaining manually produced transform databases as well as the competing emergence of reaction retrieval systems. Zefirov and Gordeeva[26] provide references and descriptions of a broad range of systems (both novel and derivative) from the early research period. Ugi *et al* [27] include a useful section on computer-aided synthesis planning in a general review of computer applications in chemistry. In addition specific detailed overviews exist for the LHASA,[28, 7] SECS,[29] SYNGEN,[30, 31, 32, 33] and SYNCHEM / SYNCHEM2 [34] synthesis planning programs.

Programs for Computer Aided Synthesis

A variety of computer aided synthesis planning methodologies has been developed by a number of computational chemistry groups over the last 45 years. After an initial surge of fundamental research spanning the first 20 years, interest eventually waned. Even though there were initial successes with the various developed methodologies, the systems did not enjoy widespread acceptance by users. Instead, user interest gradually refocused on using tools such as reaction retrieval systems and reaction databases to solve much simpler questions related to providing precedents for single step reactions.[25] Part of this lack of acceptance can be attributed to the chemist's belief that they could do better than any computer program. Ihlenfelt and Gasteiger go as far as to state that "*computer-assisted synthesis planning has been met with utter scepticism, even hostility, from the majority of chemists*".[25]

Part of the reason for the lack of acceptance can be attributed to weaknesses in the two main methodologies used to capture or express reaction knowledge. In the original empirical rule based systems retrosynthetic transforms were collated by hand and distilled into declarative computer readable descriptions incorporating the scope and limitations of each type of reaction along with the changes the reaction brings about. This was a laborious and highly skilled job with the result that the systems that employed this approach never acquired the breadth and depth of knowledge of a skilled human chemist. This problem was compounded by the concurrent and continued rapid discovery of new reactions, new methods and more selective reagents. There was little hope of catching up and the work of manually building the knowledge bases came to a virtual standstill.

The systems that chose to use formal methods to automatically generate reactions from first principles conversely suffered from the need to apply rigorous constraints to remove unlikely reactions and limit the inevitable combinatorial explosion of the synthesis tree. Even though these systems were capable of suggesting novel chemistry, the lack of hard precedents, with accompanying scope and limitations, was not attractive for the end user. This problem was partially addressed by linking in a reaction database to provide precedents for each suggested reaction.[35]

In recognition of these fundamental problems, research into the automated production of transforms from reaction databases has begun to make inroads over the last few years.[36,37,38,39,40] This newest approach promises to revitalise this important research area and is covered later in the introduction.

Topics

Perception of molecular features and the use of generalised chemical knowledge provide a vital entry into strategy and transform selection, as well as the means to recognise and filter out inappropriate precursor molecules generated when executing a synthesis plan. The recognition of certain structural features may strongly suggest a particular approach to solving a synthetic problem. *Strategies* provide the vital focus aimed at reducing the enormous combinatorial search space that is the main obstacle in any synthesis planning exercise. Strategies are used to develop the detailed plans that dictate what should be done and in what order, whereas retrosynthetic transforms or constrained reaction generators provide the tactical means to execute a plan. Elegance and economic efficiency imparted into a synthesis plan are a measure of a system's ability to select good strategies. The following sections highlight the scope and limitations of the techniques used to handle perception, tactics (via transforms, reaction rules or reaction generators) and strategy in a selected set of computer-aided synthesis research programs.

Target and precursor perception

Before a synthesis planning program can choose a retrosynthetic strategy or select an appropriate transform it must be aware of the important chemical features that make up the target molecule. Algorithms have consequently been devised for many aspects of molecular perception for a whole variety of applications including computer aided synthesis design. Other key applications performing molecular perception include molecule or reaction query and retrieval systems and structure elucidation programs. In the majority of these systems molecules are represented by some form of undirected graph (typically as a connection table). In graph theory terms vertices are used to represent atoms and the edges that connect together vertices represent the bonds. Consequently molecular perception algorithms are based mainly on graph theoretic principles.[41]

Functional groups

At the outset, the earliest programs OCSS, LHASA (Logic and Heuristics Applied to Synthesis Analysis)[28] and SECS (Simulation and Evaluation of Chemical Synthesis)[29] relied on transform selection by recognition of functional groups and the relationships between them. Two methods of functional group recognition have been described in the literature, one based on decision trees, the other on matching substructure patterns.

Corey described a decision tree approach[13] that initiates a search in the molecule connection table at selected atom types (principally primary hetero atoms) and walks through a series of tests comparing them to the atoms and bonds attached to the origin atom. If the walk successfully reaches a terminating branch in the tree, a functional group name is assigned to the matching substructure in the target molecule. Each atom in the matched substructure is assigned a name (ATOM*1 etc.) to allow it to be explicitly referenced during transform execution (*vide infra*). Each functional group is also assigned a descriptor that provides information about its reactivity and electron-withdrawing or donating effects. The construction of the decision tree was not straightforward and adding additional tests for new functional groups required detailed planning to locate the relevant insertion points in the tree.

Myatt describes an alternative generalised pattern based approach used in the CAESA[42] and ARChem[40] systems. This method utilised the PATRAN language (*vide infra*) to directly describe the substructure units that define each functional group. Adding new functional groups requires only a new independent pattern to be written and compiled. A subgraph isomorphism algorithm[41] was used to exhaustively match the compiled PATRAN patterns to the target molecule to locate all occurrences of functional groups.

Ring perception

Knowledge of the locations of synthetically important rings provides the means for synthetic analysis programs to devise strategies for ring construction from cyclic and acyclic precursors.[43] Sophisticated synthesis strategies can utilise the stereoselective differentiation of rings in fused polycyclic ring systems to plan the construction of specific arrangements of stereocentres on rings or in ring appendages formed later in the synthesis by cleavage of a ring.[44] Only a few cycle perception algorithms were known when the first retrosynthetic planning programs were developed. Consequently the cycle algorithms devised or modified by the Corey, Wipke and Gasteiger groups (amongst others of the period) are no longer considered the most efficient. Two basic approaches are available: a deconstructive method based on finding all rings and filtering these down to the synthetically useful rings; or a constructive method of finding a basis set of rings and selectively combining these to derive the synthetically useful set.

Corey described a number of ring perception methods[45,13] based on Paton's algorithm[46] which was extended and tailored to discover synthetically "important" rings. Based on a path growing algorithm devised to find the fundamental cycles, the modified algorithm selected only cycles that included all simple rings and envelope rings up to size 8. This rather arbitrary selection was based on the then known scope and limitations of ring forming reactions.

The logical selection of a ring set based on purely mathematical criteria such as choosing a minimum basis set (the smallest set of smallest rings or the SSSR), while ideal for structure naming, presented problems in the context of synthesis planning. For example the SSSR for structure A (Figure 2) includes only two six membered rings out of the three in evidence. The selection of these two rings is arbitrary (and wholly dependent on the numbering of the molecule graph vertices) so the addition of non-basis rings up to some "useful" limiting size provided a pragmatic solution. Similarly the norbornane skeleton (B in Figure 2) contains only two five membered basis rings and not the obviously synthetically important six membered ring attainable by application of the Diels-Alder reaction. These problems required the derivation of new classes of ring sets.

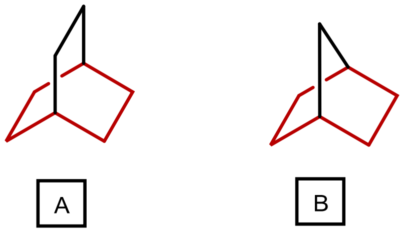


Figure 2 The highlighted rings are not selected by basis ring set algorithms.

Jorgensen has described a modified SSSR finding algorithm[47] that added in all symmetrically equivalent rings to the minimum basis set to solve some of the issues but this still did not directly address the norbornane problem. Plotkin[48] provided an alternative solution defining the $\mathcal{K}$ ring set[b] as the union of all possible minimum basis sets. This approach eliminated the arbitrary selection of a single basis set. To date the most efficient algorithm that finds the $\mathcal{K}$ set is Vismara's algorithm[49] that runs in a low order polynomial time.

To date the best overall performing SSSR algorithm is that due to Balducci and Pearlman which uses a breadth first search technique running in polynomial time.[50] Syslo[51] has improved the Gibbs linear combinations algorithm although the solution is limited to the planar graph set, meaning that certain uncommon classes of molecule are not amenable to analysis by this method.

Downs has described an algorithm for finding the extended set of smallest rings (ESSR)[52] used in a chemical patent search and retrieval application. The ESSR appears to be closely allied to the concept of synthetically important rings as it directly includes the problematic norbornane class

[b] The $\mathcal{K}$ ring set has the useful property that it is graph-invariant. Unlike a basis set its selection is not dependent on the numbering of the vertices of the molecular graph.

of six membered rings. However problems with the original ESSR definition have been identified. [53]

A general review of cycle perception algorithms drawn from the domains of mathematics, computer science and chemistry is provided by Downs *et al*[54] while Berger[53] provides a critical analysis for a wide sample of published algorithms that includes notes on algorithm performance and provides counter examples for the flawed methods.

Aromatic rings

The recognition of aromatic stabilisation in all types of carbocyclic and heterocyclic rings helps to determine if a retrosynthetic strategy should avoid breaking up the ring or if it should select appropriate reactions applicable to the chemistry of aromatic rings.

In the LHASA and SECS systems knowledge of the location of aromatic rings is used when keying the pattern based transforms (*vide infra*)[55, 56] that are concerned with aromatic substitution or the construction of hetero-aromatic ring systems. Jorgensen describes a detailed aromatic classification procedure used in the CAMEO program[47] based on the application of the Hückel (4n + 2) π electrons rule with a number of modifications. It ignores solutions where n > 5 recognising that large conjugated rings cease to exhibit aromatic behaviour and revert to regular polyene chemistry.[57] Care was taken to recognise tautomer forms that gave rise to aromatic stabilisation. An aromatic cycle detection procedure has been described for LHASA[13] also based on the Hückel method but with some limitations involving polarisable carbonyl bonds where, for example, cyclopropenone is only recognised as aromatic if drawn in its polarised form (Figure 3).

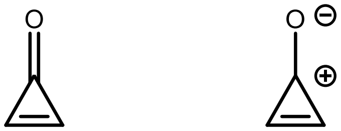


Figure 3 Problem cases in aromatic perception.

Later versions of LHASA and the CAESA program utilised an improved algorithm that abandoned the formal Hückel approach which tended to fail in complex ring systems, resulting in the overestimation of aromatic ring assignments. The revised approach was based on matching sets of generalised substructures written in the PATRAN language[58, 42] but was conversely open to underestimation of aromatic ring characterisation if a necessary pattern was missing. To overcome this problem, Seghal reported a hybrid solution[59] that utilised the formal Hückel approach for the first level of assignment and a second level application of a set of pattern rules to *eliminate* those ring systems overestimated by the first step.

Stereochemistry

Recognition of stereochemistry and the relationships between neighbouring stereocentres is vital to the appropriate selection of stereo-simplifying strategies and the application of stereoselective or stereospecific reaction transforms. Corey has described a large number of stereochemical based strategy rules[11] ranging from the general principle of identifying *clearable* stereocentres to the preservation of stereocentres selected from the chiral molecule pool.[60] For example elimination of stereocentres or modification of stereo relationships can be achieved by skeletal disconnective transforms or non-disconnective transformations converting one or more stereocentres to non-stereocentres (such as converting sp^3 to sp^2 centres). Stereo perception appropriate to synthetic analysis describes stereocentres in terms of ordered lists of substituents around each stereocentre or alternatively as local parity values related to the atom numbering used in the connection table.[61]

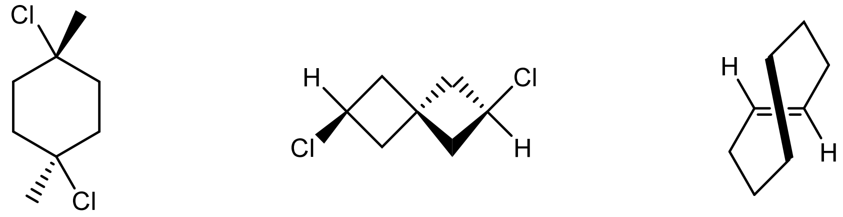


Figure 4 Axial and planar chirality in symmetric molecules.

Detailed methods for recognising stereocentres have not been widely published in the context of synthesis planning programs as many of those programs were not equipped to handle stereochemistry. Wipke and Dyott give a comprehensive description of the algorithms employed in the SECS program for the handling, representation and comparison of tetrahedral carbon and olefin stereocentres based on the ordered substituent list approach.[62] Corey and Long described LHASA's handling of stereochemistry[43,63] in a papers primarily concerned with multistep look-ahead and olefin synthesis strategies. Detailed algorithms for the handling of a comprehensive set of stereocentre types including sulphoxides, allenes, cumulenes and coordination complexes has been described by Blower *et al* in the context of the CAS Registry system.[64] A detailed algorithm for matching relative stereochemistry in a substructure search system has been described by Lipkus and Blower[68] and is readily applicable for solving the problem of identifying a stereo-specified retron in a target molecule. Günther and Zügge have described the advanced recognition and handling of axial and planar chirality in symmetric molecules[65] in the context of the Beilstein molecule registry system (Figure 4).

Molecular equivalence

The retrosynthetic analysis problem generally requires many alternative routes to be explored. A problem occurs if more than one route converges on the same precursor molecule as time is subsequently wasted duplicating an already explored route. Testing each newly generated precursor for equivalence against all others via a graph isomorphism algorithm can be computationally prohibitive. An effective solution to this problem employs the Morgan canonical naming algorithm[66] (or one of its derivatives[67]) to generate a sequence of characters that can be rapidly compared to others. Wipke and Dyott enhanced the method by incorporating stereochemical descriptors, defining the Stereochemical Extended Morgan Algorithm used to generate so-called SEMA codes.[68] The search times for establishing precursor uniqueness can be reduced from linear to constant lookup times by employing hashing techniques.[69] A useful review of the theory of the Morgan algorithm has been provided by Figueras.[70]

Molecular symmetry

The recognition of topological symmetry in the computer representation of molecules is essential to the selection of two-directional[71, 72, 73] and desymmetrisation synthesis strategies.[74, 75, 76] Knowledge of symmetry is used to prune out redundancy when matching retrons to molecules where either the retron or the molecule or both exhibit symmetry. A method that accounts for symmetry including stereo chemistry has been described by Agarwal for application in the SYNCHEM program. [77] A number of authors have described general methods based solely on the atom partitioning induced by the Morgan class of algorithms,[78, 79, 80] but this method can be flawed unless extreme care is taken. [81, 82, 83, 84] Ivanciuc in his review on canonical numbering and constitutional symmetry[61] points out that the origin of this problem is due to the Morgan algorithm failing to distinguish between topologically non-equivalent atoms in some cases. Figueras provides the mathematical basis of this failure.[70] More powerful algorithms have been devised to overcome this issue based on matrix multiplication techniques[85] or employing a highly constrained and efficient graph automorphism algorithm[c] in a post processing step after an initial application of the Morgan algorithm. [86]

Reaction representation

The essence of a synthetic plan is contained in the reactions that must be performed to execute the sequence of changes required to convert a starting material through to the target

[c] A graph automorphism is a mapping of a graph onto itself. The set of all automorphisms reveals the symmetries present in the graph.

compound. Systems that rely on the retrosynthetic approach code reactions in the reverse direction as *transforms*. To date four methods have been studied. The OCSS prototype coded a few reaction transforms directly in the implementation language of the program. This served the needs of a prototype but it was of course impractical, inefficient and rapidly abandoned. To enable flexibility and allow non-programmers to add new reaction transforms to the knowledge base, the LHASA and SECS programs adopted specialised declarative languages aimed to be friendly to non-programmer chemists. More recently, a new approach of automatically creating reaction transforms from databases of known reactions has been explored by programs such as ARChem.[40]

The EROS[87] and SYNGEN[31] programs took a different route, recognising the inherent difficulty of curating an effective database of known reaction rules. Instead these systems adopted what they regarded as a more fundamental approach. Using basic rules of feasible electron and bond redistribution they used algorithmic methods to generate the required reaction.

Reaction transformations

SECS used the ALCHEM[29] language to describe reaction transforms. ALCHEM was a purely declarative language and as such it did not support looping or subroutine calls. References to other named transforms were prohibited, ensuring each transform was self-contained and independent. Further ALCHEM deliberately did not support the specification of strategies to ensure strict independence between them and the tactics employed to execute them. The language supported the declaration of chemical facts and used situation-action statements to validate the applicability of the transform to a particular target. An example of an excerpted ALCHEM transform is shown in Listing 1.

```
; HA-C-C-W => A=C H-C-W
; Aldol condensation
   DGROUP WGROUP PATH 3 PRIORITY 100
   CHARACTER BREAKS CHAIN BREAKS RING
      IF GROUP 1 IS ESTERX THEN KILL
      IF GROUP 2 IS A NITRILE THEN SUBT 20
      IF AN RGROUP IS ALPHA TO ATOM 2 THEN SUBT 10 FOR EACH
      CONDITIONS BASIC AND NUCLEOPHILIC OR
      CONDITIONS ACIDIC AND NUCLEOPHILIC THEN SUBT 20
      ...
      BREAK BOND 1
      ...
      MAKE BOND FROM ATOM 1 IN GROUP 1 TO ATOM 2 IN GROUP 1
      ...
      END
```

Listing 1 A sample ALCHEM transform.

The general structure of an ALCHEM transform consists of a keying retron substructure, a priority rating, and sets of transform characteristics, reaction conditions, synthetic scope/limitations statements and structural manipulations. The transform character statement acted as a selection screen allowing the strategy executive to select only relevant transforms when attempting to satisfy a strategy goal (*vide infra*). The retron substructure described the minimum applicable structural unit that must be present in the target molecule to key the actions of the transform.

A number of retron representations were supported. Either the substructure could be specified as a named functional group, or a pair of functional groups separated by a path distance or via generic functional group names such as OXO (ketone, aldehyde or ester), DGROUP (alcohol, amine, thiol *etc*.) or WGROUP (OXO, cyano, nitro, acid halide etc.). An alternative retron form allowed an explicit pattern of atoms and bonds to be used to extend the ability of ALCHEM to key transforms with more complex non-functional group based substructures. The keying retron for the aldol reaction, for example, could use either “KETONE ALCOHOL PATH 3” or “O=C-C-C-OH”.

The scope and limitation of the transform were managed by situation-action statements (IF situation THEN action). These used set and graph theory operations [d] on perceived structural features in the target molecule to establish and validate specific locations in the molecule relative to the matched retron. If the situation was favourable the rating (PRIORITY) was incremented by a suitable amount. Conversely if the situation of the match was unfavourable the rating was decremented (SUBT 70) or the transform was abandoned (KILL). If the transform proved applicable after the scope and limitations analysis the manipulation statements were used to modify the structure to create a new precursor, which was then added to the synthesis tree.

The situation-action statements could also compare steric congestion at the reaction site via a supporting model builder[88] and electronic properties of conjugated systems via a Hückel molecular orbital calculation module.[29] These were expressed in situation-action statements such as:

“IF STERIC AT GROUP 1 BETTER THAN KETONE ANYWHERE THEN …” and “IF ELENERGY ON ATOM 1 BETTER THAN ATOM 2 …”.[e]

A special compiler program[29] was used to convert files of ALCHEM transforms into an efficient binary representation that was subsequently loaded into the SECS program at run-time.

[d] For example the intersection and union of sets of atoms or bonds and the generation of paths from a known starting point.

[e] ELENERGY refers to electrophilic localisation energy.

The LHASA program used the CHMTRN transform language to represent and evaluate the applicability of reaction transforms.[89, 90] CHMTRN constantly evolved during the long period that this program was used as a research tool. There are strong similarities with ALCHEM and some significant differences.

```
TRANSFORM 569
NAME GRIGNARD OPENING OF EPOXIDE
...PATH 2 BONDS
RATING 35
GROUP*1 MUST BE ALCOHOL
...
KILL IF DONATING GROUP ON CARBON*2  ...unstable epoxide
....
   ATTACH BROMIDE TO CARBON*3
   JOIN CARBON*2 AND HETERO1*1
   BREAK BOND*2
   INVERT AT CARBON*2
....
KILL IF CARBON*1 IS LESS*HINDERED THAN CARBON*2 ... undesired attack
```

Listing 2 A typical CHMTRN transform.

Listing 2 shows a typical (but incomplete) CHMTRN transform. CHMTRN supported qualifier (situation-action) tests for the generated precursor that could appear after the structure manipulation statements. This enabled inappropriate precursor structures to be detected and eliminated on mechanistic and steric grounds (Figure 5). For example the stabilisation of charges at various atoms in the generated precursor could be compared to determine if the regioselectivity implied was appropriate for the mechanism in question.

CHMTRN was extended to support looping and subroutine calls. For example: FOR EACH CARBON ALPHA TO CARBON*2 CALL CHCKFG.[f] These facilities were extensively used to mechanise decision tree based pattern matching in the long range strategy packages (*vide infra*). Further extensions were added to CHMTRN to enable the writing of strategy rules[55] in addition to its normal role as a reaction transform language.

[f] CHKFG is a function that checks for the presence of an interfering functional group.

A facility to utilise generalised substructure patterns in a transform retron description was supported with the PATRAN language (Listing 4).[55, 56] A linear PATRAN pattern was used to describe the required retron atoms and bonds as an alternative to the one-group or two-group keys. The PATRAN language is versatile and supported atom and bond qualifiers such as setting the location of supporting functional groups. During transform evaluation the pattern is matched to a target molecule and the matched atom locations are named[g] in the same manner as one-group and two-group keys so that CHMTRN situation-action statements could refer to the components of the retron pattern. This generalised approach was particularly useful for supporting a broad range of aromatic and heterocyclic reactions or any situation not adequately

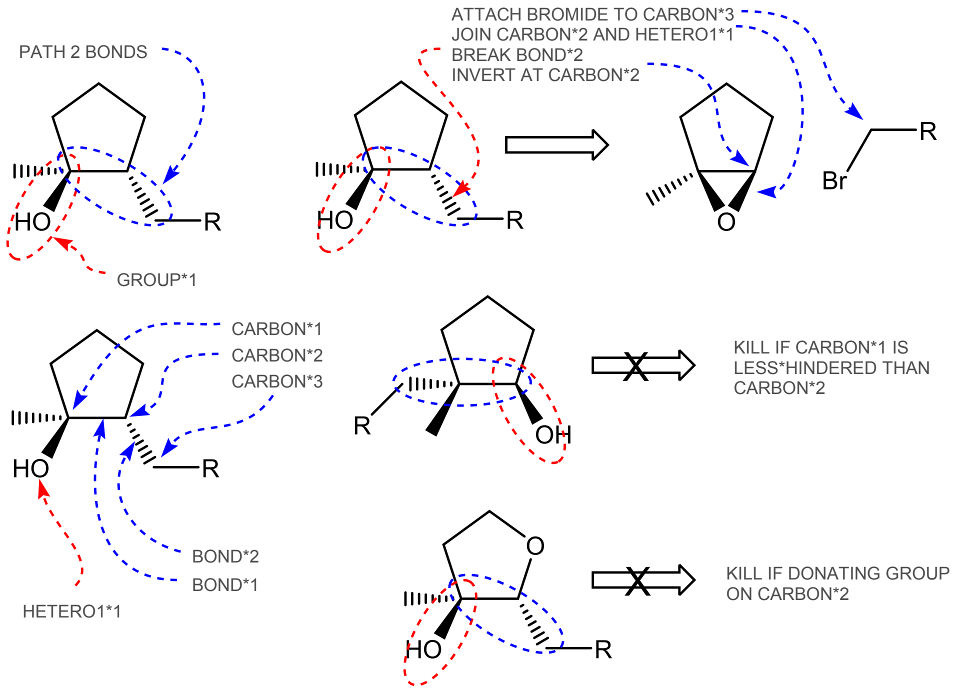
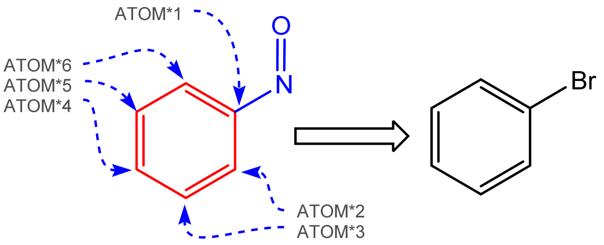


Figure 5 **The application of the epoxide ring opening transform (Listing 2) to various target molecules. The left-hand structures identify the relationship of named atoms and bonds with the functional group and path keys of the transform.**

handled by named functional groups. Efficient transform selection was supported by a filtering system[55] that indexed each transform by a set of substructure screens also written in the PATRAN language. This enabled the absence of structural features in the target molecule to be used to filter out non applicable transforms prior to transform evaluation.

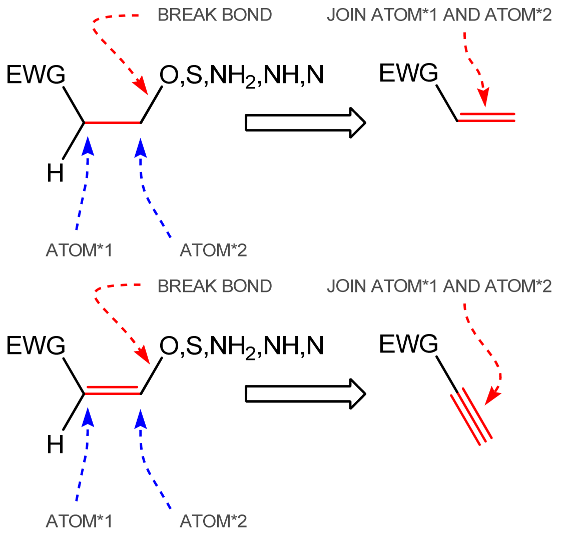
[g] Target molecule atoms that match a PATRAN statement are named ATOM*1, ATOM*2 etc. corresponding a left to right ordering of the pattern atoms.

```
TRANSFORM 1302
NAME AROMATIC NITROSATION
...STARTP
...C[FGS=NITROSO]%C%C%C%C%C%@1
...{1,6,5,4,3,2}
...ENDP
   REMOVE THE NITROSO GROUP ON
ATOM*1
   ATTACH A BROMIDE TO ATOM*1
```



Listing 4 An example PATRAN retron pattern, as used in LHASA aromatic substitution and heterocyclic ring forming reaction transforms. The righthand diagram is an interpretation of the transform retron and the action of the mechanistic commands.

```
...NAME MICHAEL ADITION OF HETERO
NUCLEOPHILE
...RATING 40
... W OR,NH2,NHR,NR2,S    W
... | |                   |
... C-C               => C=C
... |
... H
...
...
...
... RATING 40
... W NH2,NHR,NR2   W
... | |             |
... C=C          => C#C
... |
... H
```



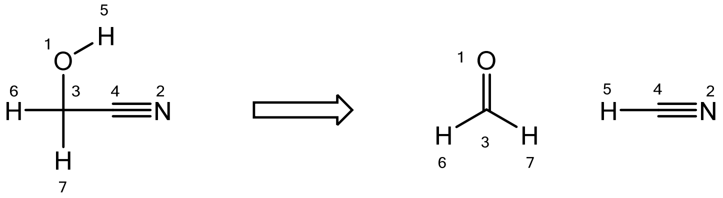
Listing 3 A LHASA 2D transform pattern. All retron keying and bond modification information is automatically extracted from the 2D transform pattern text. The righthand diagrams are interpretations of the transform retrons and the actions of the implied mechanistic commands.

Many LHASA transforms were rewritten using a newer highly generalised 2D pattern language incorporating both the retron and the corresponding precursor substructure[91] (Listing 3 illustrates a generalised 2D retron for the Michael addition). The addition of the precursor substructure enabled the transform to automatically determine the changes required to convert a target to a precursor obviating the need for structure manipulation statements. A “change code” based on the pattern of bond changes was computed for each transform and used as its index. These keys could be used to select appropriate reactions to satisfy a requested for a sub-goal. Sub-goals were invoked by transforms to rectify small mismatches between a detected substructure in the target molecule and the desired retron. The difference between retron and the located substructure was converted into a "change code" key and this was used to attempt to select a sub-goal transform to make the correction.

The change code mechanism was used to efficiently select the individual transforms needed to execute a particular tactical combination.[91, 92] The LHASA transform compiler was updated to compute change codes for each transform. Likewise the tactical combination compiler indexed the sequences of 2D reaction patterns in each tactical combination with a set of corresponding change codes enabling efficient cross referencing with the transforms. New transforms could then be accessed from a tactical combination rule without recompilation of the tactical combinations database.

Automated reaction generators

The application of formal approaches to algorithmic reaction generation provides an alternative to the use of transforms derived from reaction precedents. The Dugundji-Ugi (DU) model is exploited by programs such as EROS[15], IGOR[14] and RAIN[93] to automatically generate reactions using a formal reaction model. This consists of a molecule representation (the BE matrix) and reaction operators (the R matrix) that can transform one molecule into another. The BE matrix is an N x N atom matrix specifying the number of bonding electrons connecting atoms in the off-diagonal elements, while the diagonal elements specify the number of non-bonding valence electrons. The R matrix specifies the electron changes needed to bring about a reaction. Adding an R matrix to a BE matrix brings about a synthetic step, subtracting brings about a retrosynthetic step (Figure 6). A basis set of five R matrices can be used in linear combination to perform all theoretically possible reaction transformations bound by the usual rules of chemical valence etc. The advantage of this approach is that it affords the possibility to explore unprecedented reactions in a retrosynthetic plan. A disadvantage is the requirement of the



$$\begin{bmatrix} 4 & . & 1 & . & 1 & . & . \\ . & 2 & . & 3 & . & . & . \\ 1 & . & . & 1 & . & 1 & 1 \\ . & 3 & 1 & . & . & . & . \\ 1 & . & . & . & . & . & . \\ . & . & 1 & . & . & . & . \\ . & . & 1 & . & . & . & . \end{bmatrix}_{BE} - \begin{bmatrix} . & . & -1 & . & +1 & . & . \\ . & . & . & . & . & . & . \\ -1 & . & . & +1 & . & . & . \\ . & . & +1 & . & -1 & . & . \\ +1 & . & . & -1 & . & . & . \\ . & . & . & . & . & . & . \\ . & . & . & . & . & . & . \end{bmatrix}_{R} = \begin{bmatrix} 4 & . & 2 & . & . & . & . \\ . & 2 & . & 3 & . & . & . \\ 2 & . & . & . & . & 1 & 1 \\ . & 3 & . & . & . & . & . \\ . & . & . & . & . & . & . \\ . & . & 1 & . & . & . & . \\ . & . & 1 & . & . & . & . \end{bmatrix}_{BE}$$

Figure 6 BE and R matrices.

method to formally account for all atoms in the transformation which is problematic in the retrosynthetic direction (requiring the generation of otherwise unimportant by-products).[94] The main disadvantage is the inevitable combinatorial explosion of unrealistic reaction pathways.

EROS includes an armoury of heuristics[95] aimed at controlling this combinatorial problem, including algorithms that estimated reaction enthalpy values, to weed out unfavourable pathways.

SYNGEN employed a formal reaction generator based on a symbolic representation of reaction changes at carbon atoms.[96] Changing bond types were divided into the following classes: carbon-hydrogen or carbon-boron (***H***); saturated carbon-carbon (***R***); unsaturated carbon-carbon (***Π***); and carbon-heteroatom (***Z***) bond formation or cleavage. The number of bonds of each type at each reacting carbon atom was represented by the terms $\boldsymbol{h}$, σ, π, and $\boldsymbol{z}$ (which must sum to 4). The components $\pi + \boldsymbol{z}$ represented the carbon functionality (***f***) and $\boldsymbol{z} - \boldsymbol{h}$ the carbon oxidation state (***x***). The change occurring at each carbon was represented by a pair of half-reaction characters indicating the type of bond created and the type cleaved. In other words the first character of the transform code represents the nucleophilic component, the second the electrophilic component of the reaction change at each carbon atom.

This scheme allowed the coding of 16 different types of fundamental reaction changes which were classified into 3 reaction type categories. Oxidative through to reductive construction reactions were coded as ***RH***, ***RZ***, ***RΠ***. The introduction, elimination or alternations of functional groups (termed as trans-functionalization reactions) were coded as ***HH***, ***HZ***, ***ZH***, ***ZZ***, ***ΠH***, ***ΠZ***, ***HΠ***, ***ZΠ*** and ***ΠΠ***. The remaining category of 4 destructive reaction types was not used by SYNGEN. Linear sequences of these symbol pairs for each reacting carbon atom were used to represent complete reactions and could be used as either a retrosynthetic reaction transform or a synthetic transform.[16] For example the reaction shown in Figure 7 is coded by the symbolic reaction ***RΠ.HΠ*** (or the retrosynthetic symbols ***ΠR.ΠH***).

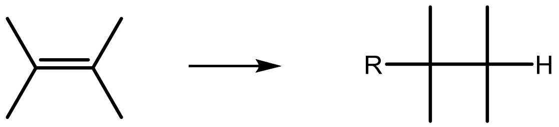


Figure 7 Addition to an olefin is coded as RΠ.HΠ

The sequence of changes in carbon functionality (***Δf)*** for each part of the reaction descriptor provided the means to determine relevance of a reaction type to the functionalised carbon skeleton scheme used in SYNGEN. The program RETRIEVE was used to map SYNGEN proposed generalised reaction pathways to literature precedents stored in reaction databases.[31]

Retrosynthetic strategies

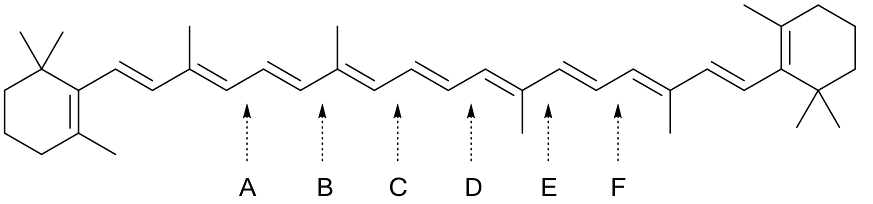
The original retrosynthetic protocol outlined by Corey (*vide supra*) represents a basic opportunistic (unconstrained) strategy. At each step every applicable reaction is tried, the only specified goal is to curtail a route when a known starting material is reached or an impossible structure is generated. As explicitly recognised by Corey, this strategy suffers from a potential combinatorial explosion of proposed routes, making it impractical unless the depth of the search is limited to a very small number of steps (typically no more than three or four steps). In particular, the functional group interchange (FGI) and addition (FGA) transforms are nearly always applicable if no constraints are applied. Consequently the aim of many synthetic planning programs is to apply goal directed strategies to significantly curtail the combinatorial growth by appropriate tree pruning, hopefully without discarding promising routes.

Strategies are represented by selecting and ordering specific goals related to the primary task of retrosynthetic structural simplification. Machine perception of the components of structural complexity in the target molecule provides the means to recognise and choose appropriate strategies. For example molecular complexity can be expressed in terms of: molecular size; ring and topological complexity; functional group content and density; stereocentre content and density; (hidden) symmetry; and regions of high chemical reactivity or instability. Each of these complexity features may be handled by independent or cooperative strategies, a general principle being the concurrent use of as many of these strategies as possible.[11] In the case of reactivity/instability an obvious strategy is to schedule the earliest removal of the reactive functional group, whereas ring and stereochemical complexity can be handled by strategies that aim at eliminating these features.

Strategies may not necessarily be related to structural complexity. A *transform-goal* strategy aims to utilise a specific high utility reaction and may need to *increase* structural complexity to generate the retron required to key the reaction transform. Conversely a *structure-goal* strategy identifies key intermediates or potential starting materials and guides the retrosynthetic analysis towards the goal molecule.

Strategy representation

The *SECS* program was designed to enforce a strong separation between program control, the transforms (*vide supra*) and strategy representation. Strategies are represented explicitly as a logic tree of individual goals such as functional group alterations, structural modifications (make or break bonds) and attention foci (use or avoid specified atoms, bonds or functional groups).[29] For example a symmetry based strategy applied to β-carotene (Figure 8) would identify appropriate strategic bonds (e.g. those constructible *via* a vinyl coupling reaction) and propose goals that symmetrically construct the molecule in a two-directional[h] manner.[73]



Goal 1: (AND (break bond A, break bond F)
Goal 2: (AND (break bond B, break bond E)
Goal 3: (AND (break bond C, break bond D)
Goal 4: (kill if fail, OR (Goal 1, Goal 2, Goal 3))

Figure 8 A goal plan for a two-directional symmetrical synthesis of β-carotene.

In this scheme strategies and goals are independent of the transform library and the program must search for appropriate transforms by a two-step matching process. First the transform must be *relevant* to the goal, and, second, it must be *applicable* to the target. Relevance is used as a screening step that filters potentially applicable transforms by comparing the characteristics of the goal to those of the transform and serves to limit which transforms are inspected. Characteristics include: breaks chain; expands or contracts ring; makes bond; alters group, amongst others. Applicability is determined by a detailed matching process where the actions of the transform on the target are compared to the objectives of the goal, resulting in either: rejection of the transform; application of the transform; or if the transform appears strongly relevant but not directly applicable, a separate sub-goal tree aimed at achieving applicability is generated. The difference between the target structure and the transform substructure is used

[h] A two-directional strategy involves the simultaneous homologation of both ends of a chain, with a desymmetrisation of the ends if required.

to generate specific sub-goals and is an example of problem solving by decomposition guided by means-end analysis. [i, 97]

It was found that this approach limited the application of otherwise frequently triggered FGI or FGA transforms to just those useful occasions that a sub-goal was required to make a non FGI transform applicable. Notably sub-goals were not invoked directly by transforms thus maintaining a strong separation between strategy and tactics.

The strategies known to SECS are coded as program subroutines and focus mainly on skeletal topology.[98] The user interface is highly interactive and allows the chemist to draw a target molecule, select an appropriate strategy, review, edit or modify the generated goals and review the resultant precursors before repeating the process with the precursor that the chemist judges as best. The chemist user is thus an integral part of strategy control.

Strategy selection

A wide variety of different planning strategies and strategy control structures were investigated in the development of the *LHASA* program. Early versions of the program were exclusively interactive and the chemist user was integral to the strategy selection process. After drawing the target molecule[99] the user was required to select an appropriate strategy from a menu of choices. The program then applied this strategy, attempting to produce a set of precursor molecules ranked by merit. The user then selected one of these precursors along with an appropriate strategy and the process was repeated until the chemist considered that a simple or known molecule had been reached.

Short range opportunistic strategies

The earliest implementation of LHASA employed a short-range search strategy.[89] This was an opportunistic approach based on recognising the presence of a specific functional group, or a pair of functional groups separated by a specified bond path distance. This strategy was associated with a library of 1-group and 2-group goal transforms focused on a wide variety of carbon-carbon bond forming reactions. Its utility was significantly enhanced by enabling the invocation of a single step FGI or two parallel FGI sub-goal transforms aimed at correcting any goal transform mismatches. If sub-goals were required, the FGI transforms were selected from a

[i] MEA is a strategy to control a search in problem solving. Give a current state and a goal state, an operator is chosen to reduce the difference between the states. A possible difference measure is an estimate of the number of remaining reaction steps or an estimate made via a molecule complexity function.

precompiled functional group cross reference table to select an appropriate correcting transform. This search strategy relied on the inherent propensity of goal transforms to simplify the target molecule and at the end of the analysis offered the chemist a comprehensive set of candidate precursors reached by one to three retrosynthetic steps. Its immediate limitation was the lack of search depth as well the limited type of chemistry performed, based only on functional groups. However the chemist could pick promising precursors and repeat the process to increase the depth of the analysis.

Goal directed strategies

A number of goal directed strategies were investigated within the LHASA framework to increase the scope of its synthetic problem solving abilities and to improve control of the search within the synthesis planning space. These advanced strategies were developed to reduce the combinatorial growth of alternative plans and in general allowed the program to avoid exploring too many weak routes.

Sub-goal tactics

The ability to efficiently find and apply longer sub-goal FGI sequences was further developed in LHASA as an extension of the approach used in the short range strategy module.[100] The application of these sequences was goal driven, aiming to set up a precursor molecule that incorporated the desired retron to key an important highly simplifying or high utility goal transform. This facility was exploited as a high performance transform subroutine offering the transform writer the ability to request a correction from one specific structural unit to another. However the usefulness of this sub-goal tactic was limited to supporting only the highest utility goal transforms such as the Aldol, Michael, Wittig, Grignard and Mannich reactions and the Claisen and Dieckmann condensations amongst a few others. It was found that general application of this supporting strategy resulted in too many questionable routes, whereas restricting its use to the above qualifying reactions generally worked to improve the generated route quality.

This sub-goal module was made available to the various strategy packages implemented in LHASA. Its ability to apply sub-goal transforms to correct small mismatches between substructure units in the target molecule and a goal transform retron proved to be immensely powerful. Sub-goal rectification was initially limited to applying functional group based transforms such as FGIs and FGAs and could not be invoked for general goal transforms. A generalised solution was implemented that allowed *any* transform category to qualify as a sub-goal.[91] This required a change to the transform keying mechanism in which a unified 2D pattern language (*vide supra*) was used to replace the use of functional group based retrons.

Topological disconnection strategies

A topological bond disconnection strategy was developed in LHASA requiring the identification of *strategic bonds*, which if retrosynthetically cleaved, would yield a greatly simplified precursor structure.[43] For example, disconnecting qualifying strategic bonds found in *spiro*, simple fused or bridged polycyclic compounds reduced the number of rings while minimising the number of generated ring appendages. It was anticipated that this simple approach would direct the search towards synthetically accessible precursors. A number of generalised rules applicable to carbocyclic and heterocyclic targets were developed to locate appropriate ring disconnection bonds based on general chemical as well as topological principles. The set of perceived strategic bonds was then used as the specific search goal. For each strategic bond in the set, a search was made for transforms that could be directly applied to cleave it. If the bond could not be directly cleaved, another search was made to apply either FGI or FGA sub-goal transforms in an attempt to set up a regular 1-group or 2-group goal transform on the strategic bond. FGA sub-goals were applied to insert functional groups an appropriate distance from the strategic bond.

The key operational features of a topological strategy are illustrated in Figure 9. In this example no goal transforms could disconnect the perceived strategic bond (in bold), so up to four sequential FGA or FGI sub-goals are applied exhaustively. Those that could key regular goal transforms are then pursued. Failed sub-goal routes are discarded without presentation to the user.

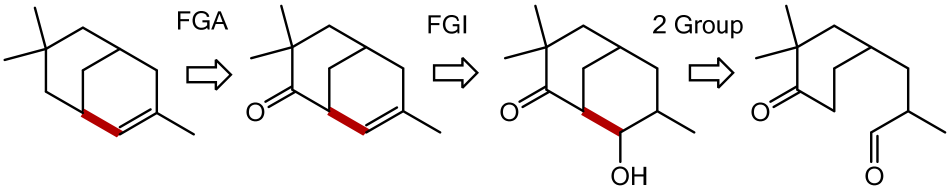


Figure 9 A strategic bond directed strategy.

The approach proved successful in many situations but had limitations. The strategic bond selection rules were not applicable to dense polycyclic fused rings systems. Routes requiring the simultaneous disconnection of two bonds (such as a Diels-Alder disconnection) were beyond the scope of this strategy. Another identified weakness related to the lack of a supporting sub-strategy that could exploit ring rearrangements prior to evaluating strategic bond disconnections.

Specific strategic bond perception rules were designed to avoid disconnections that created appendages carrying stereocentres, but this necessary tactic often curtailed what would be promising routes. This limitation was necessary because at the time no sub-goal strategy was available to direct the prior elimination of any problematic stereochemistry. This problem was identified as an area of further research.

The need to use an exhaustive application of sub-goal sequences and the consequent combinatorial growth, limited the practical depth of this goal directed strategy. Goal representation was simplistic as it considered a set of bonds from which only the cleavage of one bond was required to satisfy the goal. The successive cleavage of many strategic bonds could be achieved by reapplying this strategy to each generated precursor but this had to be guided by the user. The program was unable to automatically forward plan to evaluate a longer bond disconnection sequence.

Appendage disconnection strategies

The topological strategy was extended to cover ring appendage and branch appendage disconnection bonds.[44] A set of rules was devised to define the appendage concept and a perception module was implemented to locate qualifying bonds. The augmented strategic bond set was then processed by the existing topological strategy executive.

Using the topological strategy (Figure 10), LHASA was able to locate strategic ring appendage bonds and direct the search towards reducing the number of rings and subsequently cleaving off the generated appendage.

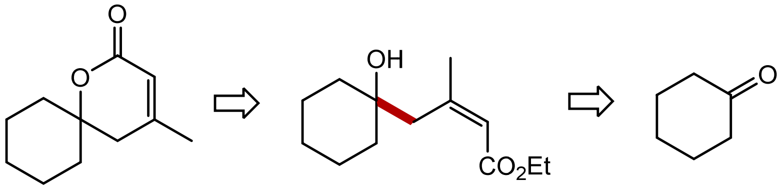


Figure 10 Disconnecting ring appendages.

Appendage reconnection strategies

A bond reconnection strategy package was implemented in LHASA to support synthetic problems requiring control of stereochemistry on rings and ring appendages (Figure 11). The reconnection strategy supported designing syntheses of medium sized rings[44] via transannular bond reconnections (Figure 12). An important principle behind the reconnection strategy was that the formation of chiral centres in a ring is generally easier to control than on a chain (Figure 13), though modern synthetic methods employing chiral reagents, auxiliaries or catalysts have opened up new avenues to solving this class of problem.[101, 102, 103]

In contrast to most other strategies the reconnection strategy increases precursor complexity as it generates new rings in the retrosynthetic direction. However the strategy strives to offset this cost by gaining access to substrate driven stereochemical control or access to otherwise hard to make medium sized rings. The strategy executive considered three classes of bond reconnection: ring appendage to ring appendage; ring appendage to ring; and non-ring reconnections that retrosynthetically form monocyclic precursors. A procedural plan systematically considered all qualifying pairwise reconnection points. Detailed analysis of the bond paths between the reconnection points was made to eliminate difficulties due to the generated ring size and ring appendage stereochemistry.

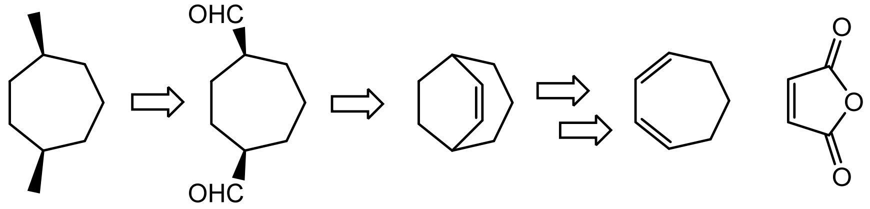


Figure 11 Ring appendage reconnection strategy.

Extensive use of sub-goal invocation was employed to attempt appropriate functionalization of

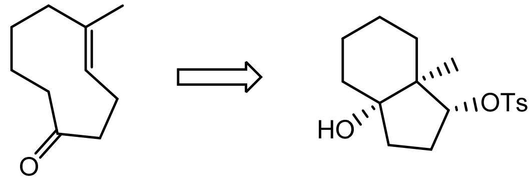


Figure 12 Medium ring retrosynthesis via a transannular bond reconnection.

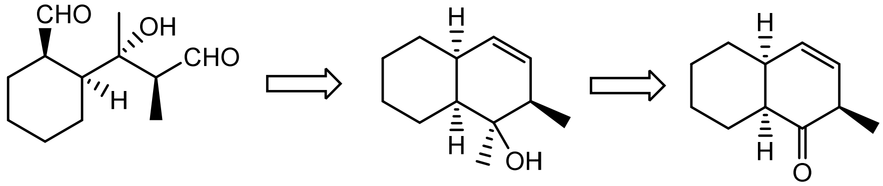


Figure 13 Controlling appendage stereochemistry via a ring reconnective strategy.

the appendage or ring reconnection points. An epimerisation sub-goal was introduced to attempt to overcome adverse stereochemical situations. Heuristics were used to determine if sub-goals were likely to be productive and, if not, to rapidly abandon that proposed reconnection and try another. The medium ring strategy also used general heuristics to avoid the generation of bridged precursors by rejecting reconnections that did not directly join two atoms across the same ring. A bias was introduced to favour fused precursors with five or six membered rings by ensuring nine membered rings or higher used reconnection paths of at least three bonds, but ideally four or five.

Long range goal strategies

Methods for implementing long range goal directed strategies were investigated.[104] The opportunistic approach was limited due to the tendency to search the synthetic space in a breadth first manner leading to a combinatorial growth of proposed routes before any depth was achieved. The alternative approach was to employ directed deep searching. In order for the depth first approach to be productive it had to be guided towards a specific goal. The chosen goal was to find a route to a powerful simplifying transformation (known as a transform goal or T-goal). T-goals represented reactions that significantly reduced molecular complexity by ring disconnection and/or the reduction of or removal of stereocentres. The user was able to choose one of these long range strategies if they thought it was applicable to the current precursor. Long range search packages were implemented for the Diels-Alder reaction,[104] the Quinone-Diels-Alder reaction,[105] halo-lactonization,[106] Robinson annulation,[107] and stereoselective olefin synthesis.[108] The implementation approach adopted used highly detailed question and answer binary decision trees written in the CHMTRN transform language (*vide supra*).

Using detailed knowledge of routes to key precursors of the T-goal retron, the transform writer selects common intermediate structural patterns (known as structure goals or S-goals) that are readily converted to the T-goal substructure by a known transformation. The decision tree is arranged to attempt to convert the current precursor into one of the selected S-goal substructures and as such consists of detailed matching procedures for each S-goal. The aim of these procedures is to use FGI or FGA sub-goals to make progressive non-simplifying corrections to drive the search towards the key S-goal. These procedures are keyed by recognising small ‘localised matching units’[106] consisting of 1 to 3 functionalised carbons and using an associated specialised subroutine to attempt to solve the sub goal. To reduce the search space, prior procedure evaluation[106] is used to estimate the cost of reaching the S-goal from the current position and curtail the sub-goal route if the estimate is too high (for example if too many steps were predicted).

These long range strategy modules proved to be highly effective and able to propose effective routes of up to 15 steps. However the procedural decision trees were difficult to write and maintain and the approach proved too prohibitive in terms of keeping up to date with new chemical knowledge. Another limitation was that multiple strategies could not be executed simultaneously due to the strict procedural execution of each long range strategy module.

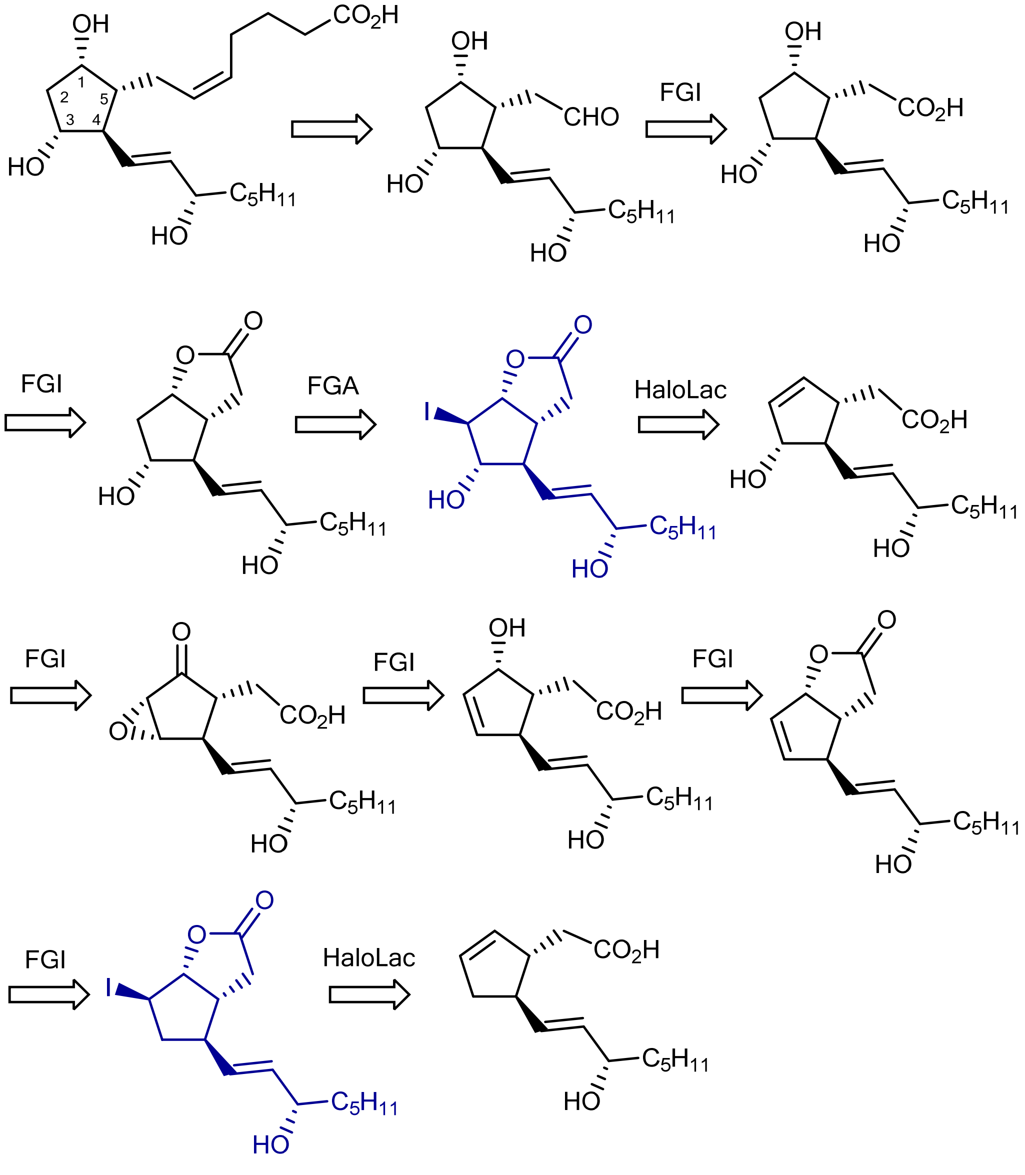


Figure 14 The LHASA long range halolactonization strategy applied to planning the synthesis of PGF2α

Figure 14 illustrates the capabilities of LHASA long range strategies with the application of the user selected halolactonization T-goal strategy to the retrosynthetic analysis of the prostaglandin PGF2α. This particular route was the fourth highest scored route merited by LHASA. The notable features in this example are: the sequential application of FGI and FGA sub-goal transforms to achieve the targeted halolactonization retron; the multiple use of the stereospecific halolactonization reaction to relay the established stereochemistry of the 5 position to the hydroxyl groups introduced at the 1 and 3 position. Underlying the use of the halolactonization T-goal is the important strategy of stereochemical simplification by the controlled systematic removal of stereocentres.

Convergent synthesis strategies

Hendrickson's *SYNGEN* program[31] employed an unsupervised hierarchical strategy planning algorithm producing first a set of generalised construction plans and then a more detailed set of functionalization plans before searching for matching transformations. The first phase of the strategy reduced the target molecule to a carbon skeleton and analysed the resultant reduced graph to determine which bonds should be strategically disconnected. The end point of the disconnections was reached either when each disconnected skeleton matched a skeleton of a known starting material or when two levels of pairwise skeletal divisions had been completed. The disconnection bonds were represented as a bond set and ordering this set as a hierarchy produced a skeletal construction plan. Selecting different bond sets and hierarchical orderings produced alternative construction plans.

This abstraction process had the advantage that it conflated groups of retrosynthetic pathways which shared common carbon skeletons and key carbon-carbon forming reactions into a single common construction plan, allowing the smaller set of generalised construction plans to be compared and ranked. The number of possible construction plans was still combinatorial even for moderately sized target molecules (the 18 carbon skeleton of estrone with up to 9 selected construction bonds generates around 107 billion unique construction plans) so it was necessary to employ a selection algorithm that chose the most economical plans first. The search for a construction plan was directed by assigning a weight to each skeletal fragment based on its size and then scoring the whole plan by combining those weights with the number of steps required to carry that fragment to the target. Consequently the best plans selected are those which are fully convergent, bringing together balanced fragments in the least number of steps.

The second phase functionalised the carbon framework of the construction plan, producing a set of alternative functionalization plans by simple permutation. The functional groups were generalised as a leaving group, electron-withdrawing group, electron-donating group *etc.*, and their codes were marked on the carbon skeletons in preparation for selecting suitable transforms created by a reaction generator (*vide supra*). Figure 15 illustrates the actions of the SYNGEN strategy controller showing the highest merited construction plan, a derived functionalization plan and matched starting materials for the synthesis of an estrone precursor.

This approach to synthesis design is claimed to have the following advantages: it chooses the most efficient fully convergent synthesis;[109] the contraction of the search space via the use of construction plans makes the search of the synthesis space more manageable; and it focuses on convergence to known starting materials avoiding a blind search. However the approach is unable to capitalise on advantageous rearrangement or degradative reactions and has no provisions for stereochemical control.

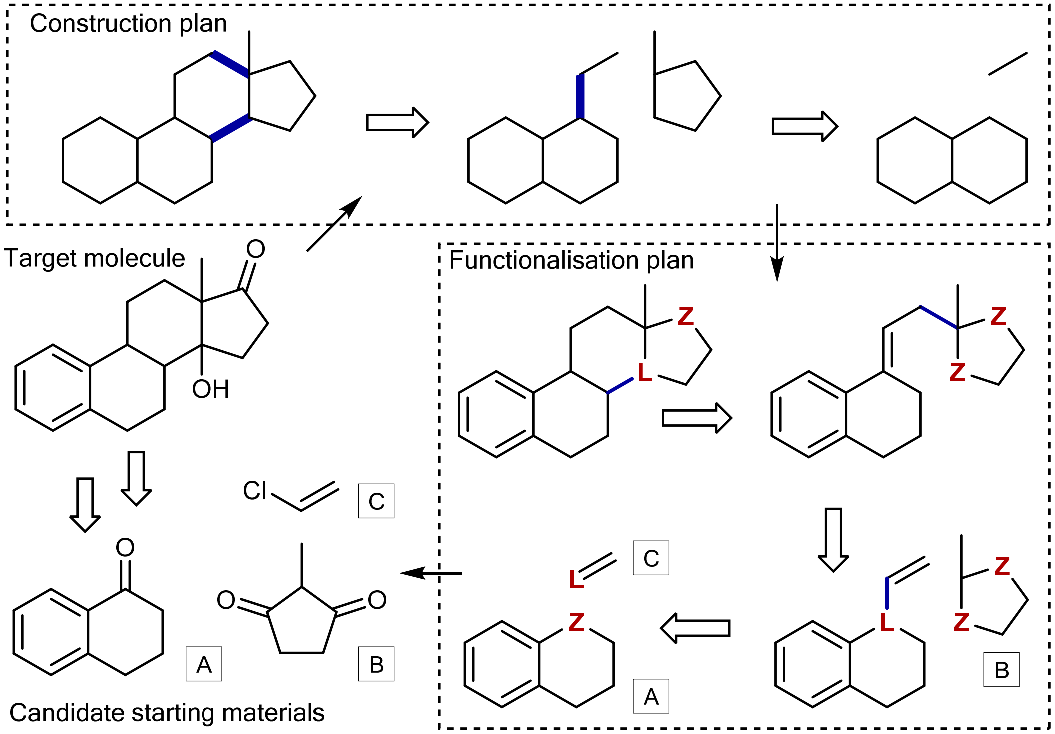


Figure 15 Construction and functionalisation plans for an estrone precursor with matched starting materials.

Starting material directed strategies

A powerful associative strategy[12] module was added to LHASA to direct a synthesis plan towards known starting materials.[110] The user drew in a specific starting material or selected one from a pool of aromatic, chiral or isotopic-labelled compounds obtained from supplier catalogues. A heuristically guided mapping algorithm[111] was employed to discover the possible embedment of the selected starting materials onto the target molecule. A scoring function[112] based on measuring "synthetic distance" was used to select the best mappings from which the user chooses one. The notion of synthetic distance was based on estimating the difficulty of performing the required modifications from the application of general chemical knowledge. Once a mapping was chosen, the retrosynthetic goals derived from the target-to-starting material mapping were expressed in terms of required functional group and bond modifications in a manner similar to the SECS goal list (*vide supra*).

Ordering goal directed strategies

The order of goal satisfaction is generally vital to the success of a synthetic plan and without ordering the constraints, N goals generate N! plans. Consequently to reduce the combinatorial growth an unsupervised planning algorithm[55] was added to LHASA. This used strategy rules written in a declarative chemist friendly language to apply ordering constraints to the unordered goal satisfaction plan generated by the application of the starting material directed strategy module.

Three problems were addressed: which precursor molecule to pursue next; which goal to satisfy next; and which transforms to select to attempt to satisfy the selected goal. Selection of a precursor was performed using the A* search algorithm[113] which relies on a two part cost estimation function,[114, 115] the estimated cost to reach the current point in the plan, and an admissible estimate[116] of the cost to continue to the targeted starting material. The backward cost included the number of transforms (reaction steps) used, coupled to an estimate of reaction yield estimated from the evaluated transform ratings. The forward cost incorporated an inertia value coupled with a measure of the remaining synthetic distance using the same function that rated the target-to-starting material mapping. The inertia constraint was important to stop the search losing focus and prematurely jumping back and forth between precursors. A change of route was thus delayed until one route was convincingly more promising than any other. This approach combined both breadth-first and depth-first searching so was able to search for precursor routes in a "best first" manner without the need for user intervention.

A goal solver was used to test strategy rules against the current precursor and determine the next best goal to attempt.[55] Each strategy rule was invoked only if an associated keying structural pattern could be matched to the precursor and the matched location corresponded to a specified goal type stated in the rule. After evaluating the qualifying statements in each strategy, the conclusions about the ability to satisfy each goal were combined to decide on an overall outcome. As the qualifying strategy rules may reinforce or contradict each other with respect to a specific goal the final outcome was determined by a combination function that weighed the arguments presented by each rule.

Strategy rules could also amend existing goals or add new ones before scheduling the assigned goals. Each strategy rule could also specify the category of transform used to attempt goal satisfaction thus speeding up the selection process. Control was then passed to the transform executive to search for appropriate goal or sub-goal transforms which were used to generate new precursors.

The capabilities of the goal solver were demonstrated with a package of aromatic substitution strategy rules (Listing 5).[55] These strategy rules enabled LHASA to design efficient synthetic plans that could capitalise on the directing effects of ring substituents and attempt to avoid premature deactivation of reactivity. This was supported by the addition of a semi-quantitative Hammett-type calculation[55] that provided the ability to compute approximated net directing effects in poly-substituted aromatic rings.

IF the substituents on the ring direct to a site THEN make substitution at that site a higher priority than others
IF a ring is electron rich THEN favour electrophilic substitution transforms
IF a ring is electron poor THEN favour nucleophilic substitution transforms
IF an FGI on a substituent alters the directing affect to another substituent THEN schedule the FGI
IF a substituent is electron-withdrawing THEN make its removal a high priority
IF a substituent Is interfering THEN make its removal a high priority
IF the removal of a substituent yields an attainable substitution pattern THEN make its removal a high priority

Listing 5 Aromatic substitution strategy rules implemented in LHASA.

Listing 6 illustrates a LHASA strategy rule representing the statement “IF a substituent is electron-withdrawing THEN make its removal a high priority”. It consists of: a keying substructure pattern that locates the ring substituent; an initial strategy rating to establish relative rule importance; a selector for the transform category to be used to attempt goal satisfaction (electrophilic substitution transforms qualify in this example); a qualifier that

```
NAME ELECTROPHILIC AROMATIC SUBSTITUTION STRATEGIC ORDERING
...STARTP
...C[HS=0]%C%C%C%C%C%@1
...{1,6,5,4,3,2}
...ENDP
RATING 20
ELECTROPHLIC*AROMATIC*SUBSTITUTION
KILL IF NOT GOAL*BOND ON ATOM*1
... Remove withdrawing group early to make other electrophilic substitutions easier
... and prefer groups that are likely not to require an FGI i.e. withdrawing groups
ADD 5 IF THERE IS A WITHDRAWING*GROUP ON ATOM*1
SUBTRACT 5 IF THERE IS A DONATING*GROUP ON ATOM*1
SCHEDULE THE STRUCTURAL CHANGE REQUIRED ON ATOM*1
```

Listing 6 A LHASA strategy rule written in the CHMTRN language.

establishes whether the pattern was matched to a required goal type (the strategy rule is irrelevant if the substituent bond is not a goal bond); qualifiers that adjust the strategy rating based on substituent electronic properties (in this case favour electron-withdrawing groups); and an instruction to schedule the qualifying goal with its final rating.

Using the goal solver with the aromatic substitution strategies, LHASA was able to plan a retrosynthetic route from bumetanide to the user selected starting material, toluene (Figure 16).

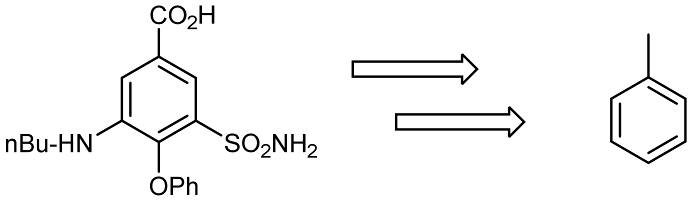


Figure 16 The synthesis of bumetanide from toluene requires four FGRs.

The full route developed by LHASA (Figure 17), corresponded to the commercial manufacturing route. LHASA was able to make each goal sequencing decision correctly by the cooperative application of the aromatic substitution strategy rules. This route was the first one found during the analysis. A notable achievement was the ability of the strategy rules to suggest the use of FGI sub-goals to alter substituent activation and directing effects appropriate to the synthesis. Strategy rules could achieve this by amending existing goals or they could insert new ones to request the needed sub-goal.

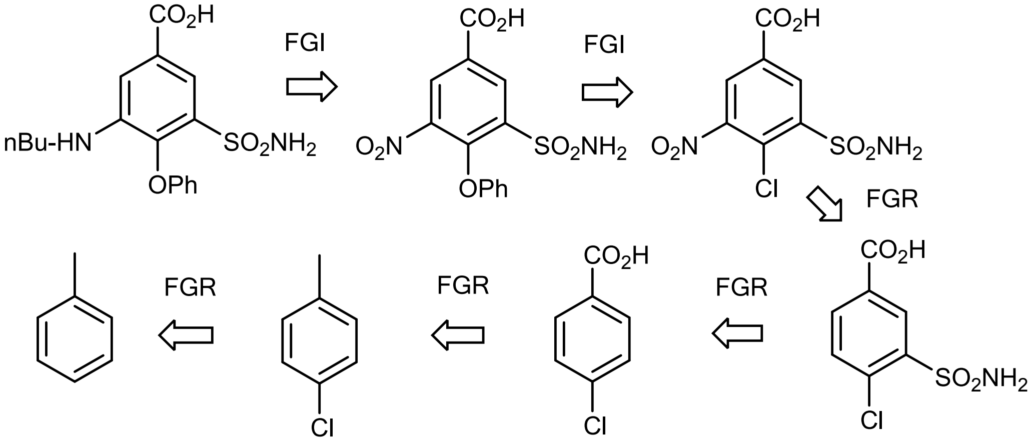


Figure 17 Retosynthetic analysis of bumetanide proposed by LHASA.

Tactical combinations of transforms

A "tactical combinations" strategy[117] in LHASA was built upon the generalised sub-goal scheme.

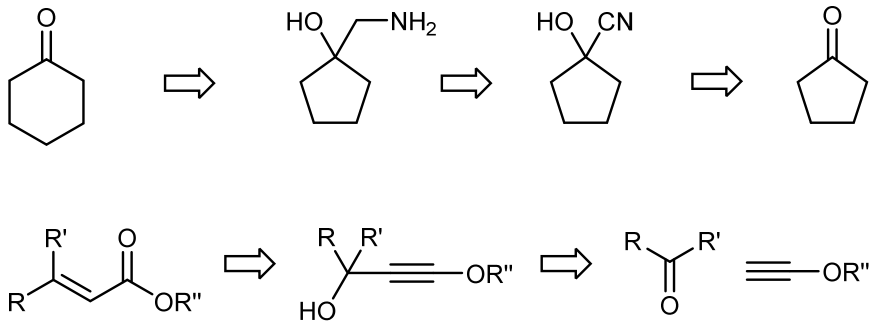


Figure 18 Example tactical combinations of transforms.

A tactical combination (TC) represented what may be termed a "synthetic cliché", a commonly used sequence of reactions repeatedly used in a range of published syntheses to achieve a particular objective such as alkylating a specific ring position, executing a functional group transposition or performing a ring expansion or contraction (Figure 18).

Over 450 TCs were added to the LHASA knowledge base and proved useful for both

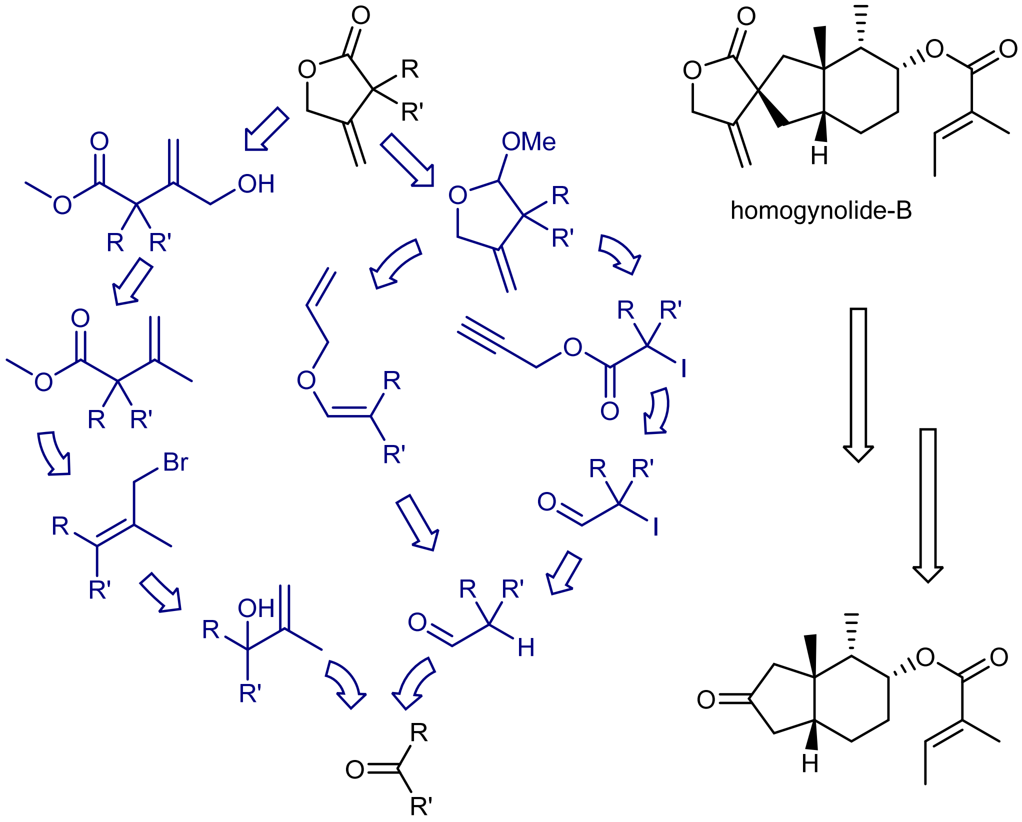


Figure 19 The application of a tactical combination of reactions to the retrosynthetic analysis of homogynolide-B.

opportunistic and tactically-guided analysis strategies.[92] The TC facility allowed LHASA to propose retrosynthetic plans comparable to the published synthesis of homogynolide-B[118, 119] (Figure 19 and Figure 20). In this particular retrosynthetic analysis LHASA was able to recognise the application of two successive tactical combination sequences each comprising a number of alternative prebuilt routes.

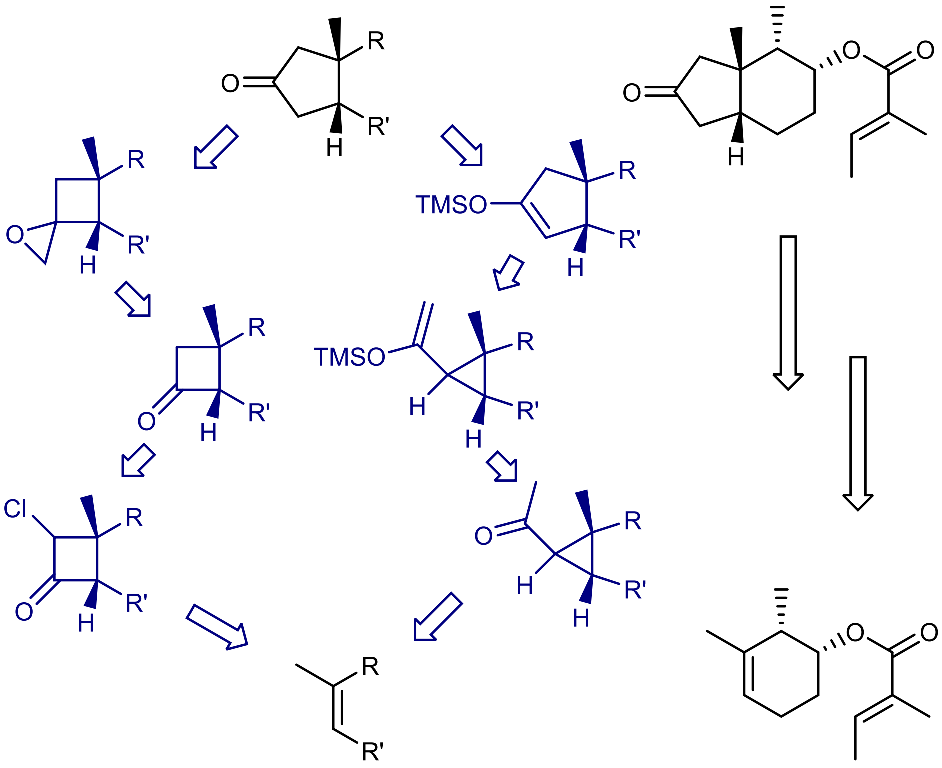


Figure 20 The application of a tactical combination of reactions to the retrosynthetic analysis of a homogynolide-B precursor.

The sequence of two TCs reduced the ring complexity by two rings and systematically removed three chiral centres, leaving a precursor with a single ring and two adjacent chiral centres for further retrosynthetic analysis. Even though LHASA was able to highlight interfering functionality in the non-modified portions of the precursors, the issues of regioselectivity and stereo-control in the application of TCs were not fully addressed and it remained with the user to judge and select the best route from the proposed alternative reaction sequences. It is evident that the application of prebuilt TCs is a strategy with the power to drive a deep analysis. In this respect each TC can be thought of as a reaction sequence template.

Forward and bidirectional strategies

The FORWARD program[16] was developed to complement SYNGEN. It was designed to investigate synthetic routes forwards from a library of starting materials using the same skeletal disconnection and re-functionalization strategy exploited by SYNGEN, but used an inverse of the SYNGEN reaction generator. The program operated by fitting sets of starting material skeletons to the target molecule skeleton working under severe constraints to control the combinatorial explosion of this process. The principle constraint was provided by the measurement of synthetic distance[120] in terms of the re-functionalization changes required for a given skeletal mapping. The program was workable as long as the synthetic distance measure was rigorously used to provide a strong strategic focus. It was found that when the constraints were ideally tuned for a specific target, FORWARD was able to find the same shortest convergent plans that SYNGEN generated. The logical union of SYNGEN and FORWARD was planned but has never been reported.

Conclusions

In broad terms the long range strategies employed by the LHASA program to focus on key simplifying reactions showed real promise that a computer could emulate the deep thinking abilities of an expert chemist. However these strategies were difficult to code and keep current. They also suffered from the inability to consider more than one strategy at a time due to the procedural nature of their implementation.

Hopkinson's approach using AI planning techniques[55] demonstrated that by assembling sets of independent strategy rules based on broad chemical concepts, a computer could simultaneously use these rules to successfully *argue* the best course of the retrosynthetic plan in terms of overall strategy and control.

The work by Kappos and Long [91,92] in preassembling tactical combinations of transforms as reusable synthetic sequences was demonstrated as providing another powerful technique able to apply a deeper analysis to a complex synthetic problem.

Hendrickson's approach of reducing the problem to abstract skeletal disconnections[1] has merit in that it provides the means to guide goal selection that favours an economic convergent synthesis. It also offers the means to guide the plan towards known starting materials while simultaneously reducing the combinatorial growth of the search space by using a powerful problem reduction approach.

Each of the reviewed systems has provided unique insights into the various problems and solutions of retrosynthetic analysis. However it is notable that no single system to date has attempted to unify the best approaches of the whole body of research.

By the late 1980s systems like the LHASA program were regarded as the most sophisticated and promising tools to aid chemists in designing chemical synthesis[121] and it was thought to be only a matter of time before they would become routine tools used in the laboratory. Indeed 10 years earlier in the late 1970s a consortium of pharmaceutical companies had initiated a collaborative project and subsequently invested much time and effort adapting the SECS program to their needs.[122] However the prediction of the routine use of such tools never truly came to fruition. One obstacle has been the widely held belief that a human chemist would routinely outperform the efforts of such a computer program. At that time this belief was reinforced by the limited depth of knowledge available to such programs and the costly work in terms of time and effort required to build the knowledge bases. Ultimately the effort to enhance the systems waned and they fell out of routine use.[25, 123]

One fruitful outcome of the cumulative research effort has been pedagogical. Retrosynthetic planning in terms of the logical repetition of the steps of selecting a strategy, recognising retrons and relating these to known reactions is now routinely taught at undergraduate level. Much of the research into retrosynthetic analysis has provided useful insights on how a chemist can apply strategic and tactical thinking to the synthesis problem. In recognition for pioneering this field E.J. Corey was honoured in 1990 with a Nobel Prize for "*the development of the theory and methodology of organic synthesis*".[7]

The ARChem program

The ARChem program is a commercial computer program designed to perform short retrosynthetic analyses of target molecules. [40] Its primary audience are process chemists seeking to devise viable synthetic routes. It currently does not support the recognition or manipulation of stereochemistry. The development of algorithms for *stereoselective* retrosynthetic analysis for use by ARChem is the subject of the rest of this thesis.

Automated generation of reaction rules

In contrast to LHASA and SECS, the *ARChem* program[40] does not require generalised reaction transforms to be written in their entirety by skilled chemists using a special language. *ARChem* automatically builds transforms by analysing the contents of reaction databases such as Methods of Organic Synthesis (MOS), Cheminform (CIRX) and Beilstein (Reaxys).

Reaction databases store the reactant and product molecules along with a set of atom-to-atom correspondences (Figure 21). *ARChem* uses a set of algorithms to locate and characterise a reaction descriptor to enable reaction examples to be clustered. A basic reaction core (Figure 22) is extracted from each database record using the atom-to-atom reaction map to identify the changed, created and broken bonds. The reaction core is extended to include relevant atoms and bonds that may have influence on the reaction mechanism. The extension algorithm

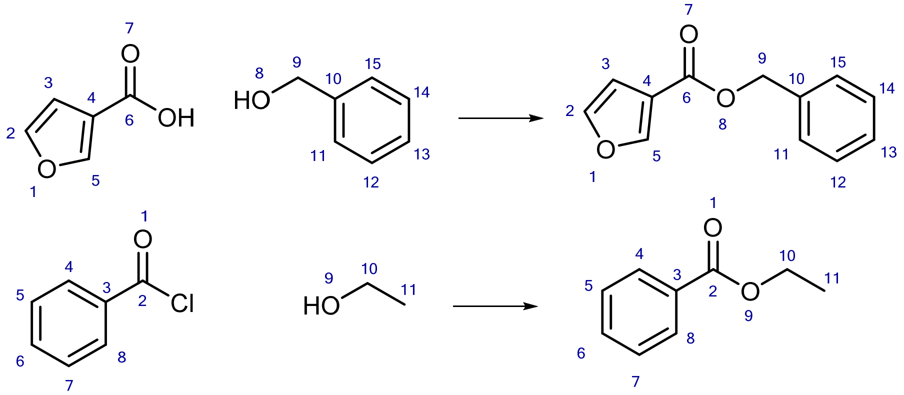


Figure 21 Example atom-atom mapped reactions.

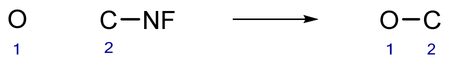


Figure 22 The extracted reaction core is only composed of the changed bonds and associated atoms.

searches for neighbouring or overlapping functional groups that are associated with the reaction core and adds them to the reaction descriptor. If present, leaving groups are identified and characterised as a nucleofuge (NF) or electrofuge (EF) and these characteristics are used to represent the extended reaction core in a generalised form (Figure 23).

Reactions with identical generalised cores are clustered together using a variant of the Morgan algorithm[66] to generate numerical comparison codes. These clusters enable the partitioning of the database into groups of related reactions. The final step is the creation of a complete reaction rule associated with the cluster, where the generalised NF or EF groups are replaced by a specific group selected as the most common representative for each of the generic leaving groups (Figure 24). This formulates the final transform containing both the retron substructure and the bond manipulation instructions needed to form the precursor molecule.

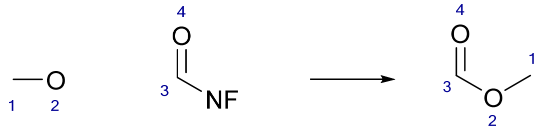


Figure 23 An extended reaction core with a generalised leaving group.

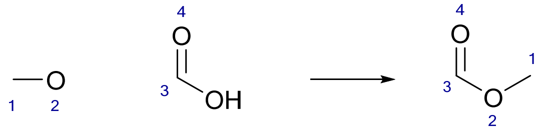


Figure 24 A finalised reaction rule replacing NF with the most frequent leaving group example.

Further statistical analysis of the example reactions, associated with each transform, provided evidence of the reaction utility (represented by the number of examples) and the range of environments (as the distance from the reaction site) that were likely to be amenable to the transform. The analysis also provided direct evidence of tolerated functional groups and hence an inference about potential interfering functionality. However the lack of published failed reactions necessitated that any such inference would be overly conservative, especially for transforms with very few example reactions. This method of automated reaction transform generation was tested on 10.8 million specific reactions provided via the Beilstein Crossfire database (now Reaxys).[97] The final transform set was selected from those generalised clusters that were supported by at least 5 example reactions and this generated roughly 105,000 unique reaction transforms.

In retrosynthetic operation *ARChem* selects transformations in a two-step process. The first step determines transform *relevance* for a particular target compound by comparing various global property counts, then checking detailed local atom/bond properties to ensure that the full requirements of the transform retron are met. The second step determined transform *applicability* by a detailed mapping of the retrons of each of the relevant transforms to the current target molecule to find matching locations to which to apply the bond modifications for precursor generation. An evidence based approach is used to evaluate if non reacting portions of the molecule may be interfering. Information about non interfering functionality is collated from the reaction database by noting the non-reacting functional groups that survive each reaction and comparing these to the situation in the current retrosynthetic analysis.

The rule building process is illustrated with an example of the Michael addition. Figure 26 shows the definition of the core reaction rule discovered by considering *only* the bonds modified during the reaction. The reaction core is a necessary but insufficient rule required to fully describe the requirements of the Michael addition reaction. Figure 25 shows the subsequent application of *ARChem* perception rules to extend a reaction core to the essential activating functionality of the reaction. The rule is extended to include neighbouring activating functional groups that are assumed to be necessary for the retron description. The reaction site extension algorithm

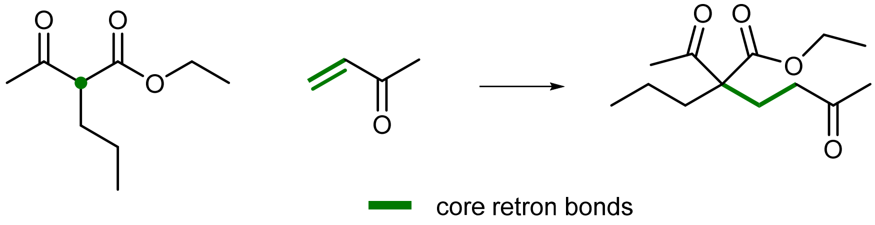


Figure 26 The reaction core of the Michael addition reaction. The changed bonds (including implied bonds to hydrogen) are coloured green.

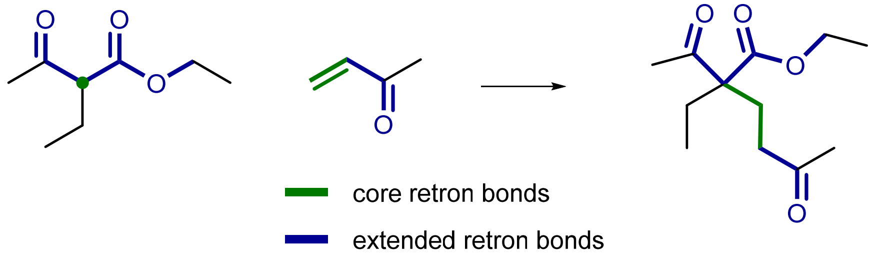


Figure 25 The extended core of the Michael addition reaction. The mechanistically essential but non-changed bonds are coloured blue. The extensions incorporate neighbouring functional groups but avoid unfunctionalised substituents.

utilises a small set of meta-rules to recognise likely mechanism types to guide the expansion of the retron site.

The *ARChem* program[40] currently employs an unsupervised heuristical search strategy to plan a synthesis. The main strategy executive largely follows Corey's original protocol (*vide supra*) with the addition of some categorised transform selection rules and a strategic transform prioritization scheme. Transforms are classified into four main categories. The first incorporates those transforms that aid the construction of the carbon skeleton (labelled as regular disconnective or RD). These can be selected by the strategy executive without restriction as long as they are applicable to the current target molecule. The second category incorporates opportunistic transforms such as FGI, bond-orientated FGI (BFGI) and non-dissociative rearrangements (NDR). These are used tactically if the application of the transform converts the current target to a known starting material or it subsequently enables an RD transform (which had otherwise failed to match). This strategy ensures that unproductive multiple sequential applications of FGI, BFGI and NDR transforms are suppressed. The third category is the simplifying functional group removal (FGR) transforms which can be applied as long as they do not follow a functional group addition (FGA) at the same site.

Transforms are prioritised based on a large number of merit factors including ranking RD and FGR above others and favouring low wastage, highest estimated yield, balanced disconnections, ring simplification, and well explored chemistry (as evidenced by the number of examples)[j]. These strategic selection and prioritization rules aim to reduce the combinatorial growth of the proposed routes, without inadvertently curtailing otherwise promising routes, by directing the analysis towards features found in (smaller) available starting materials. The strategy control is not yet adapted to exploit strategic features present in the current target molecule as found in an explicit goal based search system.

[j] Each transform rule is linked to a set of published examples from which the scope of the reaction can be extracted.

Chapter 2

Computer Representation of Stereochemistry

Introduction

This chapter describes an approach taken to recognise, represent, manipulate and compare a wide variety of stereogenic types (centres, axes or planes) typically found in molecules associated with organic synthesis. This includes the common stereogenic types found in target and precursor molecules, starting materials and also those found in chiral reagents and catalysts.

The computer representation of stereochemistry is essential for such tasks as: identifying and retrieving chiral starting materials from a supplier's database; [68] searching for potential asymmetric starting materials from the chiral pool by identifying molecules with appropriate stereochemical subunits; [19] recognising the subset of available starting materials that are *meso*-symmetric for potential application in *desymmetrisation* strategies; increasing the variety of reaction types that can be described in stereochemical terms to include amongst others, chiral sulphoxide chemistry and chiral allene chemistry; and selecting stereo-controlled rules from a database of retrosynthetic transforms. [55]

This chapter describes the scope of the stereogenic recognition problem; the types of structure diagram the recognition system will process; the representation of stereochemistry within computer systems as stereo descriptors; the representation and handling of framework models that describe the properties of stereo descriptors; algorithms for pattern recognition of stereogenic centres, axes or planes in chemical diagrams; the subsequent translation of these recognised stereogenic units into stereo descriptors; and the description of a novel concise and efficient algorithm for the rapid comparison of stereo descriptor pairs. The latter function is necessary to match structure and substructures queries incorporating stereochemical constraints to files of molecules and reactions.

The efficient retrieval and selection of reaction examples that support a specific reaction rule necessitates that a screening step is used to remove as many non-matching reactions as possible before a detailed substructure match is performed. An algorithm is described that perceives the set of stereochemical changes brought about by the reaction. The concept of a reaction hyperstructure is briefly described as is a novel adaptation of the Morgan algorithm used to code bond changes and stereochemical changes. The generated codes are used as screens for selecting the reaction examples.

Stereochemical Perception

Problem Scope

This section presents a brief survey of the types of stereogenic units found in molecules and discusses their relevance to synthetic organic chemistry particularly in the context of computer aided retrosynthetic analysis.

By far the most important types of stereogenic unit in organic molecules are those based on a tetrahedral configuration of ligands around carbon and the planar rectangular configuration of ligands around a carbon-carbon or carbon-nitrogen double bonds. These two types of centres are behind the vast majority of problems related to stereocontrol in synthesis design. Most if not all of the retrosynthetic analysis programs written to date have been equipped to deal with nothing other than these *classic* stereogenic units. Possibly the only exception was Choplin and Wipke's extension to the SECS program to handle pentavalent phosphorous chemistry[124] and octahedral coordination complexes.[125] However this study went no further than demonstrating canonical name and isomer generation algorithms. This lack of scope was partly due to the absence of a perceived need and partly due to limitations in the techniques used to represent molecules. The stereogenicity of allenes, sulphoxides,[126] sulphinimines,[127] sulphoximes,[128, 129] and other related functional groups has significance in modern synthetic organic chemistry and requires treatment in any modern computer implementation of stereochemical perception.

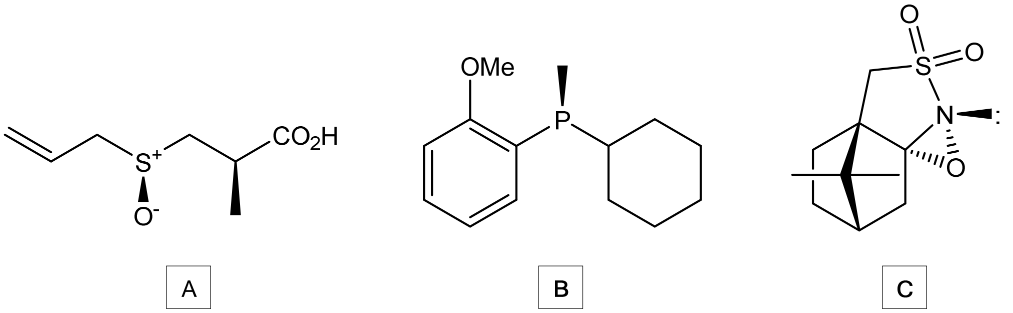


Figure 27 Examples of chiral sulphur, phosphorus and nitrogen stereogenic centres in natural products, and reagents: A) Alliin; B) (*S*)-CAMP; C) (1*R*)-(-)-(10-camphorsulfonyl)oxaziridines.

A ‘ligand’ at a stereogenic unit may include a lone pair of electrons occupying a position at a vertex of the tetrahedral or rectangular figures which circumscribe these stereogenic units. For example lone pairs actively participate in defining the stable tetrahedral configuration of sulphoxides, phosphines and oxaziridines, particularly where there is a significant energy barrier to inversion of the lone pair (Figure 27: *vide infra*).

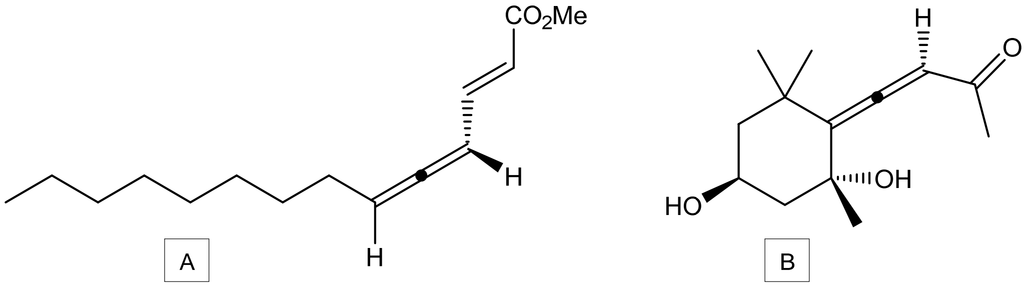


Figure 28 Stereogenic axes found in chiral allene natural products: the insect pheromones *of* A) Acanthoscelides obtectus; B) Romatea microptera. [46]

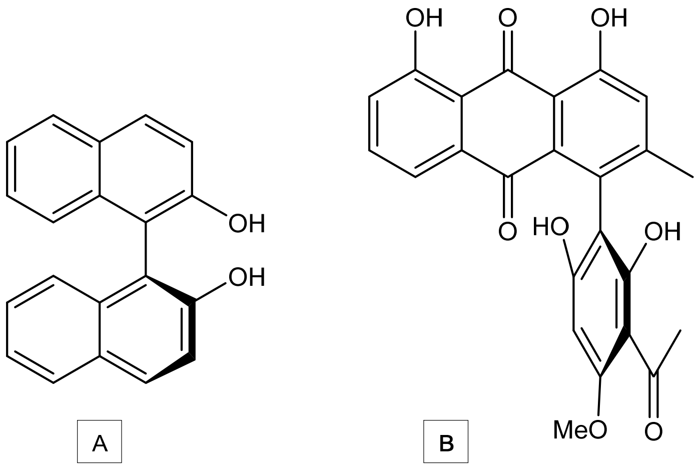


Figure 29 Stereogenic axes found in chiral atropisomers: A) (*S*)-Binol [49] **is representative of an important class of asymmetric catalyst ligand; B) the natural product Knipholone.** [50]

Allenes and their analogues represent a class of stereogenic centre based on a geometric figure with D_{2d} symmetry (Figure 28). Although allenes are generally uncommon in organic synthesis they occur in some natural products [130] and are occasionally useful synthetic intermediates.[131, 132] They have sufficient merit to be included within the scope of this project.

A common chiral ligand class employed in many asymmetric catalysts is based on the atropisomerism of 1,1’-biphenyl or 1,1’-binaphthyl moieties with additional bulky metal coordinating substituents largely at the *ortho* positions (Figure 29). The origin of the stereogenic axis in this case is a significant energy barrier to rotation around the 1,1′-biaryl linking bond due to steric repulsion by substituents.

Modern asymmetric synthetic methods frequently employ chiral catalysts that exhibit stereogenic centres based on a wide variety of coordination geometries other than tetrahedral, such as the square planar, trigonal bipyramidal, square pyramidal and octahedral configurations, where the ligands carry immutable 'off-metal' chiral centres or more rarely a 'chiral-at-metal' arrangement of ligands.[133] The latter class of catalyst is relatively rare due to the relative ease of non-dissociative epimerisation at many metal centres. [134]

Metallocene complexes (Figure 30) can exhibit planar chirality where the metal ligands are coplanar and dissymmetric. This type of chirality has wide applications in catalytic asymmetric synthesis. [135, 136]

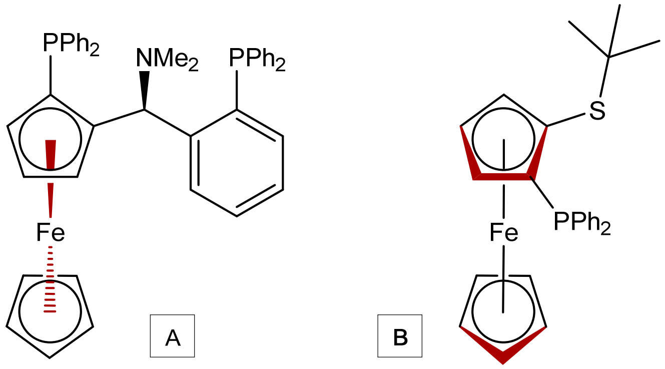


Figure 30 Planar chirality in some ferrocenes used as asymmetric catalysts: A) Taniaphos; [55] B) Fesulphos. [56, 57] Note the variations in drawing styles used to convey the presence of a stereogenic plane (highlighted in red).

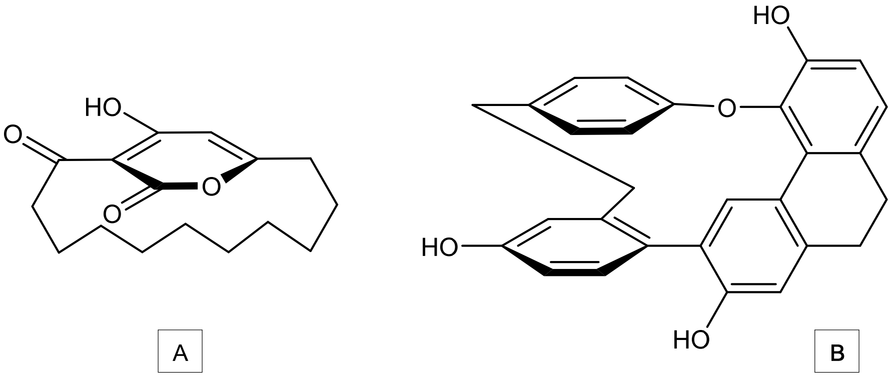


Figure 31 Molecules with planar chirality: A) A synthetic chiral cyclophane developed as a potential asymmetric catalyst ligand; [62] B) the natural product Cavicularin which exhibits both axial and planar chirality. [63]

Rarer forms of molecular chirality include the planar chirality exhibited in cyclophanes (Figure 31) or the helical axial chirality found in molecules such as helicenes and trans-cyclooctene (Figure 32). Rarer still is the topological chirality exhibited in catenane, [137] moebius [138] and knot molecules.[139] Sauvage's metal template synthesis of a chiral molecular trefoil knot is a recent highlight in this exotic area. [140]

It is appropriate that a stereochemical recognition system should at least attempt to represent as many of the types of the stereogenic units that occur in natural products, drug-like molecules, chiral reagents, auxiliaries and catalysts to support functions such as database retrieval, molecule feature perception and synthesis strategy selection. Equally it can be argued that solving the problem of computer recognition and representation of rare topological chirality has very little relevance in the context of general synthesis planning. Consequently planar, helical and topological chirality is considered outside the scope of this project.

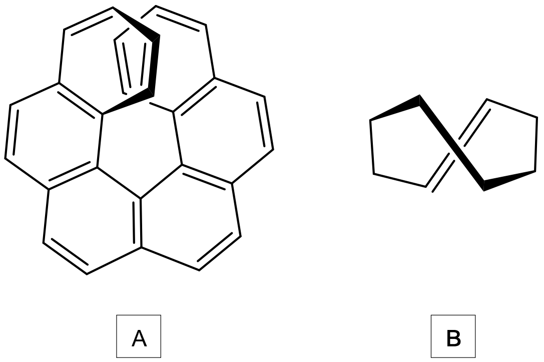


Figure 32 Chiral molecules with helical axes: A) (M)-hexahelicene; B) trans-cyclooctene.

Representations of Stereochemistry in Diagrams

The limitation of any method used to recognise stereogenic centres in a molecule diagram is ultimately dictated by the representational capabilities of the file formats used to transfer molecules and reactions between software programs. The ARChem program currently uses the MDL Molfile and RDfile formats to accept input of target molecules for synthetic analysis, load supplier's catalogues of available starting materials and to build the reaction transforms database. Certain types of structural moieties such as the representation of π bonded metal ligands are poorly represented in these formats which consequently limit the ability to recognise planar chirality and 'chiral-at-metal' complexes that involve metallocenes.

File exchange formats encode the stereochemical aspects of the drawing at a number of different levels of abstraction: in unprocessed 'as drawn' forms, either as the 2D coordinates of individual atoms with stylistic attributes (wedge, dash *etc.*) assigned to bonds; as 3D atom coordinates; in a processed form such as CIP designators;[141, 142] or as system specific stereo parity values.[143, 68, 64] Even within a single exchange format it is common that a number of alternative methods are available for coding stereochemical features with no requirement for the source application to use all or any of them. The most reliable approach for handing stereochemical information in these diverse exchange formats is to perceive the information directly from the 2D atom drawing coordinates and the drawn bond styles and ignore the processed forms. This approach currently precludes using 3D and linear notation formats as input sources. [k]

There are a number of distinct drawing conventions used to convey the three dimensional aspect of stereogenic centres within a 2D style structural diagram. Each of these drawing styles needs to be distinguished from the others before a specific recognition method can be selected for processing. Representatives of the most frequently used modern drawing styles are shown in Figure 33.

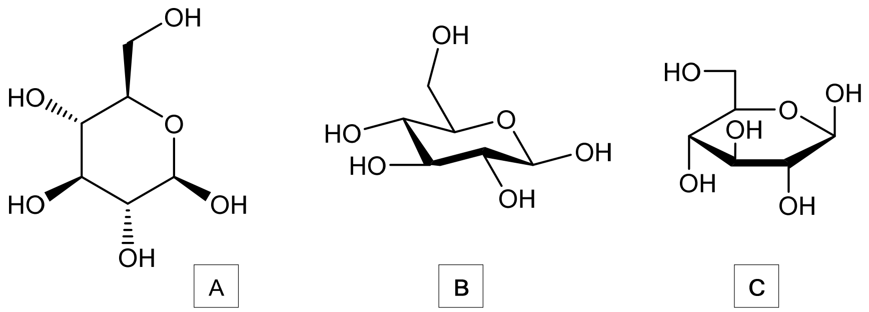


Figure 33 Representations of D-glucose: A) Skeletal formula (Natta projection), B) conformation projection, C) Haworth projection.

The *Natta projection* (A) is the diagram style used in the vast majority of published drawings of molecules and reactions. It is an unrealistic rendering style used to rapidly communicate the constitution of the molecule. Configuration information is overlaid on the skeletal framework by using wedge style bonds. Solid wedged bonds are used to denote bonds that rise up out of the surface plane and hashed wedge bonds denote those that descend below the surface plane in which the rest of the molecule has been flattened and projected. The narrow end of both styles of wedge bond is usually directed towards the stereogenic centre.[145] However in drawings predating the modern convention popularised by Natta, the bond style used to denote below plane bonds may be drawn solid with the wide end placed at the stereogenic centre. [146]

[k] So called “1D” formats such as SMILES[144]

The *Conformational projection* (B) is designed to emphasise the (pseudo) axial and equatorial placements of substituents in a specific ring conformation. Wedged and solid bonds are occasionally used on the ring bonds to denote which side of a ring is nearest the observer, usually in a slightly *above-and-side-on* projection. This highlighting style is not mandatory and is only used for emphasis. A ring is drawn in a particular conformation (*e.g.* chair, boat, twisted boat or envelope) and the relative magnitudes of the angles and alternate directions of the concertinaed bond sequence around the ring are used to deduce the axial and equatorial assignments of the substituents. It is notable that both the Natta and conformation styles may be used in a single diagram (Figure 34) which may occasionally lead to *hard-to-interpret* or erroneous situations if care is not taken (Figure 35).

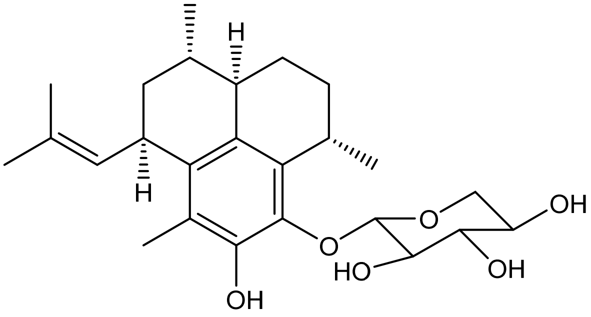


Figure 34 Mixed Natta and conformation styles used in a diagram of Pseudoptersin A.

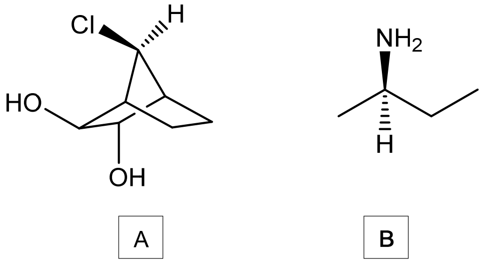


Figure 35 Examples of poor use of bond symbols: A) Ambiguous Natta style stereo bond symbols applied to a conformation projection style drawing. B) Conflicting bond symbols at the same stereo centre.

The *Haworth projection* (Figure 33: C) is largely limited to the representation of cyclic saccharides and is a stylised version of the conformation projection that dispenses with the ability to clearly denote the axial and equatorial placement of substituents.

Newman projection; Fischer projection; Saw-horse projection: These drawing styles are rarely encountered in reaction and starting material databases and will not be considered within the scope of this project. IUPAC actively discourages these drawing styles for the purposes of publishing diagrams of reactions or molecules. [68]

The application of stylistic drawing elements can be variable and inconsistent within one particular drawing convention. For example the portrayal of atropisomerism and planar chirality is often artistic in nature (Figure 36, also see Figure 29 to Figure 31). Notably this is prevalent in the situations where the IUPAC guidelines have not provided any specific recommendations. [68]

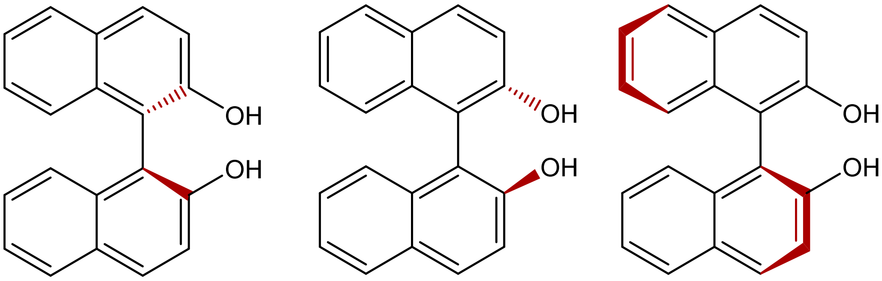


Figure 36 Varieties of artistic styles used to convey the atropisomer configuration of (*S*)-BINOL.

Stereo Perception Algorithms

This section describes an approach that recognises and encodes stereogenic centres found in a chemical diagram from the 2D atom coordinates and the 'wedge' and 'dash' stylistic attributes of the bonds. The types of stereogenic centre detectable by the method includes: tetrahedral centres; olefin-like and allene-like functional groups; biaryl type atropisomers; tetrahedral, square-planar, trigonal-bipyramidal, square-pyramidal and octahedral coordination complexes. Perceiving topological, planar and helical chirality was considered outside the scope of the project due to the exotic nature of these species and the lack of relevance to general synthetic problems. In principle the approach adopted could detect and represent the planar chirality of metallocenes but was not implemented due to current limitations in the machine representation of π bonded metal ligands.

A solution has been developed for recognising stereogenic centres in Natta style skeletal diagrams. In future this will be extended to handle the recognition of stereogenic centres in conformation projection style diagrams.

The stereo perception module consists of five main modules (Figure 37). The main stereo perception executive (B) uses a small database (D) of stereo centre recognition patterns designed to map the stereogenic features found in a chemical diagram (A) to an internal stereo descriptor representation suitable for computer manipulation (C). Each recognition pattern is linked to an associated reference model (E) to direct the construction of the matching stereo descriptor. Module E accesses a database of named 'framework' models characterised by their symmetry operations. These named models ('tetrahedron', 'octahedron' *etc.*) define the positions and interchanges of ligands around a stereogenic centre in terms of rotation and reflection permutations. The association of a stereo descriptor to a named model allows stereo descriptors to be compared for substructure and rule matching purposes.

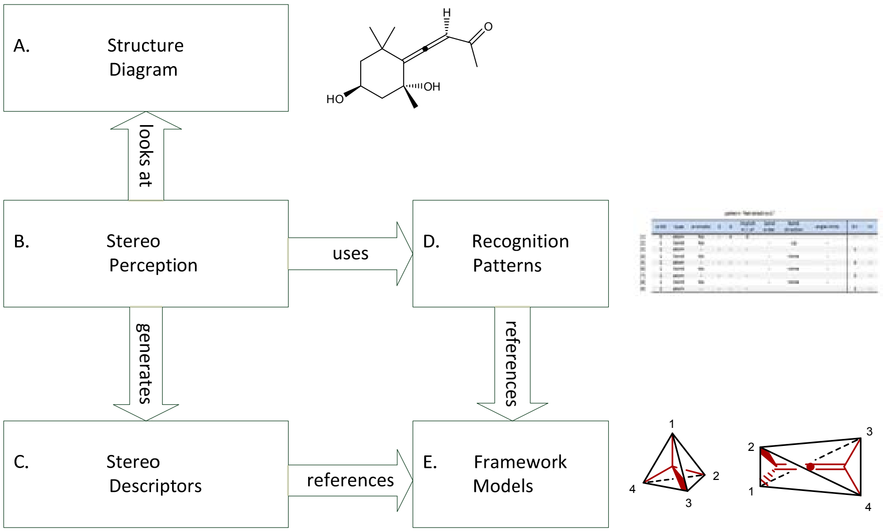


Figure 37 The Stereo Perception module architecture.

Stereo Descriptors

A variety of stereochemical descriptors have been defined for applications in chemical representation, systematic nomenclature and information retrieval systems. The most familiar descriptors are the Cahn–Ingold–Prelog (CIP) codes used in systematic nomenclature. In general the CIP descriptor method has proven hard to implement completely and correctly on a computer, [147] and attempts to do so have exposed flaws in the rules.[148, 149] Petraca, Rush and Lynch defined an alternative approach for generating parity descriptors [143] based on the Morgan naming algorithm [66] and extended this beyond tetrahedral carbon and olefins to some coordination geometries.[150] This method (in contrast to CIP descriptors) was readily amenable to implementation and use within a computer program without suffering the complexity and deficiencies of the CIP rules. Wipke and Dyott extended this Morgan based approach further and created the linear SEMA code [68] as an efficient tool for retrieving exact stereoisomers from a database of molecules.

None of the above descriptor techniques are suitable for the representation of stereochemistry in substructure queries or patterns. They require a complete description of the molecule to assign priorities as information remote from each stereogenic centre may be needed to differentiate ligand priorities.

Stereo descriptors capable of describing the local ligand arrangements in full structures and substructure representations are used in the SMARTS pattern and SMILES molecule notations, [144] and have been described for the SMD and MIF [151] molecule file formats. Internal ordered list representations of local ligand arrangements have been described for the LHASA [104] and CAMEO [152] programs. The Petraca, Rush and Lynch representation method also describes an ordered list approach.

An adaptation of the ordered list technique was adopted for the purposes of this research project. Each descriptor consists of a sequence of atom indices derived from the molecule connection table. The atoms cited in the descriptor are those in the ligand that connect to the stereogenic centre and the order of the atoms is determined by reference to a framework model that describes the geometry and placement of the ligands around the centre.

Special provision is made for ligands that are hydrogen atoms or lone pairs of electrons. It is common practice in cheminformatics applications to internally represent molecules in a reduced form where hydrogen atoms are removed to reduce storage space and improve algorithm running times. Instead hydrogen is represented as a count on the atoms to which they are attached. Exceptions occur when the bond to the hydrogen atom is a stereodefining bond or the atom is a hydrogen isotope. To normalise these possibilities all hydrogen atoms cited in a stereo descriptor are represented by a special code, as are lone pairs. For illustrative purposes these codes are represented as ‘H’ and ‘**:**’ respectively. Internally these are represented by special integer codes.

Each descriptor is labelled using the name of the referenced framework model and the descriptor is indexed by the atom or bond at the centre of the stereogenic unit. The role of the framework model is fully discussed in the following two sections. The general form of the descriptor is:

$[v_1, v_2, v_3, \ldots]_{model}$ where v_n is the index of the atom at position *n* in the named model.

Table 1 illustrated the representation of a selection of stereo descriptors derived from some sample molecule diagrams. The generated stereo descriptors are created in a non-canonical form and determining equivalence between a pair of descriptors requires the application of a comparison algorithm which is discussed in a subsequent section. The method of generating the descriptors is presented in the following section.

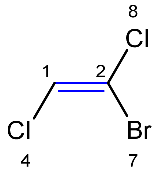
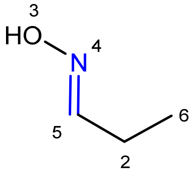
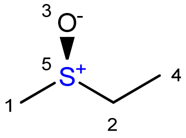
Molecule Diagram	Stereo Descriptors
	$[8\ 7\ 4\ \mathrm{H}]_{ol}$ *or* $[4\ \mathrm{H}\ 8\ 7]_{ol}$
	$[3 : 2\ \mathrm{H}]_{ol}$ *or* $[2\ \mathrm{H}\ 3 :]_{ol}$
	$[3\ 1\ 2 :]_{tet}$ *or* $[3\ 2 : 1]_{tet}$ *or* $[3 : 1\ 2]_{tet}$

Table 1 **A stereo descriptor cites the indexes of ligand atoms attached to a stereogenic centre (blue) in an order that is determined by a reference framework model. 'ol' identifies the olefin framework; 'tet' identifies the tetrahedron framework. Equivalent orderings of the descriptor are possible depending on the initial ligand atom chosen in the molecule diagram. 'H' and ':' are codes for hydrogen and lone pairs.**

Stereo Recognition Patterns

Stereogenic centres drawn in Natta style diagrams are discovered by applying a pattern matching algorithm using a set of structural patterns to find tetrahedral, olefin, allene, bi-aryl atropisomer and some common configurations of metal coordination complexes. The patterns are divided into two groups, one concerning stereogenic centres rooted at an atom (tetrahedron, allene and coordination complexes) and the other are for those rooted at a bond (olefin, some atropisomers).

The first step in this pattern matching approach is the elimination of atoms and bonds that cannot give rise to a stereogenic centre. This includes all single bonded terminal atoms (-CH_3, -NH_2, -OH, -SH etc.) and single bonded non-branched atoms (-CH_2-, -NH-, -O-, -S-, etc.) and all atoms and double bonds drawn in aromatic systems.[13, 153, 47]

The algorithm proceeds by visiting in turn each qualifying atoms in the structure, testing the environment around the root atom according to the attributes of each atom pattern. Likewise each qualifying bond is visited and the set of bond rooted patterns are tried. If a pattern is found to match, instructions contained within the pattern are used to construct a stereo descriptor. The stereo descriptor is stored and indexed by the root atom or bond.

The structure of a recognition pattern is based on an ordered graph walk starting at a root atom or bond. Each pattern consists of a series of records that follow the ordered graph walk and specify required atom and bond attributes that must match the tested structure for the walk to continue (Figure 38).

pattern “tetrahedron1”

	orbit	type	aromatic	Z	D	implicit H / LP	bond order	bond direction	angle limits	EV	IV
[1]	0	atom	No	-	4	0				-	-
[2]	1	bond	No				-	Up	-		
[3]	2	atom	-	-	-	-				1	-
[4]	1	bond	No				-	none	-		
[5]	2	atom	-	-	-	-				4	-
[6]	1	bond	No				-	none	-		
[7]	2	atom	-	-	-	-				3	-
[8]	1	bond	No				-	none	-		
[9]	2	atom	-	-	-	-				2	-

Figure 38 A representative stereo recognition pattern. Atoms and bonds surrounding a candidate stereocentre are sorted into a specific order and must match the rows of the pattern to confirm the presence of the stereocentre. If a match occurs the explicit vertex (EV) and implicit vertex (IV) indices are used to build the descriptor. Z is atom type; D is the connectivity count; EV is the position the matched atom is placed in a stereo descriptor; IV is the position an implicit hydrogen or lone pair code is placed in the stereo descriptor.

Attribute values in the pattern can be explicit and must match the atom or bond attribute or they can be omitted to indicate that the test can be skipped. The primary attribute is the ‘*orbit*’ or shell number that gives the distance of the atom or bond from the pattern root. These orbit numbers alternately reference atoms and bonds as the walk from the root is traversed and this is illustrated in Figure 39.

The ‘*type*’ attribute (Figure 38) has the values ‘*atom*’ or ‘*bond*’ and is used to confirm the pattern is correctly aligned on the right orbit. Atom rooted patterns have all atoms located on even orbits while in bond rooted patterns atoms are located on odd orbits. The ‘*aromatic*’ attribute has values ‘*yes*’ or ‘*no*’ and applies to both atoms and bonds. It is used to rapidly curtail unproductive pattern matching in aromatic regions of the molecule.

The next three attributes apply only to atoms. ‘*Z*’ lists the allowed atom types (as a list of atomic symbols). A value may be absent meaning that any atom type qualifies. The ‘*D*’ (degree) attribute specifies the number of bonds that must be directly drawn to the atom. This is used primarily on root atoms to efficiently filter out patterns that would otherwise fail on a later pattern record. The ‘*implicit H/LP*’ attribute handles drawings where implicit lone pairs or hydrogen atoms are present at the stereogenic centre. This frequently occurs in drawings of sulphoxides and oximes and at abbreviated tetrahedral carbon centres when hydrogen atoms are deliberately omitted (Figure 40). The number of implicit hydrogen atoms and lone pairs is calculated for the atom and the sum of these quantities is matched to the value in the pattern record.

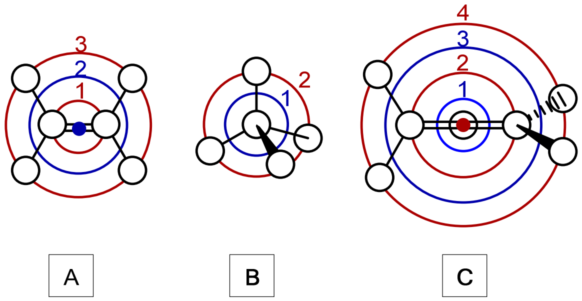


Figure 39 **Stereogenic atom and bond orbits. Blue circles denote bond orbits and red circles denote atom orbits: A) illustrates the orbits for an ‘olefin’ pattern; B) a ‘tetrahedron’ pattern; C) an ‘allene’ pattern. The numeric values give the orbit distance from the root atom or bond at orbit 0**

The final three pattern attributes are applied only to bonds. The ‘*bond order*’ attribute is only checked if it is not overridden by the ‘*aromatic*’ attribute. When the latter is set to ‘*yes*’, it takes priority over ‘*bond order*’ and serves to remove double bonds drawn in aromatic rings from consideration. The ‘*bond direction*’ attribute enables the stereo drawing style to be tested. The values ‘*up*’ and ‘*down*’ indicate that the bond was drawn in a solid wedge or hashed wedge style with the pointed end directed towards the pattern atom immediately preceding the bond record. The value '*none*' is used for in-plane bonds The '*angle limits*' attribute sets limits on the minimum and maximum angular separation of the pattern bond to a proceeding bond in the pattern. Both must be connected to the same parent atom. Angle limits are required whenever the '*implicit H/LP*' attribute is on a parent atom. For example constraining a pair of bonds to subtend an angle in the range 5 - 175 degrees around a tetrahedral centre allows us to establish that the implicit hydrogen atom or lone pair of electrons lies in the region formed by the (larger) opposing angle subtended by the same bonds (Figure 40, A). Likewise the '*angle limit*' constraint allows the position of implicit hydrogen and lone pairs to be deduced in olefin, imine (Figure 40: B, C) and allene drawings.

The discovery of a stereogenic centre is confirmed when all atom and bond records in a pattern are matched, otherwise the remaining recognition patterns are tried one by one until they are either all exhausted or a matching pattern is found.

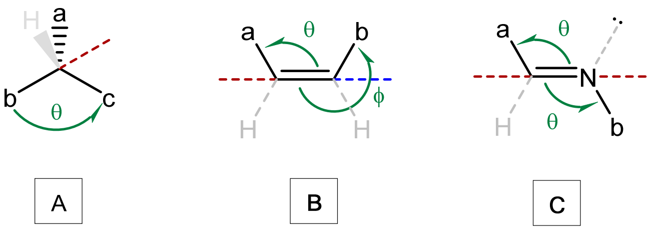


Figure 40 Stereo recognition pattern angle limits. The angle limits θ (5 – 175 degrees) and ϕ (185 – 355 degrees) constrain the positions of bonds in stereo recognition patterns. The red dashed lines indicate the *maximum* extent of θ. The blue dashed lines give the *minimum* extent for ϕ. The grey dashed bonds indicate the deduced position of implicit hydrogen or lone pairs.

Once a match is confirmed the perception executive uses the values in the *EV* (explicit vertices) and the *IV* (implicit vertices) fields to construct a stereo descriptor. The matched pattern is scanned record by record and for each non-zero *EV* field the atom index of the corresponding matched atom is placed in the stereo descriptor at the position specified by *EV*. If a non-zero value is found in the *IV* field a check is performed on the associated atom to determine if the implicit item is hydrogen or a lone pair. A special code is placed in the descriptor at the location indicated by *IV* to indicate the type of implicit entity found.

When a root atom or bond is selected to be matched to a recognition pattern, a tree walk[l] through the molecule connection table is generated in preparation for the record by record comparison checks. The walk is represented by a tree rooted at the atom or bond under consideration.

The first step in growing the tree is to extend out two orbits in the case of atom roots and one

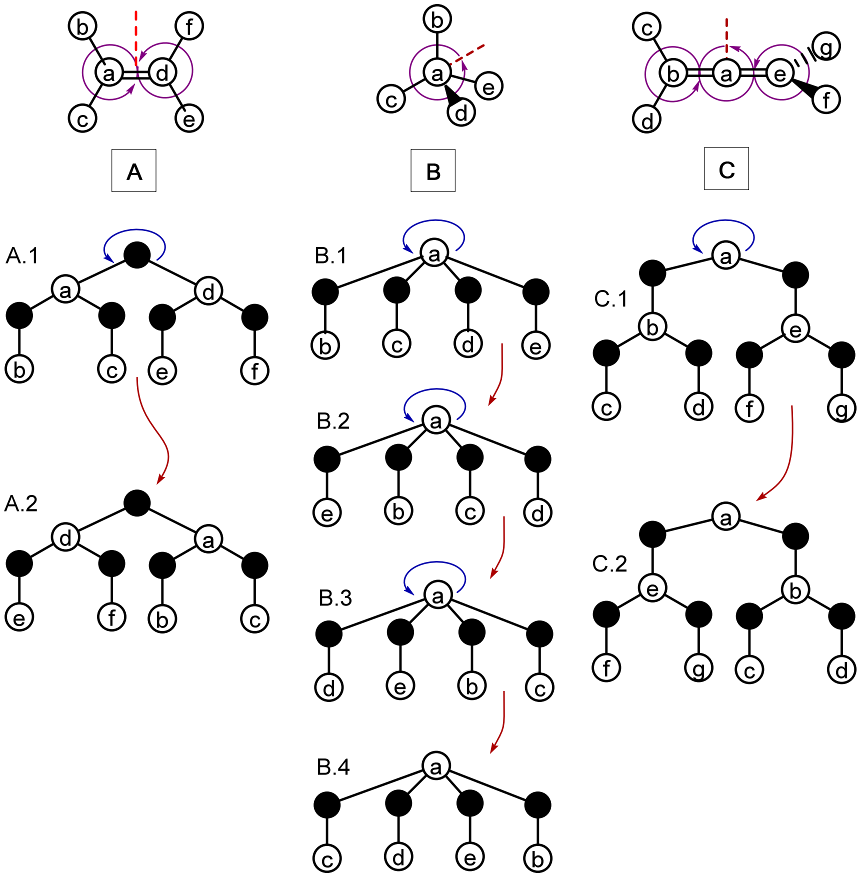


Figure 41 **Trees derived from a candidate stereogenic unit. The tree branches are ordered by increasing anticlockwise angle from a parent bond. Separate trees are generated for each alternative starting point at the root node.**

Solid black nodes are bonds; open nodes are atoms. Dashed red lines mark the start of the first angle ordering around the root. The purple circular arrows show the sorted angle direction used to transform structures A, B, C into the rooted trees A.1, B.1 and C.1. The blue arrows show the right hand side to left hand side branch rotations used to transform the tree A.1 to the equivalent tree A.2 *etc*.

[l] A 'walk' through a graph is represented by an ordered list of atoms and bonds. A 'tree walk' includes branches off the main walk were each branch is appended to end of the parent walk. This is applied recursively when branches occur on branches.

orbit in the case of bond roots. In both cases the growth of the tree is halted on an atom orbit. The tree is then sorted so that the branches are arranged in increasing anticlockwise order based on an analysis of the position of each atom in the chemical diagram data. The starting point for this ordering at a root is arbitrary and so it may not initially align correctly with the pattern under test. To take this into account a series of trees are grown, each choosing a different branch bond as a starting point (Figure 41). Thus the maximum of number of trees needed to completely test a pattern is equal to the degree of the root atom or two in the case of root bonds.

In practical terms the expensive growth of multiple trees is avoided by growing and then rearranging a single tree. When the next tree orientation needs to be tested (because the prior one failed to match), the right most branch attached to the root node is detached and reattached to become the leftmost branch (see Figure 41 sequences A.1 to A.2; B.1 to B.4; C.1 to C.2).

Tree growth is incremental and executed on demand to improve processing efficiency (Figure 42). If a match test fails early in the process there is no need to grow the tree any further and

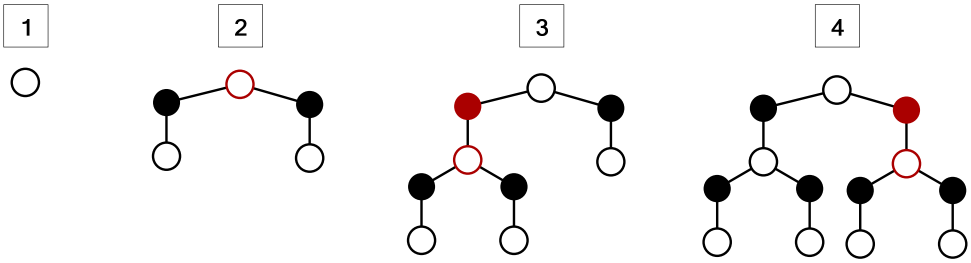


Figure 42 The incremental construction of a stereo recognition tree showing an ‘allene’ tree grown in four steps. Filled nodes are bonds, open nodes are atoms. Growth points are from the parent atoms (open red) and angle sort orders are measured from the grandparent bonds (filled red). All nodes are horizontally aligned in the diagram if they occupy the same orbit.

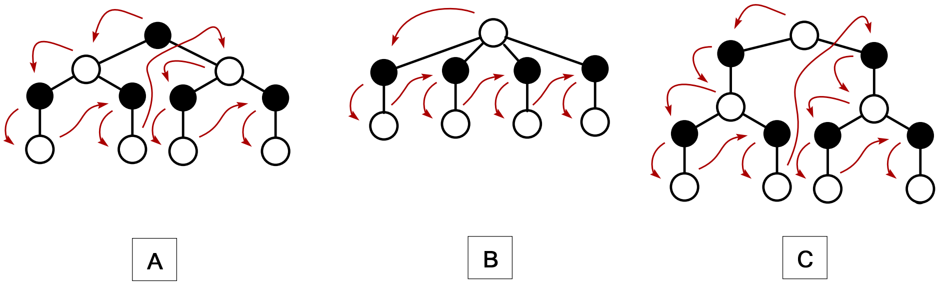


Figure 43 Depth first pre-ordered sequence traversal of the stereo recognition trees from their root nodes. A) ‘olefin’; B) ‘tetrahedron’; C) ‘allene’.

the partial tree is abandoned. When a recognition pattern requests a bond that is not present in the partially grown tree, the next two orbits are grown from the parent atom of the bond and sorted into an anticlockwise angle order measured from the grandparent bond (Figure 42).

The expanded search tree is traversed in a depth first pre-ordered manner (Figure 43), a process recursively described as: *visit the root; for each sub-tree (from left-most to right-most) traverse the sub-tree*. During tree expansion the left-most branch is grown first to satisfy the depth first pre-order traversal scheme.

Each recognition pattern is arranged to follow the same traversal order as a search tree so that a predictable correspondence is maintained during the comparison phase.

Geometry Frameworks

The elements of a stereo descriptor are ordered and this ordering is associated with a named geometric framework (Figure 44). These frameworks represent an idealised solid geometric figure on which atoms that are in the immediate orbits of a stereogenic centre are placed at the

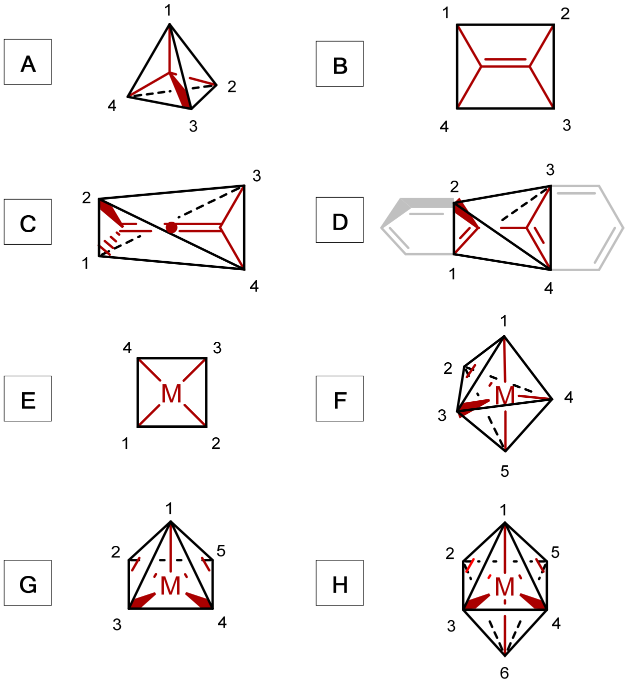


Figure 44 **Framework models that define the configurations of ligands around common stereogenic centres:**
A) 'tetrahedron'; B) 'olefin'; C) 'allene'; D) 'atropisomer'; E) 'square planar';
F) 'trigonal bipyramid'; G) 'square pyramid'; H) 'octahedron'.

The red figures show the mappings of molecule substructures onto the frameworks. Note that structural motifs C and D have identical framework geometries but with different substructure mappings.

vertices. The vertices of these frameworks are numbered uniquely but arbitrarily and the numbering maps the framework model to associated stereo descriptors such that each atom cited in a descriptor is located by its sequence position to the corresponding numbered vertex in the framework. Blower *et al* [154] briefly describe this approach as the basis for handling stereochemistry in the CAS Registry File, but very few details are described. Alternative approaches have been described by Petraca *et al* where the stereo descriptors are reduced to a single parity value based on the ordering of connection table atom numbers when mapped to an appropriate geometric framework. [143, 150] These ideas were later adapted by Wipke and Dyott as the basis for their stereochemically extended Morgan algorithm (SEMA) codes. [68] Dietz describes an approach that dispenses with frameworks and formulates stereo descriptors that are mapped to '*conformational paddle wheels*' [155] in an attempt to overcome a perceived limitation of framework models. Rohde describes an alternative approach were all geometries are decomposed to their constituent arrangements, as sets of 3-atom idealised angles, 4-atom idealised dihedral angles and 4-atom '*orientation*' helices. [156]

The approach adopted for the ARChem system was to use a flexible adaptation of the geometric

	description	*point group*	*rotation operations*	*rotations*	*vertex count*	*descriptor permutations*	*cosets*	*mirror pairs*
				r	*v*	*v!*	*v!/r*	*v!/2r*
A	tetrahedron	T_d	$E + 8C_3 + 3C_2$	12	4	24	2	1
B	olefin	D_{2h}	$E + C_{2(z)} + C_{2(y)} + C_{2(x)}$	4	4	24	6	-
C	allene	D_{2d}	$E + C_2 + 2C'_2$	4	4	24	6	3
D	square planar	D_{4h}	$E + 2C_4 + C_2 + 2C'_2 + 2C''_2$	8	4	24	3	-
E	trigonal bipyramid	D_{3h}	$E + 2C_3 + 3C_2$	6	5	120	20	10
F	square pyramid	C_{4v}	$E + 2C_4 + C_2$	4	5	120	30	15
G	octahedron	O_h	$E + 8C_3 + 6C_2 + 6C_4 + 3C_2$	24	6	720	30	15

Table 2 Symmetry characteristics of the framework models.

pframework model described by Petraca *et al*. Any number of framework geometries can be defined in an external file enabling new geometric frameworks to be added in the future. Each framework object is named and characterised by its symmetry in the form of permutation operators. These operators describe how the vertices in the geometric framework can be interchanged by rotation and reflection. These permutations can be used to reorder the atoms in a stereo descriptor to enable a pair of stereo descriptors to be sequentially compared. The comparisons that are of interest are: determining if a pair of stereo descriptors is identical; if they are mirror images of each other; or they are some other isomer of each other. This stereo descriptor comparison function is an integral part of the solution for stereochemical structure and substructure matching. The substructure matching algorithm determines the overall stereochemical relationship of a pair of structures and is described in more detail in a subsequent section.

A stereo descriptor with *v* atoms can be sequenced in *v!* different permutations in which v!/r cosets exist whose members are interchangeable by the *r* proper rotations belonging to the point group of its associated geometric framework (Table 2). There also exist v!/r non-rotation permutation operations that transform a member belonging to one coset to a member of another coset. These latter transformations are not necessarily symmetry permutations that preserve ligand configuration and in general may correspond in chemical terms to geometric

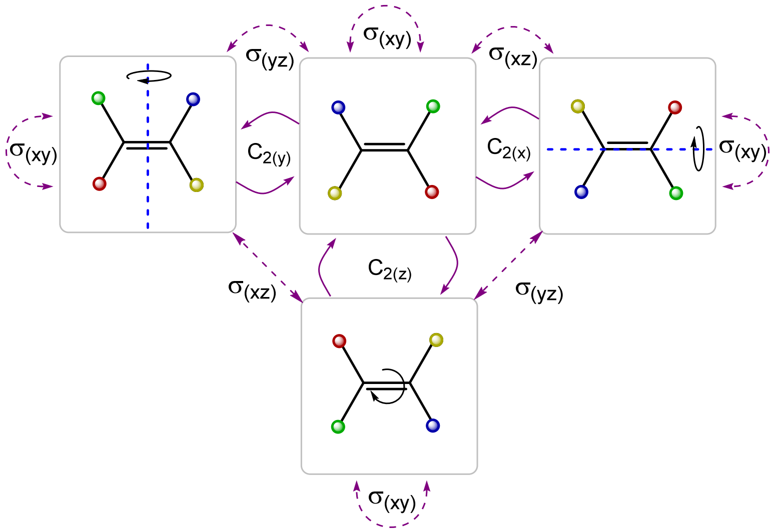


Figure 45 The vertices of the 'olefin' framework can be placed in 4 alternative positions by rotation as shown by the cycle graph of the D_{2h} point group. These interrelated positions constitute all 4 members of one of 6 possible cosets. In planar geometries the members of each coset are interrelated by reflection operations.

isomerism or constitutional isomerism.[m] Cosets may be symmetrically paired by a reflection permutation, though in planar geometries (such as ‘olefin’ and ‘square planar’) the reflection operation is degenerate and transforms a permutation into another of member the same coset.

The interrelationships by rotation operations are depicted as a cycle graph. Reflection transformations in planar geometries are between members of the same rotation coset as shown in Figure 45 for the ‘olefin’ framework while non-planar geometry frameworks associate the rotation cosets in pairs as shown in the cycle graph for the ‘allene’ framework (Figure 46). The existence or non-existence of reflection relationships between cosets is directly related to the observation that ligands arranged around non-planar stereogenic centres can exhibit chirality (depending on the numbers and placement of non-equivalent ligands) while planar stereogenic centres are non-chiral.

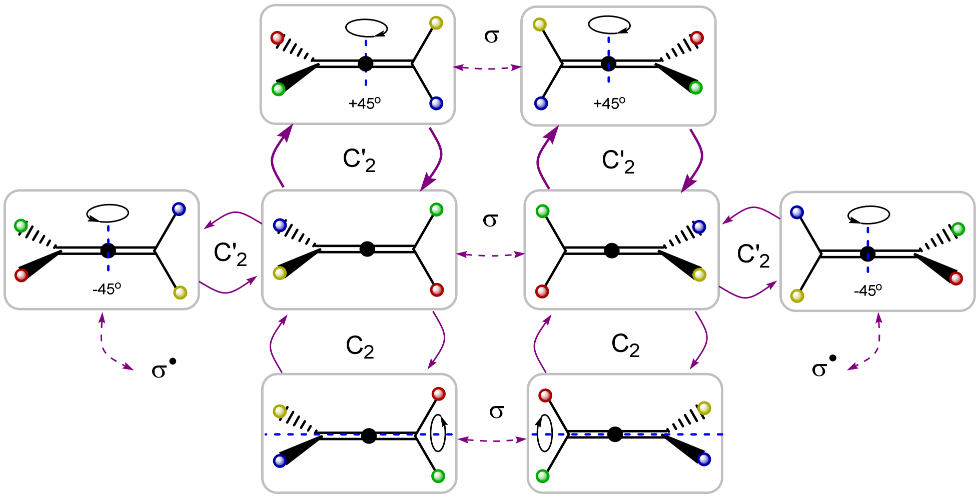


Figure 46 The vertices of the ‘allene’ framework can be placed in 4 distinct positions by rotation as shown by the cycle graph of the D_{2d} point group. These interrelated positions constitute all 4 members of one of 6 possible cosets. In non-planar geometries the left coset is paired with its mirror coset by reflection operations between pairs of coset members.

Comparing Stereo Descriptors

The stereochemical correspondence between two molecules is determined in three phases. First an atom and bond embedment must be found using a (sub) graph isomorphism algorithm to establish constitutional equivalence between two molecules or a substructure and a molecule.

[m] In Figure 44 a permutation that exchanges ligands at vertices 1 and 4 in the ‘olefin’ framework corresponds to cis-trans stereoisomerism. However a permutation that exchanges ligands at 1 and 2 corresponds to constitutional isomerism.

Second using the established embedment, the relationships between each pair of corresponding stereogenic centre are determined. Finally the entire group of stereo relationships are compared to determine if the molecules are either identical; enantiomers; or diastereomers. For example if it is established that all stereo descriptor pairs belong to mirror cosets we have enantiomers.

This section describes a fast algorithm that satisfies the second requirement, comparing two stereo descriptors. The algorithm is applicable to all geometries and requires only the set of proper rotation operations and a reflection operation of the framework model. A precondition for the application of the algorithm is that the stereo descriptor pair has a complete atom correspondence, a condition established in the graph isomorphism phase. The purpose of the algorithm is to establish to which cosets the descriptor pair belong: the same coset; a mirror pair of cosets; or some other pair of cosets.

An algorithm for the comparison of tetrahedral centres has been described that performs pairwise swaps of ligands until congruence between descriptors is achieved. [62, 104, 143] An even number of swaps establishes that the stereocentres are identical while an odd number establishes the centres are mirror images. This approach is applicable to olefins where an even number of swaps establishes equivalence. However this simple procedure cannot be extended to other geometries. Petraca and Rush have described a series of complex algorithms for

A	E	1	2	3	4
B	C^2_3	1	3	4	2
C	C^1_3	1	4	2	3
D	C_2	2	1	4	3
E	C^1_3	2	3	1	4
F	C^2_3	2	4	3	1
G	C^2_3	3	1	2	4
H	C^1_3	3	2	4	1
J	C_2	3	4	1	2
K	C^1_3	4	1	3	2
L	C^2_3	4	2	1	3
M	C_2	4	3	2	1
		1	*2*	*3*	*4*

σ	2	1	3	4
	1	*2*	*3*	*4*

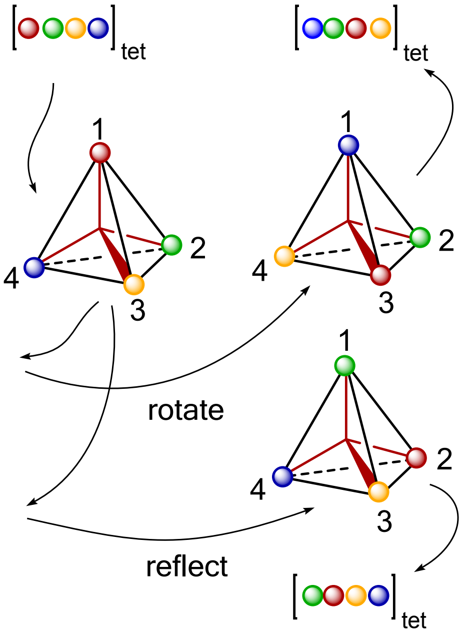


Table 3 **The rotation and reflection permutation table for the tetrahedron model. The figures illustrate the actions of rotation and reflection on a stereo descriptor (shown in cartoon form) in relation to the equivalent actions on the framework model.**

comparing pairs of square planar centres or octahedral centres using a hierarchy of coaxial vertex groupings, adjacent pair swapping and ligand reordering procedures to establish equivalence or differentiation by counting the number of ligand exchanges. [150] The principles underpinning the two methods could in theory be used to extend the swapping approach to other geometries but a simpler solution was required for the ARChem project.

An efficient comparison algorithm applicable to *any* geometry model was developed that selects the required rotation operator without the need to conduct a pairwise swapping search. The algorithm picks two atoms in the first stereo descriptor as probes and notes their positions. The second descriptor is searched to find the corresponding positions of the probes and from these pairs of position correspondences, the single rotation needed to bring both of these descriptors into congruence can be derived from a lookup table. A number of factors slightly complicate this

A	E	1	2	3	4	5	6
B	C^1_4	1	3	4	5	2	6
C	C^2_4	1	4	5	2	3	6
D	C^3_4	1	5	2	3	4	6
E	C_2	2	1	5	6	3	4
F	C^1_3	2	3	1	5	6	4
G	C^1_3	2	5	6	3	1	4
H	C^1_4	2	6	3	1	5	4
J	C^2_3	3	1	2	6	4	5
K	C^1_4	3	2	6	4	1	5
L	C_2	3	4	1	2	6	5
M	C^1_3	3	6	4	1	2	5
N	C^3_4	4	1	3	6	5	2
P	C^1_3	4	3	6	5	1	2
Q	C^2_3	4	5	1	3	6	2
R	C_2	4	6	5	1	3	2
S	C^2_3	5	1	4	6	2	3
T	C^3_4	5	2	1	4	6	3
U	C_2	5	4	6	2	1	3
V	C^2_3	5	6	2	1	4	3
W	C^2_4	6	2	5	4	3	1
X	C_2	6	3	2	5	4	1
Y	C^2_4	6	4	3	2	5	1
Z	C_2	6	5	4	3	2	1
		1	*2*	*3*	*4*	*5*	*6*

σ	1	5	4	3	2	6
	1	*2*	*3*	*4*	*5*	*6*

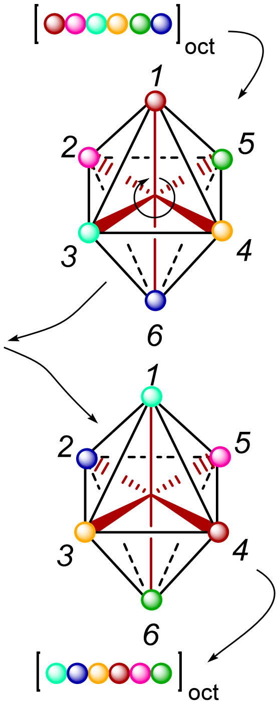


Table 4 Rotation and reflection permutations for the octahedron model. The figures demonstrate the action of rotation M on a stereo descriptor (shown in a cartoon form) in relation to the equivalent action on the framework model.

approach. First: stereo descriptors may contain multiple 'place holder' atoms representing hydrogen and lone pairs. For example *cis* or t*rans* di-substituted olefins require two hydrogen place holders which cannot be distinguished from each other when searching for a correspondence between descriptors. The probe search must skip these place holders. Second: the two probe atoms must lie on different rotation axes to allow the congruence transformation to be fixed. Choosing two coaxial atoms would allow the axis passing through them to remain rotatable preventing the fixed alignment of the two descriptors.

The algorithm requires a number of data tables to be constructed in advance. A rotation permutation table was set up for each geometric framework. [64] The identity permutation was omitted from the table and a single reflection permutation was added for non-planar geometries. As an example Table 3 and Table 4 list the rotation and reflection permutations for the tetrahedron and octahedron framework models and demonstrates their action on cartoons of a stereo descriptor. These permutation tables are loaded at program start up and

Inputs: *$\{R_1 \ldots R_k\}$ is the set of k rotation permutations for the geometric framework model M containing V vertices.*

Operation: *S is a set of coaxial vertices; s is a vertex; r is a rotation permutation (as a vector)*

Outputs: *$A_M[v]$ is the set of coaxial vertices of vertex v in framework model M*

```
FOR ALL: $u \in \{1 \ldots V\}$
  $A_M [v] \leftarrow \emptyset$                 (empty all coaxial sets)

FOR ALL: $r \in \{R_1 \ldots R_k\}$
   $S \leftarrow \emptyset$
   FOR ALL: $v \in \{1 \ldots V\}$
     IF: $r [v] = v$                (is the vertex v stationary under the application of rotation r?)
       $S \leftarrow S \cup \{v\}$

   IF: $|S| > 1$                    (there is an axis if there is more than one stationary vertex)
     FOR ALL: $s \in S$
       $A_M [s] \leftarrow S - \{s\}$       (record all the other vertices on the axis for vertex s)
```

Algorithm 1 Constructing the coaxial lookup tables.

	1	2	3	4	5	6
A_{tet}	{ }	{ }	{ }	{ }	{ }	{ }

	1	2	3	4	5	6
A_{oct}	{6}	{4}	{5}	{2}	{3}	{1}

Table 5 The generated coaxial tables for the tetrahedron and octahedron frameworks derived from Table 3 and Table 4 respectively by Algorithm 1. Note that the tetrahedron framework model has no coaxial vertices.

immediately processed to generate the coaxial and the congruence lookup tables. The coaxial table enables the algorithm to avoid the selection of inappropriate coaxial probes while the congruence table enables rapid lookup of the rotation permutation required to attempt to bring two stereo descriptors into alignment.

Algorithm 1 builds the coaxial lookup tables for each framework by inspecting each rotation permutation and noting which positions are stationary. If more than two positions in any one permutation are stationary they must be coaxial and are collected and stored in the vector ***A***. At completion the element ***A [v]*** holds the set of all the positions coaxial with position ***v*** in either a stereo descriptor or a framework model (Table 5).

Inputs: *{R_1 … R_k} is the set of k rotation permutations for the framework model M containing V vertices.*

Operation: *u, v are vector indices; r is a rotation permutation (implemented as a vector)*

Outputs: *C_M [u][v] is the set of rotation permutations that transform vertex u to v in the framework M*

```
FOR ALL: u ∈ {1 … V}
    FOR ALL: v ∈ {1 … V}
        C_M [u][v] ← ∅              (all congruence sets start empty in the VxV congruence matrix)

FOR ALL: r ∈ {R_1 … R_k}
    FOR ALL: u ∈ {1 … V}
        v ← r [u]                   (rotation r transforms vertex u to vertex v)
        C_M [u][v] ← C_M [u][v] ∪ {r}   (collect together all the permutations that transform u to v)
```

Algorithm 2 Constructing the congruent lookup tables.

1	{A, B, C}	{D, G, K}	{E, J, L}	{F, H, M}
2	{D, E, F}	{A, H, L}	{C, G, M}	{B, J, K}
3	{G, H, J}	{B, E, M}	{A, F, K}	{C, D, L}
4	{K, L, M}	{C, F, J}	{B, D, H}	{A, E, G}
C_{tet}	1	2	3	4

1	{A, B, C, D}	{E, J, N, S}	{F, L, Q, T}	{H, M, R, V}	{G, K, P, U}	{W, X, Y, Z}
2	{E, F, G, H}	{A, K, T, W}	{D, J, V, X}	{C, L, U, Y}	{B, M, S, Z}	{N, Q, R, S}
3	{J, K, L, M}	{B, F, P, X}	{A, H, N, Y}	{D, G, Q, Z}	{C, E, R, W}	{S, T, U, V}
4	{N, P, Q, R}	{C, L, U, Y}	{B, M, S, Z}	{A, K, T, W}	{D, J, V, X}	{E, F, G, H}
5	{S, T, U, V}	{D, G, Q, Z}	{C, E, R, W}	{B, F, P, X}	{A, H, N, Y}	{J, K, L, M}
6	{W, X, Y, Z}	{H, M, R, V}	{G, K, P, U}	{E, J, N, S}	{F, L, Q, T}	{A, B, C, D}
C_{oct}	1	2	3	4	5	6

Table 6 The congruence tables generated for the tetrahedron and octahedron framework models derived from the rotation permutations in Table 3 and Table 4 respectively by Algorithm 2.

Algorithm 2 builds the congruence matrix ***C*** for each framework model by inspecting each rotation permutation. If the model vertex at position ***u*** is moved to position ***v*** by the permutation, it is added to the set of permutations stored in the matrix element ***C [u] [v]***. At completion each element ***C [u] [v]*** contains the set of *all* the permutations that transform model vertex ***u*** to ***v*** (Table 6).

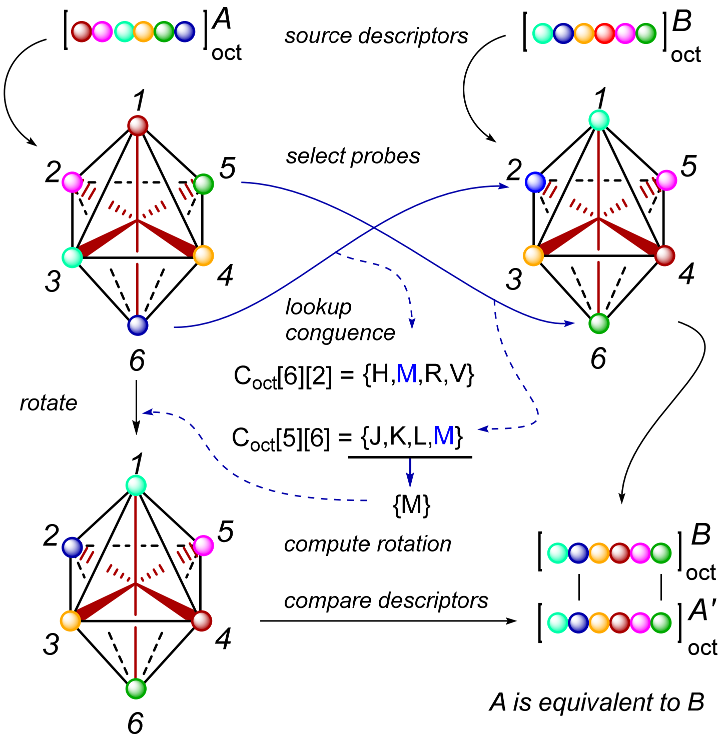


Figure 47 **Finding a congruent rotation permutation. Two non-coaxial probes are selected in descriptor A and located in descriptor B. The noted probe transpositions (6 →2, 5 → 6) are used to select two sets of congruent permutations from matrix C_{oct}. The intersection of the extracted sets {H,M,R,V} and {J,K,L,M} selects the single rotation (M) which is used to generate descriptor A'. A' is compared element by element to descriptor B to prove equivalence.**

The comparison of two stereo descriptors (**S, T**) is handled at run-time by the executive **MatchStereoDescriptors (S, T)** (Algorithm 3). This performs the basic steps presented at the beginning of this section, namely first testing for 'same-coset' membership by rotation and if necessary, by reflection and rotation for 'mirror-coset' membership. Each of these two coset membership testing phases are conducted by **InSameCoset (S, T)** (Algorithm 5) which selects the required congruence rotation and checks if the subsequent descriptor alignment leads to a match. The selection of congruence rotation is conducted by **SelectCongruenceRotation (Q, T)** (Algorithm 6) which seeks two corresponding probe atoms in both stereo descriptors and lookups up a congruence rotation that line up the two probe atoms. Each pair of corresponding probe atoms is used to lookup the group of rotations that can align the probe pair. Intersecting the permutation groups associated with both pairs of probes may yield a single rotation permutation that aligns both sets of probe atoms. The descriptor alignment steps are depicted graphically in Figure 47. If the corresponding probe atoms in the second descriptor are coaxial then the intersection of the permutation groups yields an empty set indicating the descriptors are not equivalent (Figure 48).

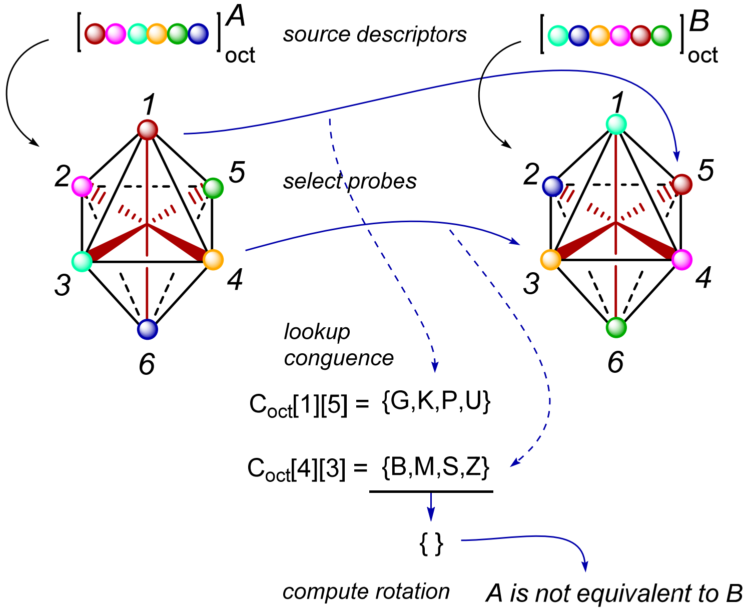


Figure 48 **Proving non-congruency. In this example two non-coaxial probes are selected in descriptor A and located in descriptor B. The probe transpositions are used to select two permutation sets from matrix C_{oct}. However the intersection of the sets is an empty set proving A and B cannot be brought into alignment and are thus not members of the same coset.**

MatchStereoDescriptors (S, T)

Inputs: *S and T are stereo descriptors vectors of dimension V mapped on the same framework model M. All the elements of S are members of T.*

Operation: *P is a stereo descriptor and the geometric reflection of stereo descriptor S. σ is the reflection permutation of M.*

Outputs: *r is one of 'same-coset', 'mirror-coset' or 'isomer-coset'*

```
IF: InSameCoset (S, T)
   r ⟵ same-coset                        (the descriptors are equivalent)
ELSE:
   P ← PermuteStereoDescriptor (S, σ)    (reflect the stereo descriptor )
   IF: InSameCoset (P, T)
      r ← mirror-coset                   (enantiomer relationship)
   ELSE:
      r ← isomer-coset                   (constitutional isomer or diastereomers relationship)
```

Algorithm 3 MatchStereoDescriptos determines the coset relationship between two stereo descriptors.

InSameCoset (S, T)

Inputs: *S and T are stereo descriptors vectors of dimension V mapped on the framework model M. All the elements of S are also members of T.*

Operation: *R is a set of permutations with either 0 or 1 member. P is a permuted stereo descriptor of S.*

Outputs: *m is true if S and T are members of the same coset*

```
R ← SelectCongruenceRotation (S, T)   (try to achieved congruence via two probe vertices)
IF: |R| ≠ 1                           (check if congruence was not achieved)
   m ← false
ELSE:
   r ← ∈ R                            (select the only congruence permutation)
   P ← PermuteStereoDescriptor (S, r) (bring two vertices into congruence)
   m ← true
  FOR ALL: v ∈ {1 … V} ∧ m
     m ← S [v] = T [v]                (for full congruence all vertices of S and T must match in order)
```

Algorithm 4 InSameCoset tests if two descriptors are in the same rotation coset.

PermuteStereoDescriptor (S, P)

Inputs: *S is a stereo descriptor vector of dimension V mapped on the geometric framework M. P is a rotation or reflection permutation of M.*

Outputs: *S is transformed into a new sequence order in T by permutation P.*

```
FOR ALL: $v \in \{1 \dots V\}$
  $T[v] \leftarrow S[P[v]]$
```

Algorithm 5 PermuteStereoDescriptors transforms a descriptor using a permutation operator.

SelectConguenceRotation (Q, T)

Inputs: *Q and T are stereo descriptor vectors of dimension V mapped on the framework model M. All the elements of Q are also members of T. Members 'H' and ':' are place holders for hydrogen and lone pairs respectively. The vector element A_M [v] is the set of coaxial vertices of vertex v in framework M. The matrix element C_M [u] [v] is the set of rotation permutations that transform vertex u to v in the framework model M.*

Operation: *r,s are selected probe vertices in Q; r',s' are the corresponding probe vertices in T*

Outputs: *R is the set of rotation permutations required to transform Q to T. R is empty if congruence cannot be achieved; otherwise R has one member that will bring at least two vertices in Q into congruence (alignment) with those in T.*

```
$r \leftarrow 0$, $s \leftarrow 0$, $r' \leftarrow 0$, $s' \leftarrow 0$        (set the probe vertices to null values)

FOR ALL: $v \in \{1 \dots V\} \wedge r = 0$
  IF: $Q[v] \neq$ 'H' $\wedge$ $Q[v] \neq$ ':'
    $r \leftarrow v$                    (find first probe in Q that is not hydrogen or a lone pair placeholder)

FOR ALL: $v \in \{r + 1 \dots V\} \wedge s = 0$
  IF: $Q[v] \neq$ 'H' $\wedge$ $Q[v] \neq$ ':' $\wedge$ $v \notin A_M[r]$
    $s \leftarrow v$                    (find second probe in Q that is not H, : or a coaxial vertex of r)

FOR ALL: $v \in \{1 \dots V\} \wedge r' = 0$
   IF: $Q[r] = T[v]$
    $r' \leftarrow v$                   (find the vertex in T corresponding to the first probe)

FOR ALL: $v \in \{1 \dots V\} \wedge s' = 0$
  IF: $Q[s] = T[v]$
    $s' \leftarrow v$                   (find the vertex in T corresponding to the second probe)

$R \leftarrow C_M[r][r'] \cap C_M[s][s']$        (lookup and select the congruence rotation)
```

Algorithm 6 SelectCongruenceRotation finds the rotation permutation that aligns a pair of stereo descriptors on two non-coaxial probe vertices.

Inspection of the **MatchStereoDescriptors** algorithm shows that it scales linearly with the number of vertices in the stereo descriptor. It is expected that the atoms for first and second probes will be selected from positions 1 and 2 in the vast majority of cases curtailing the need to search, with the exception of the occasions were implicit hydrogen or lone pairs occupy these positions. Careful design of the recognition patterns can be used to avoid placing implicit hydrogen or lone pairs in descriptor positions 1 and 2. However this cannot be avoided in some olefin patterns. The expectation is **MatchStereoDescriptors** will effectively behave as a constant time function built around a simple table lookup as the vast majority of detected stereocentres will contain a constant number of vertices (olefin, allene, biaryl atropisomer and tetrahedral geometries all have 4 vertices). The key feature of the algorithm is that it is applicable to any type of stereo descriptor that is supported by a framework model consisting of a set of rotation permutations.

Reaction Representation

Reaction indexing systems [157, 158] and reaction databases [159, 160] internally represent reactions as a set of educt and product molecules accompanied by a list of atom to atom correspondences between educt and product. This correspondence list is commonly represented as an associated atom-atom map. The atom-atom map serves to precisely define the changes that have occurred to transform the educts into the products according to the known or presumed mechanism of the reaction. This mapping information is exploited in reaction search systems by allowing the user to target particular reaction types or specific transformations. This is done by creating educt or product molecule queries in which specific bonds can be marked as 'made', 'broken' or 'changed'.

Atom-atom maps are created either manually as part of the reaction curation process carried out by expert human chemists or they are created after submission using auto-mapping algorithms. [158] Auto-mapping algorithms are generally based on finding a maximal common subgraph between educts and products to first identify non reacting portions [161, 162] followed by a search amongst the remaining atoms for atom-atom assignments that result in a minimum set of bond changes. [163] Consequently automated map generators are highly susceptible to erroneous assignments as they are based on the mistaken assumption that all reactions involve the smallest possible number of bond changes (Figure 49 and Figure 50).

Failures often occur when reactions are drawn in a casual unbalanced form (Figure 50) with reagent atoms only introduced into the product and absent as an atom-atom mapped educt, or with leaving groups omitted as an atom-atom mapped by-product. In these situations using a maximum common subgraph algorithm to locate non reacting substructures may generate a confused solution.

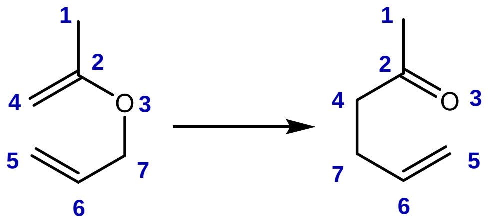


Figure 49 A typical reaction mapping error produced by the assumption that minimum bond changes occur in a reaction.

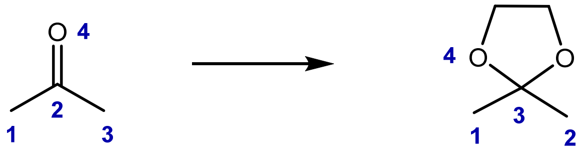


Figure 50 A typical reaction mapping error produced by assuming a maximum common subgraph match will identify all non-reacting atoms.

The balance between the two approaches is a trust that a human curator can identify atom mappings reliably using extensive mechanistic knowledge versus the use of computer algorithms with inadequate supporting knowledge of reaction mechanisms. Thus the indexing policy of the reaction database producer can have a major impact on the overall utility of a particular reaction database as a source of reliable information for a synthesis planning program.

Representing Reactions as Hyperstructures

Vladutz observed that the set of created, broken and modified bonds occurring in a reaction and the relationships between those bonds had the potential to be used as a systematic means for uniquely identifying or naming reactions when coded as a hyperstructure.[164] A similar approach was independently devised by Fujita for the purposes of deriving a canonical coding scheme for reactions using an equivalent diagram termed an imaginary transition state (ITS). [165]

Vladutz outlined a process in which the graphs of educts and products of a reaction are superimposed to form a hyperstructure (Figure 51) according to the educt-product atom-atom mappings with an appropriate relabeling of the bonds using a new set of differential bond symbols (Figure 51, C). These bond symbols are designed to indicate forming bonds (or more generally bonds with increasing order) and breaking bonds (or more generally bonds with decreasing order). Removal of non-reacting bonds derives a subgraph that is characteristic of the particular reaction and can be used for indexing purposes. Reactions having identical characteristic reaction cores (CRC) can be clustered as examples of a common chemical transformation. However this approach is limited in that many reaction types which may differ by mechanism, activating groups or types of reagent may share the same CRC subgraph meaning that any cluster of reaction examples may contain a subset of inappropriate examples for the desired transformation.

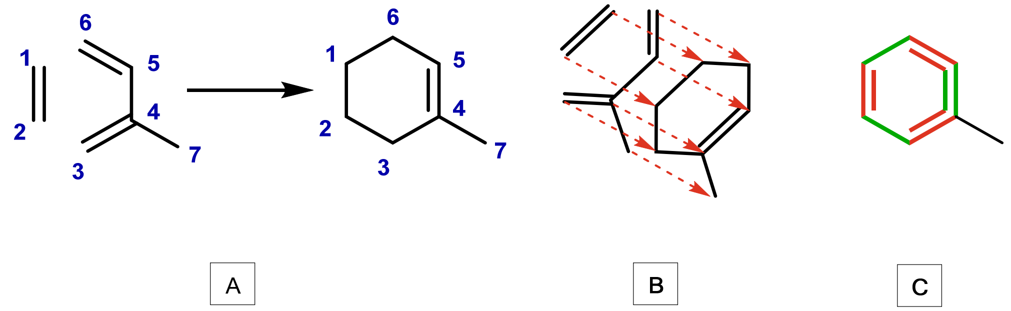


Figure 51 **The processes of superimposing atom-atom mapped educts and product graphs (A) via superimposition (B) to generate a reaction hyperstructure graph labelled with differential bond symbols (C).**

Applications of Hyperstructures

The literature concerning the construction and use of hyperstructures is sparse. Vladutz and Gould proposed that dense hyperstructures could be used to provide a compact storage mechanism for very large molecule and reaction databases. [166] In addition it was suggested that increases in search efficiency would be possible as each single substructure match within a hyperstructure would retrieve multiple molecules.

Hyperstructures are attractive data structures for certain computer processing purposes, particularly for representing and searching reactions. Reactions coded as hyperstructures are more compact as the total numbers of atoms and bonds can be reduced by up to half of that occupied by individual educt and products graphs. This has an important impact on the performance of algorithms such as subgraph isomorphism (substructure search). For example a regular reaction substructure search would require the educt and product portions to be applied independently to the educt and product graphs to identify a match. This must be followed by a post check to determine that the correct atom-atom correspondences are present. If the atom-atom correspondence test fails under a particular match, then further alternative solutions must be sought and post checked until a match is eventually found or all matches are exhausted. In contrast a search over a single hyperstructure reduces the number of individual atom atoms visited in the backtrack search and avoids the need to iterate alternative solutions as the need for the atom-atom post check is eliminated.

It is common practice to suppress hydrogen in both chemical diagrams (for readability) and in

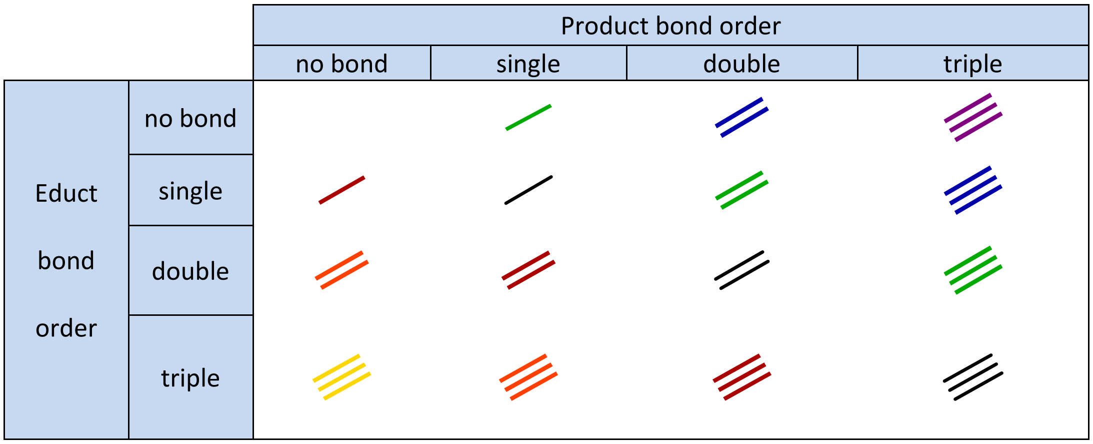


Figure 52 Colour coded differential bond types used to label bonds in reaction hyperstructure diagrams in this thesis.

		Product H Counts			
		-A	-AH	-AH$_2$	-AH$_3$
Educt H Counts	-A	—A	—A[+H]	—A[+H$_2$]	—A[+H$_3$]
	-AH	—AH[-H]	—AH	—AH[+H]	—AH[+H$_2$]
	-AH$_2$	—AH$_2$[-H$_2$]	—AH$_2$[-H]	—AH$_2$	—AH$_2$[+H]
	-AH$_3$	—AH$_3$[-H$_3$]	—AH$_3$[-H$_2$]	—AH$_3$[-H]	—AH$_3$

Figure 53 Differential hydrogen symbols used in reaction hyperstructure diagrams.

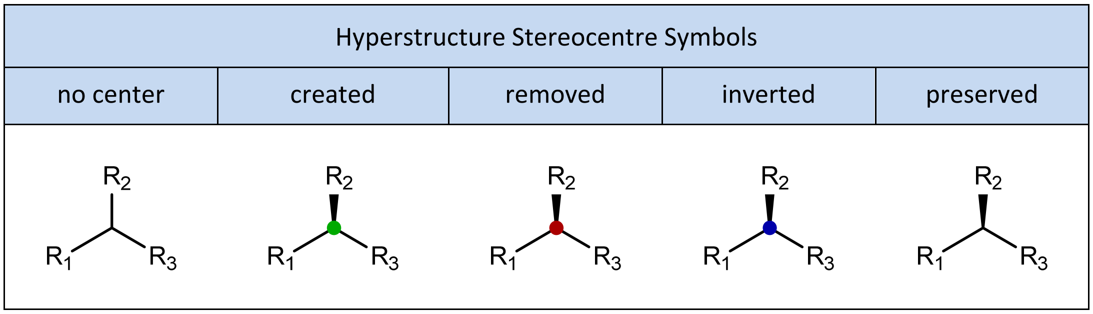


Figure 54 A selection of differential stereo centres indicators used in reaction hyperstructure diagrams.

corresponding computer representations (for processing and storage efficiency). To cater for this, additional hyperstructure attributes are introduced to specify changes in the counts of hydrogen attached to their parent atoms (Figure 53).

Attributes indicating how and where stereocentres are changed in a reaction are introduced for the purpose of aiding the hyperstructure discussion in this section (Figure 54). A separate 'chemist friendly' graphical language has been devised to construct stereo selective reaction rules where perception algorithms determine the locations and nature of stereochemical changes. This latter approach is discussed in a later chapter.

Examples Reactions Represented as Hyperstructures

The application of reaction hyperstructures to characterise reaction types is demonstrated using a number of examples. Figure 55 demonstrates the application of the hyperstructure bond and stereocentre labelling in a Cope rearrangement reaction. The hyperstructure of the reaction (B) is formed by the direct superimposition of the mapped educts and products (A). The bond changes induced by the reaction are indicated[n] by substituting the educt and product bonds with the differential bond types listed in Figure 52 (*vide infra*). The non-reacting bonds are identified using regular bond types (drawn in black). The removal of the non-reacting bonds generates a *hyper substructure* that is characteristic of all Cope rearrangements (C). This is the characteristic reaction core (CRC) [164 , 167] which is identical for all Cope rearrangement reactions. The CRC is a useful tool that can be used to rapidly retrieve multiple examples of the desired reaction from a suitable reaction database[o]. The CRC subgraph can also be augmented by a set of stereochemical characteristic reaction core (SCRC) subgraphs (E) that can be used to narrow the search for example reactions containing identical patterns of stereochemical change. For example the SCRC subgraph D codes the stereochemical transpositions that match the specific substrates used in reaction A. This screening approach is utilised to select relevant supporting example reactions for the hierarchies of stereo selective reactions rules developed for ARChem (*see chapter 5*).

[n] Limitations with ChemDraw mean that the desired annotations are absent and replaced with a colour scheme.

[o] Any reaction database that supports atom-atom maps

The oxy-Cope reaction (Figure 57) illustrates a case were an individual reaction bond change is masked during the course of the overall reaction such that it is eliminated from the CRC. In the

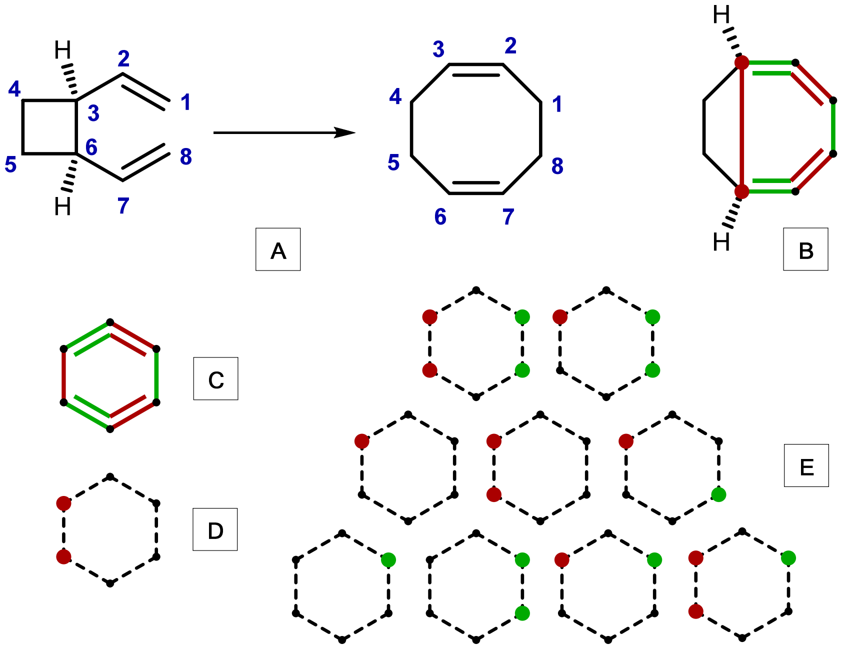


Figure 55 An example of the Cope rearrangement reaction A: with its corresponding reaction hyperstructure B; the extracted CRC substructure C; and the extracted SCRC substructure D. Depending on the locations of substituents any SCRC substructure from the set E may be combined with the CRC.

oxy-Cope rearrangement the [3,3] sigmatropic rearrangement is followed by an enol-keto tautomerisation to yield the more stable ketone product. It can be observed that the CRC has eliminated one of the [3,3] sigmatropic reaction bonds. The enol-keto proton shift step is coded with appropriate hydrogen count differences. This highlights a specific behaviour of the CRC in that it does not code the *sequences* of bond changes but only the overall outcome. The CRC remains characteristic as it is shared by all oxy-Cope rearrangements.

Another example illustrates that independent representations of the same reaction transformation may yield different hyperstructures with different CRCs. The inconsistent nature in which reactions are curated and entered into databases allows for some reacting moieties to be treated as either educt structures or cited as reagents with no accompanying structure.

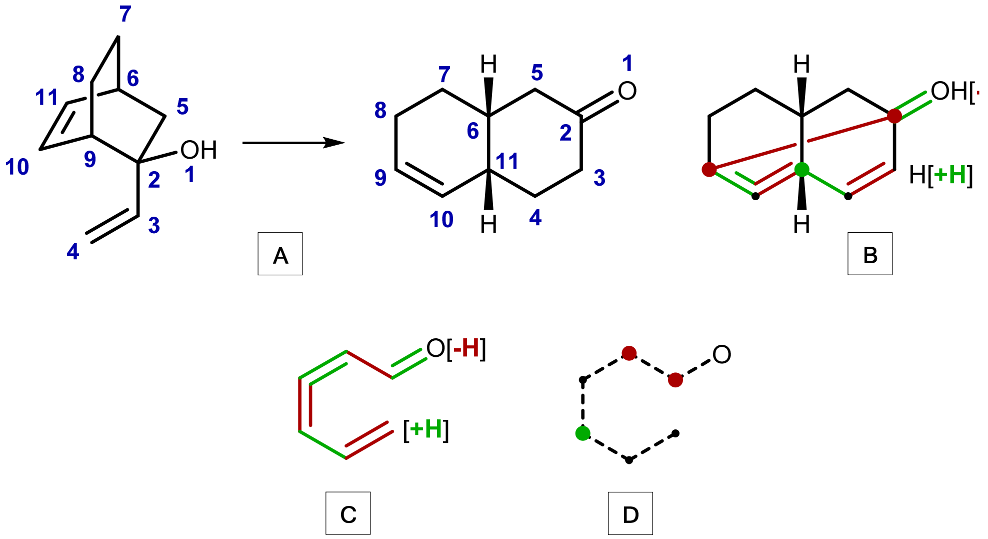


Figure 57 **An example of the Oxy-Cope rearrangement reaction A: with its corresponding reaction hyperstructure B; the extracted CRC substructure including hydrogen changes C; and the extracted SCRC substructure D specific to the example reaction A.**

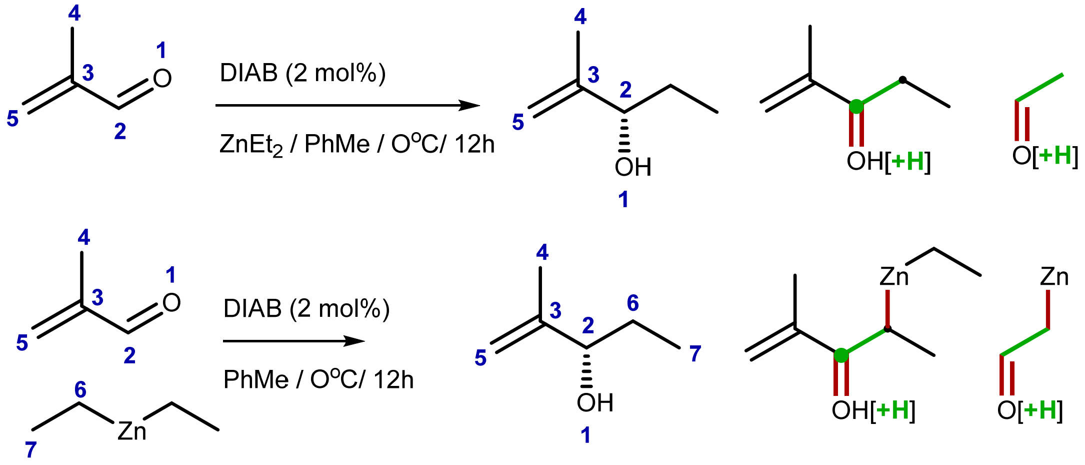


Figure 56 **Alternative representations of the same reaction. The alkylating reagent may or may not be represented as a direct atom mapped participant which leads to different CRCs.**

Figure 56 illustrates two forms of the same enantioselective alkylation reaction, with and without explicitly representing the diethyl zinc reagent. A comparison of the reaction hyperstructures and CRCs clearly shows a difference in that one hyperstructure graph includes a breaking carbon-zinc bond, while the other graph does not. This suggests that either a *normalisation* process is needed in which recognisable nucleofuges or electrofuges are removed to yield a common CRC or that a particular reaction transformation is represented by a *set* of CRCs varying only by the specific nature of the observed leaving group or nucleophilic additive. The current approach employed by ARChem is to normalise to a single representation of a CRC containing a generic nucleofuge (NF) or electrofuge (EF) using built in substitution rules. [40]

The reaction cores can be used to screen a database during search by extracting and coding the CRC or SCRC of the query or rule hyperstructure and limiting the detailed search for matches to the subset of database records with identical CRC and SCRC codes. The perception of the stereochemical changes necessary to code the SCRC and the method of coding both the CRC and SCRC subgraphs is described in detail in the following sections.

The Perception of Stereochemical Changes in Reactions

The REACCS system supported a constrained form of stereochemical reaction perception derived via the reaction atom-atom map. This was limited to the detection of substitution with retention or inversion at carbon sp^3 centres. [158] This was developed to support queries so users could constrain the type of stereochemical modification occurring in an educt or product search. The precise details of the algorithm were not published.

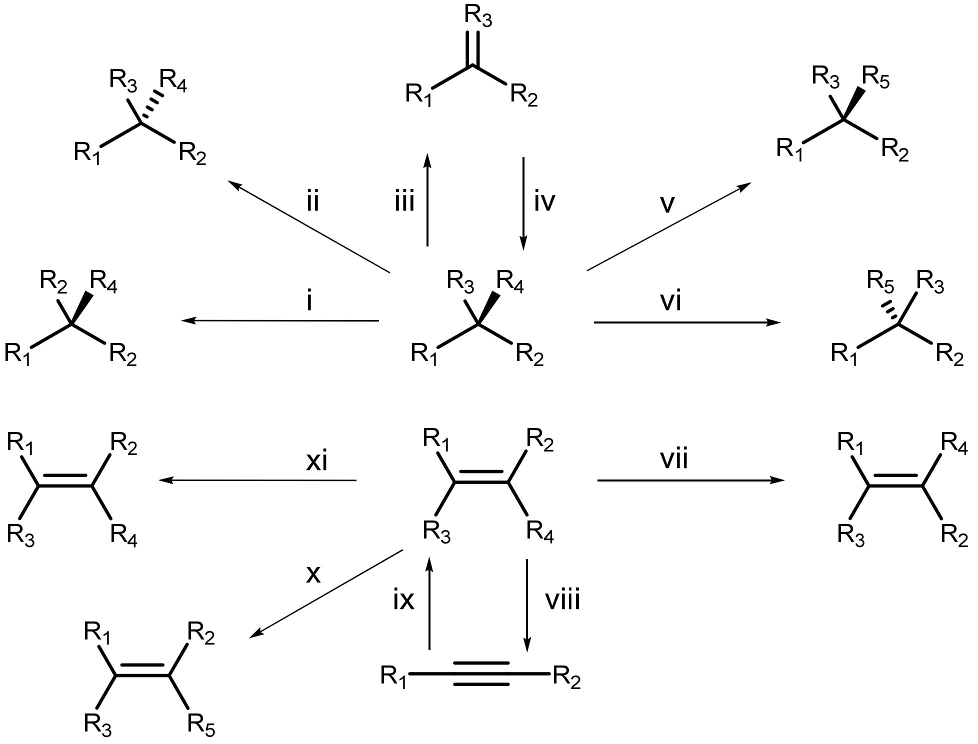


Figure 58 **Examples of stereo transformations recognised by the reaction perception module: (i)(xi) preserved; (iii)(viii) cleared; (iv)(ix) created; (ii) epimerised; (vii) isomerised; (v)(x) substituted with retention; (vi) substituted with inversion.**

The operation of a novel algorithm for the perception of stereochemical changes begins by performing stereocentre perception on the educt and product aspects of the reaction hyperstructure. Reaction changes are then discovered by pairing up corresponding educt and product stereo descriptors directly from the reaction hyperstructure. The classifications of the stereochemical changes detected by the algorithm are listed and described in Table 7. Examples of the generalised representations of these transformation classes at tetrahedral carbon centres and in alkenes are shown in Figure 58. The algorithm is applicable to all types of stereogenic centres and stereo axes and can thus be applied to chiral allenes and coordination complexes.

The algorithm is divided into two parts. A top level executive (Algorithm 7) processes the classification of atom centred stereochemical changes (sp^3 centres, allenes and metal centred coordination complexes, lines 1 - 4) and then bond centred stereochemical changes (alkenes, biaryl atropisomers, imines, oximes etc., lines 5 - 8). The sets *SA* and *SB* are those atoms and bonds in the hyperstructure where the stereo perception executive has located stereogenic centres. Functions are called to retrieve the identifiers of the corresponding stereo descriptors (lines 2 – 3 and 6 – 7). Both atom and bond based stereo descriptors are now classified by a common procedure (lines 4 and 8) detailed in Algorithm 8 (classifyStereoChanges).

Stereocentre Modification Sets			
Educt sets	Description	Product Sets	Description
eductPreservedStereos	*All educt stereogenic centres not changed by the reaction*	productPreservedStereos	*All product stereogenic centres not changed by the reaction*
eductDestroyedStereos	*All educt stereogenic centres cleared in the reaction*	productCreatedStereos	*All product stereogenic centres created in the reaction*
eductRetainedStereos	*All educt stereogenic centres with a ligand substitution with retention of configuration*	productRetainedStereos	*All product stereogenic centres with a ligand substitution with retention of configuration*
eductInvertedStereos	*All educt stereogenic centres with a ligand substitution with inversion of configuration*	productInvertedStereos	*All product stereogenic centres with a ligand substitution with inversion of configuration*
eductSubstitutedStereos	*All educt stereogenic centres with a ligand substitution with neither inversion or retention of configuration*	productSubstitutedStereos	*All product stereogenic centres with a ligand substitution with neither inversion or retention of configuration*
eductEpimerisedStereos	*All non-substituted educt stereogenic centres with inversion of configuration*	productEpimerisatedStereos	*All non-substituted product stereogenic centres with inversion of configuration*
eductIsomerisedStereos	*All non-substituted educt stereogenic centres with a non-inverted change of configuration*	productIsomerisedStereos	*All non-substituted product stereogenic centres with a non-inverted change of configuration*
eductRemovedStereos	*All non-destroyed educt stereogenic centres removed within leaving groups etc*	productAddedStereos	*All product stereogenic centres introduced from reagents etc*

Table 7 Stereochemical changes. The set names and definitions used by the reaction stereo perception module.

The classification procedure first checks to determine if either one of the educt or product stereo descriptors (ES, PS) is absent (lines 1 and 6). If this is the case we have detected a stereocentre that has no correspondent. At this point a distinction has to be made between centres that have genuinely created or destroyed or have been introduced via a reagent or lost as part of a leaving group. To cater for this a further check is made by comparing the bonds contained within the stereocentre against the *addedBonds* or *removedBonds* sets of the reaction hyperstructure. If all the bonds of the stereocentre are contained within the added or removed bonds sets the classification is set to the special case of 'added' or 'removed' stereocentres, otherwise the classification chooses 'created' or 'destroyed' stereocentres (lines 2 – 5 and 7 - 10).

PerceiveStereoChanges (SA, SB)

Inputs: SA *is the set of perceived stereogenic atoms in both educts and products; SB is the set of perceived stereogenic bonds in both educts and products*

Operation: *eductAtomToStereo, productAtomToStereo, eductBondToStereo, productBondToStereo are functions that retrieve associated stereo descriptors indexed by parent atoms or bonds. ES is a educt stereo descriptor; PS is a product stereodescriptor.*

Outputs: *the sets listed in* Table 7 *contain the perceived stereochemical changes*

```
1.  FOR ALL: $A \in SA$
2.      ES ← eductAtomToStereo (A)
3.      PS ← productAtomToStereo (A)
4.      classifyStereoChanges (ES, PS)

5.  FOR ALL: $B \in SB$
6.      ES ← eductBondToStereo (B)
7.      PS ← productBondToStereo (B)
8.      classifyStereoChanges (ES, PS)
```

Algorithm 7 The stereocentre transformation perception executive. Stereo descriptors are retrieved from both atom and bond centres and passed to a common classification algorithm.

The remaining classifications involve a pair of correspondent educt and product stereocentres. A pairwise classification procedure is now conducted in two further phases. The first phase discovers relationships that do not involve ligand substitution and the second phase discovers relationships dependent on a single ligand substitution. Higher order substitutions are not sought.

The two stereo descriptors are directly compared (line 11) using the MatchStereoDescriptors comparison function (*vide infra*). A direct comparison is possible because corresponding educt and product atoms derived from a reaction hyperstructure representation share the same atom index. Four possible outcomes are possible indicating the descriptors belong to: the 'same-coset' relationship meaning the stereocentre is *preserved* in the reaction (line 13); the 'mirror-coset' relationship meaning the stereocentre has been *epimerised* (line 15); an 'isomeric-coset' relationship meaning centre has been *isomerised* (line 17); or the ‘no-coset’ relationship meaning at least one ligand substitution has occurred.

If a ligand substitution is detected a search is made within the two stereo descriptors for all non-

ClassifyStereoChanges (ES, PS)

Inputs: *ES and PS are paired stereo descriptors from before and after a reaction change. Either can be absent if a stereocentre was created or removed.*

Operation: *SB is the set of bonds contained by the atoms of a stereo descriptor. ES’ and PS’ are ligand substituted stereo descriptors. The sets addedBonds and removedBonds are extracted from the hyperstructure representation of the reaction.*

Outputs: *the sets listed in* Table 7 *contain the perceived stereochemical changes*

```
1.      IF: ES is absent
2.          SB ← productStereoToBonds (PS)
3.          IF: SB ⊆  addedBonds
4.              productAddedStereo ← ∪ {PS}
            ELSE:
5.              productCreatedStereo ← ∪ {PS}
6.      ELSE IF: PS is absent
7.          SB ← eductStereoToBonds (ES)
8.          IF: SB ⊆ removedBonds
9.              eductRemovedStereo ← ∪ {ES}
            ELSE:
10.             eductDestroyedStereo ← ∪ {ES}
        ELSE:
11.         result ← MatchStereoDescriptors (ES, PS)
12.         IF result = ‘same_coset’
              ( the two stereocentres can be superimposed )
13.           eductPreservedStereos  ← ∪ {ES}, productPreservedStereos ← ∪ {PS}
14.         ELSE IF: result = ‘mirror_coset’
              ( there has been an inversion with no ligand substitution )
15.           eductEpimerisedStereos ←  ∪ {ES}, productEpimerisedStereos ←  ∪ {PS}
16.         ELSE IF: result = ‘isomer_coset’
              ( a geometric isomerization has occurred with no ligand substitution )
17.           eductIsomerisedStereos ←  ∪ {ES}, productIsomerisedStereos ← ∪ {PS}
            ELSE:   ( at least one ligand has been substituted )
               ( replace all substituted ligands with the same dummy code)
18.            ES’, PS’  ←  SubstituteLigands (ES, ES)
19.            result ←  MatchStereoDescriptors (ES’, PS’)
20.            IF: result = ‘same_coset’
                  ( there has been retention with ligand substitution )
21.               eductRetainedStereos ← ∪ {ES’}, productRetainedStereos  ← ∪ {PS’}
22.            ELSE IF: result = ‘mirror_coset’
                  ( there has been an inversion with ligand substitution )
23.               eductInvertedStereos ← ∪ {ES’}, productInvertedStereos ← ∪ {PS’}
24.            ELSE IF: result = ‘isomer_coset’ then
                  ( an unclassified isomerization has occurred with ligand substitution )
25.               eductSubstitutedStereos ← ∪ {ES’},  productSubstitutedStereos ← ∪ {PS’}
               ELSE:
                  ( more than one ligand was substituted. This result is not classified )
```

Algorithm 8 Perception of stereochemical changes at a specific stereocentre.

corresponding ligands and these are substituted with a common dummy ligand (by using a reserved atom index, lines 18). The descriptor comparison is now repeated with the substituted descriptors (line 19) and the result inspected. The outcomes are: the 'same-coset' relationship meaning the stereocentre was *substituted with retention* (line 21); the 'mirror-coset' relationship means *substitution with inversion* has occurred (line 23); the 'isomer-coset' relationship results in an unclassified *substitution*. The classification process terminates at this point. If it is shown that after a single dummy substitution in the stereo descriptors that there is still no coset relationship then it can be concluded that at least a double substitution has occurred. This state is not catalogued. The results are stored in the named sets listed in Table 7 (*vide supra*).

Characteristic Reaction Core Codes

Coding the characteristic reaction core (CRC) code of a reaction hyperstructure graph (*vide infra*) as a simple numerical value provides an effective means to rapidly search for all reactions of the same core type. This section describes an approach used to create CRC and SCRC hash codes and their use as an indexing tool used to select subsets of reactions examples from a database prior to substructure matching. [69] These reaction examples are used to support the scope and limitations of specific stereo selective transformation rules (*vide supra*).

At least three classes of algorithms for chemical structure canonicalization or coding have been reported. These are the Morgan algorithm[66], the Principal Eigenvector (PE) algorithm[168], and the Smallest Binary Code (SBC) algorithm[169].

Each of these classes of algorithm exhibits practical difficulties in either implementation or operation. Morgan and PE can suffer from oscillation or poor atom environment partitioning. PE can also exhibit poor convergence or suffer local minima. The SBC has been shown to have imperfections [168] as it suffers local minima problems only solved by performing complex 2 way and 3 way cyclic exchanges on rows and columns in the graph adjacency matrix. The PE and SBC algorithms are not widely used and there is little information in the literature revealing practical implementations. The Morgan algorithm on the other hand has been used extensively in many applications. [69, 170, 68, 171]

The Morgan Algorithm

Morgan describes a coding method [66] using *only* non-hydrogen connectivity in the coding scheme and in this simple form the algorithm did not differentiate realistic chemical graphs well. This situation was readily improved by others, incorporating invariant atom, bond and ring properties into the coding scheme. [170] The next advancement was the development of methods to code stereochemistry in various canonicalisation schemes.[68, 143, 150] The Morgan algorithm and its derivatives are also used to discover topological symmetry using partitions of the final atom codes to assign symmetry equivalence. [78 , 85]

A detailed analysis of the original form of the Morgan algorithm by Figueras showed that the following matrix power equation succinctly describes the algorithm's operation: [70]

$$v_r = A^r v_0$$

$\mathbf{V_0}$ is the vector of initial code values assigned to the *n* atoms of the structure and $\mathbf{A}^r$ is the r$^{\text{th}}$ power of the adjacency matrix **A**, the latter being an *n* x *n* matrix where element a_{ij} = 1 if atoms *i* and *j* are connected, otherwise a_{ij} = 0. The equation is separable in two parts; the first part depends only on the atom to atom adjacency ($\boldsymbol{A^r}$); the second part only depends on assigning an initial code value to each atom ($\boldsymbol{V_0}$). This is useful behaviour as it allows the algorithm to be readily adapted to code different types of graph and different features within the graph. For example the adjacency matrix can be constrained to the reaction core in a hyperstructure while the initial codes can incorporate atom types and bond orders amongst others. This flexibility allows distinct indexing codes to be generated that represent specific patterns of bonds changes or patterns of stereocentre changes.

A Flexible Morgan Algorithm

An improved version of the Morgan algorithm was developed to provide a flexible means to code CRC and SCRC subgraphs for the purposes of indexing and retrieving reaction examples.

The original Morgan algorithm counted the number of different code values (k) in $\mathbf{V_r}$ at the end of each pass of the coding loop and when this count ceased to change the algorithm terminated. In certain cases oscillatory behaviour had been observed [70] where the count k increases and decreases on alternate steps. To overcome this problem a novel termination method was developed that traces the origin of the information content of each atom code at each step of the algorithm. When it is known that an atom has received information from every other connected atom via all connecting bonds, the atom can be removed from the coding process. When all atoms have been removed, the coding process is complete and the algorithm terminates. This technique avoids the need to monitor changes in the number of different atom codes and avoids the known oscillation problem.

An implementation of this tracking technique associates an '*InfoSet'* with each atom to monitor the origin of received information. The *InfoSet* is empty when the algorithm is first entered. During the coding phase of the algorithm each *InfoSet* is updated to include the bonds immediately attached to the atom and the bonds stored in the *InfoSet* of each adjacent atom. At each step the membership of an atom's *InfoSet* consists of the bonds that have been used to route information to the atom code. When an *InfoSet* ceases to change, the coding of the associated atom is complete as all accessible bonds have been traversed. An advantage of this alternative approach is that it guarantees that *all* atoms in the graph have a code value containing information derived from *all other* connected atoms. At termination the information content of any one atom code is complete for the connected graph and can be used as a representative of the whole graph.

Algorithm 9 presents the improved Morgan algorithm. The algorithm operates in three phases; an initialisation phase where the $\mathbf{V_0}$ code vector and the auxiliary *InfoSet* are set up (lines 1-5); a coding loop (line 5 – 16) where new code values for each atom in $\mathbf{V_r}$ are calculated and the information sources are tracked until each atom is completely coded; and a termination phase (line 17) where a final hash key is selected or computed from the values stored in the final $\mathbf{V_r}$ code vector.

At any point in the algorithm only the code values at iterations *k* and *k+1* are needed, so the storage space can be reduced from *n* vectors to just 2 vectors which are swapped with each other after each traversal of the outer loop is completed (lines 15-16).

Coding Reaction Hash Keys

The algorithm design incorporates functions that can be overridden so that different coding strategies can be employed for a variety of purposes such as coding a reaction CRC or SCRC.

The function 'codedAtoms' constrains the coded region of the graph by selecting the subset of atoms to be coded. For example selecting all atoms except explicitly drawn hydrogen is appropriate for hash coding a complete molecule (for starting material lookup). The generation of CRC or SCRC hash code requires that the algorithm is constrained to the atoms of the CRC or SCRC subgraph. The function 'atomToCodedAtoms' restricts access to just the adjacent atoms that are to be coded.

The function 'calcInitialCode' performs the initialisation of the $\mathbf{V_0}$ vector with seed codes. The initial seed code can be varied by choosing from different sets of atom and bond properties, including for example atom type, charge values, hydrogen counts and counts of attached bond orders. A CRC seed code is derived from combining the counts of the different differential bond types (Figure 52) connected to an atom together with the atom type. A SCRC seed code is derived by assigning an index to each of the modified stereocentre sets (Table 7) and using this index when the atom or connected bonds are present in a particular set.

The function 'calcNextCode' is called for each element of the coding vector $\boldsymbol{V_r}$ during each iteration r of the main coding loop. This function allows the Morgan coding formula $\boldsymbol{V_r} =.\boldsymbol{A}^r\boldsymbol{V_0}$ to be replaced with alternatives and is generally implemented according to Equation 1. The original Morgan function simply summed the codes of adjacent atoms with the previous atom code to derive a new atom code. This was found to be a problem in small CRC subgraphs due to a propensity to produce hash collisions where different CRCs were assigned the same final code. The cube term in the replacement formula was selected to significantly reduce the chance of code collisions at the expense of a slightly increased computational effort.

$$v_{i,r+1} = 3v_{i,r} + \sum (a_{i,j} v_{j,r})^3$$

Equation 1 A hashing formula designed to avoid hash collisions when coding small molecule and reaction graphs. $v_{i,r}$ is the code assigned to atom *i* at iteration *r; $a_{i,j}$* is the adjacency of atoms *i* and *j* (values 0 or 1).

The function 'calcFinalCode' is used to pick the smallest or largest atom code from $\boldsymbol{V_r}$ as the hash key for the whole molecule or CRC code. An alternative strategy sums the terms of $\boldsymbol{V_r}$ to generate a hash code.

The computer word size chosen for the intermediate and final hash codes is a 64 bit unsigned integer type. This yields a hash space for the CRC and SCRC codes of $2^{64} – 1$, sufficient space for representing all possible reaction types. The tactic of using an excessively large hash space ensures an extremely low risk of CRC hash key collisions.

Inputs: *The molecule structure or reaction hyperstructure is represented by graph G.*

Operation: *morgan_codes and morgan_codes' are vectors of numerical codes indexed by an atom in the graph G. morgan_codes stores the codes from the prior iteration. morgan_codes' stores the new values computed in the current iteration.*
info_sets and info_sets' are vectors of bond sets used to trace the information sources in the prior and current iterations.
codedAtoms(G) is a function returning a set of codeable atoms in graph G.
atomToCodedAtoms (G, atom) and atomToCodedbonds (G, atom) are functions returning the set of codeable atoms or codeable bonds adjacent to atom in graph G.
calcInitialCode is a function that computes an initial code for atom.
calcNextCode is a function computes a new code from an atom code and its neighbours codes. calcFinalCode is a function that selects or computes a final hash code from the code vector.

Outputs: *The function returns a numerical hash code value characteristic of the graph G*

```
coded_atoms ← codedAtoms (G)

FOR ALL: atom ∈ coded_atoms
    morgan_codes [atom] ← calcInitialCode (atom)
    info_sets [atom] ← ∅

atoms_left ← coded_atoms
WHILE: atoms_left ≠ ∅

   FOR ALL: atom ∈ atoms_left
      alpha_atoms ← atomToCodedAtoms (G, atom)

      info_sets' [atom] ← info_sets [atom] ∪ atomToCodedBonds (atom)
      FOR ALL: alpha_atom ∈ alpha_atoms
           info_sets' [atom] ← info_sets' [atom] ∪ info_sets [alpha_atom]

      morgan_codes' [atom] ← calcNextCode (atom, alpha_atoms, morgan_codes)

      IF: info_sets' [atom] = info_sets [atom]
           atoms_left ← atoms_left – { atom }

   morgan_codes ← morgan_codes'
   info_sets ← info_sets'

RETURN: calcFinalCode (coded_atoms, morgan_codes)
```

Algorithm 9 A reliable version of the Morgan algorithm used to code molecule structures, reaction hyperstructures and CRCs. The implementation of the underlined functions can be varied so that algorithm can be adapted to different applications.

Results and Discussion

Stereochemical Recognition Patterns

A total of 47 stereo recognition patterns were created to recognise a variety stereocentre types in Natta style diagrams. It is anticipated that more will be added when new drawing situations are discovered. The patterns where written in an XML notation and transformed into C++ code using a custom XSLT transformation script. The generated code was compiled into the unit test programs ‘*test-stereo-finder*’ and ‘*test-matcher*’ to validate that the design worked. The patterns covered the following configurations: tetrahedron (12), olefin-like (6), allene-like (12); bi-aryl atropisomers (4); square planar (3); trigonal bipyramid (2); square pyramid (6); and octahedron (3). The test programs were provided with unit tests to cover all the created patterns to validate pattern correctness. Exemplars are presented for ‘tetrahedron’ (Figure 60) and ‘olefin’ (Figure 64). The test results for tetrahedral, olefin, allene and atropisomer patterns are presented. The remaining patterns are validated but not reported here.

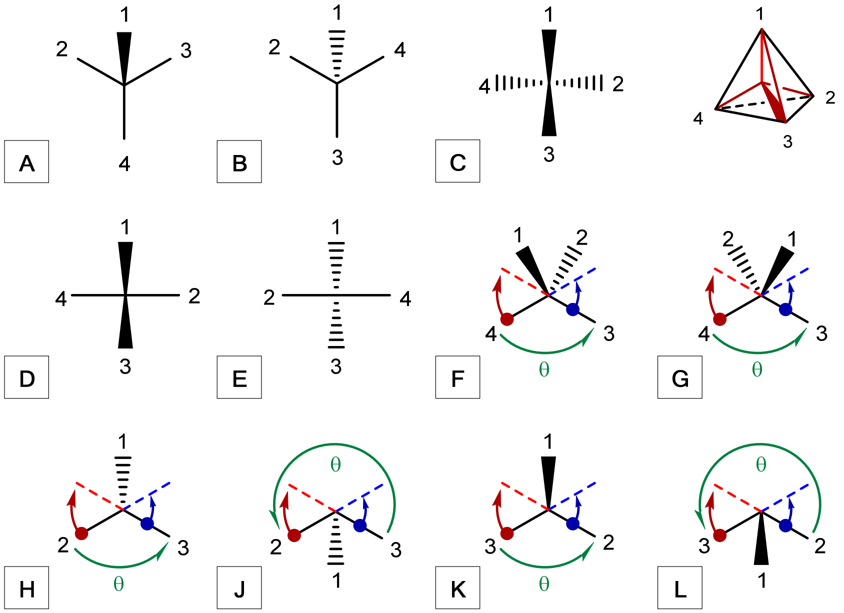


Figure 59 **Tetrahedral stereogenic centre recognition motifs. The numbered atoms are the corresponding positions in the generated ’tet’ stereo descriptors as defined in the reference framework. The red and blue arrows and dashed lines mark the minimum and maximum extents of the ligand angle limits.**

Figure 59 presents the set of substructure motifs used as the basis for creating each 'tetrahedron' pattern. The set of motifs were developed from the IUPAC structure drawing guidelines [68] and the manual inspection of sample supplier catalogue databases. Motifs A and B cover the most common method of drawing stereogenic centres with four ligands and one stereodefining bond. Motifs D to G cover those centres with four ligands and two stereodefining bonds. The need for both motifs F and G is due to 'up' and 'down' stereodefining bonds occurring in two different anti-clockwise orderings. Motifs H to L cover three ligands and one stereodefining bond. The missing fourth ligand is interpreted as implied hydrogen or lone pair depending on the nature of the central atom. The position of the implied hydrogen or lone pair is coded in the pattern as occurring in the larger of the two angles subtended by the in-plane bonds. The out-of-plane bond can occur between either the smaller or larger angles of the in-plane bonds and this gives rise for the need of the motif pairs H, J and K, L. Motif C covers the rare situation that all four ligands are drawn using stereodefining bonds.

The conversion of the motif H (Figure 59) into a corresponding recognition pattern ‘*tetrahedron7a*’ is used as an for the tetrahedral class (Figure 60). The pattern is specifically designed to cover at a minimum both tetrahedral carbon drawn with 3 ligands and in particular sulphoxides drawn in either an ylid or multiple bond form (Figure 61) hence ‘*D*’ is set to 3 and ‘*H/LP*’ to 1. To cater for oxaziridines, phosphines and other possibilities, the atom types ‘Z’ was left unrestricted. To improve performance the ‘*down*’ bond was cited before the in-plane bonds

pattern “tetrahedron7a”

	orbit	type	Aromatic	Z	D	implicit H / LP	bond order	bond direction	angle limits	EV	IV
[1]	0	Atom	no	-	3	1				-	4
[2]	1	Bond	no				-	Down	-		
[3]	2	Atom	-	-	-	-				1	-
[4]	1	Bond	no				-	None	-		
[5]	2	Atom	-	-	-	-				2	-
[6]	1	Bond	no				-	None	5 < θ < 175 from bond [4]		
[7]	2	Atom	-	-	-	-				3	-

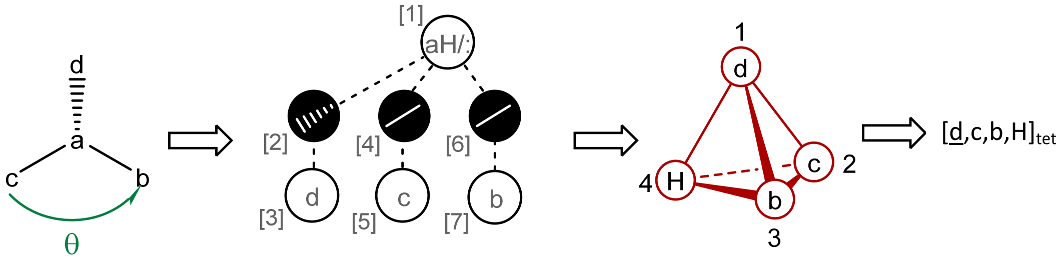


Figure 60 The stereo recognition pattern ‘tetrahedron7a’ representing tetrahedron motif H (Figure 59). The mapping of the pattern to a rooted tree is illustrated along with the atom mapping to the corresponding stereo descriptor (visualised via the associated reference framework).

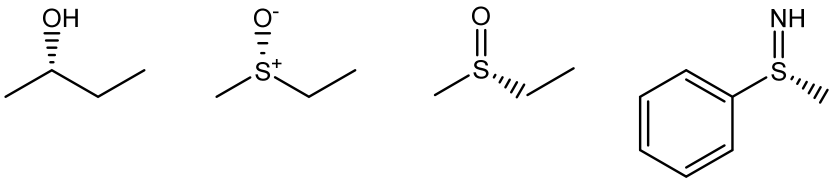


Figure 61 Example tetrahedral stereogenic centres covered by pattern ‘tetrahedron7a’

to promote early pattern rejection when the pattern could not possibly match due to the absence of the required stereo bond. The remaining in-plane bonds are set to subtend an angle between 5 and 175 degrees to ensure that the implicit hydrogen/lone pair is correctly placed in the first position of the stereo descriptor. These angle limits deliberately avoid a near collinear arrangement of the bonds at which point the intent of the drawing becomes ambiguous. The bond orders fields are also left open specifically to allow matching to sulphoximines, sulphimides type centres. The *EV* and *IV* fields are set up to map the matching atoms to positions corresponding in the '*tet*' stereo descriptor.

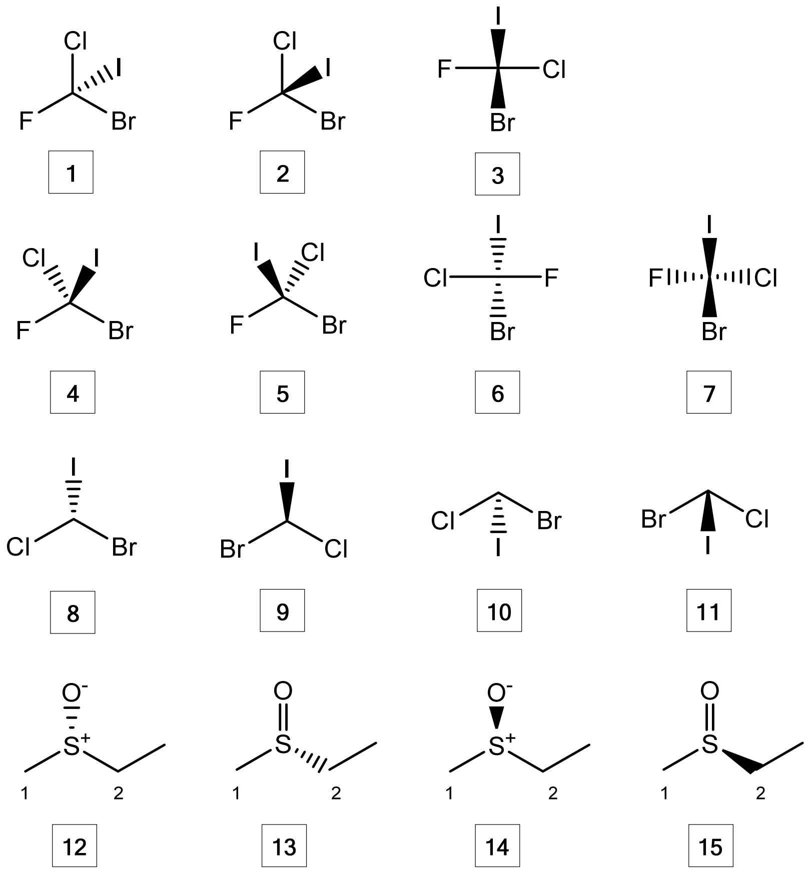


Figure 62 Tetrahedron test cases. In all cases the internal atom numbering is: F = 1; Cl = 2; Br = 3; I = 4; O = 3. Carbon atoms are explicitly numbered in the drawing.

The 11 tetrahedron patterns were validated using the test cases listed in Figure 62. These include tests specifically designed to validate the generation of stereo descriptors containing implicit hydrogen and lone pairs. Additional tests were added to check alternative forms of drawing sulphoxides in both ylid and double bond form. The results are listed in Table 8. Validation was determined by expected outcomes coded in the unit tests.

Test Molecule	Tested Motif	Matched pattern	Generated Descriptor	Validated
1	B	'tetrahedron2'	[4 2 1 3]$_{\text{tet}}$	✓
2	A	'tetrahedron1'	[4 2 3 1]$_{\text{tet}}$	✓
3	D	'tetrahedron4'	[3 1 4 2]$_{\text{tet}}$	✓
4	G	'tetrahedron3b'	[4 2 3 1]$_{\text{tet}}$	✓
5	F	'tetrahedron3a'	[4 2 3 1]$_{\text{tet}}$	✓
6	E	'tetrahedron5'	[4 2 3 1]$_{\text{tet}}$	✓
7	C	'tetrahedron6'	[3 1 4 2]$_{\text{tet}}$	✓
8	H	'tetrahedron7a'	[4 2 3 H]$_{\text{tet}}$	✓
9	K	'tetrahedron8a'	[4 2 3 H]$_{\text{tet}}$	✓
10	J	'tetrahedron7b'	[4 2 3 H]$_{\text{tet}}$	✓
11	L	'tetrahedron8b'	[4 2 3 H]$_{\text{tet}}$	✓
12	H	'tetrahedron7a'	[3 1 2 :]$_{\text{tet}}$	✓
13	H	'tetrahedron7a'	[2 3 1 :]$_{\text{tet}}$	✓
14	K	'tetrahedron8a'	[3 2 1 :]$_{\text{tet}}$	✓
15	K	'tetrahedron8a'	[2 1 3 :]$_{\text{tet}}$	✓

Table 8 **Outcomes produced by the stereo-perception module. The test input molecules are listed in Figure 62 and the motifs in Figure 59.**

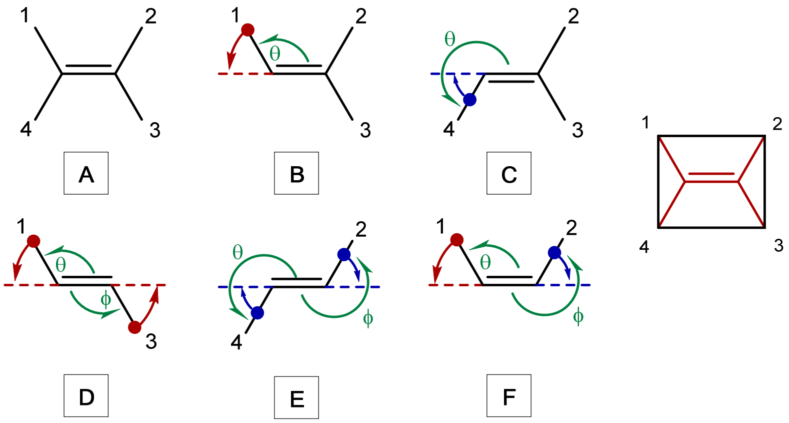


Figure 63 **'Olefin' recognition motifs. The numbered atoms are the corresponding positions in the generated 'ol' stereo descriptors as defined in the reference framework. The red and blue arrows and dashed lines mark the minimum and maximum extents of the ligand angle limits.**

Olefin-like stereogenic centres were covered by the motifs listed in Figure 63. Terminal olefin motifs are omitted as an optimisation due to the impossibility of isomers when two *gem* hydrogens are present.

The conversion of the olefin motif F (Figure 63) into a corresponding recognition pattern '*olefin4*' is used as an exemplar for the olefin class (Figure 64). The pattern is specifically designed to cover trans-substituted C=C, C=N and N=N double bonds. In contrast to the generalised tetrahedral patterns, the olefin patterns apply rigorous constraints for the bond orders. The efficiency of the pattern matching approach is improved if specific values can be assigned to as many attributes as possible, especially those nearest the root of the pattern tree. The atom '*D*' and '*implicit H/LP*' attributes in conjunction with the '*angle limits*' attribute of the attached bonds are used to define the trans relationship between the explicit ligands.

pattern "olefin4"

	orbit	Type	aromatic	Z	D	implicit H / LP	bond order	bond direction	angle limits	EV	IV
[1]	0	Bond	no				2	None	-		
[2]	1	Atom	no	C N	2	1				-	4
[3]	2	Bond	no				1	None	5 < θ < 175 from bond [1]		
[4]	3	Atom	-	-	-	-				1	-
[5]	1	Atom	no	C N	2	1				-	3
[6]	2	Bond	no				1	None	185 < ϕ < 355 from bond [1]		
[7]	3	Atom	-	-	-	-				2	-

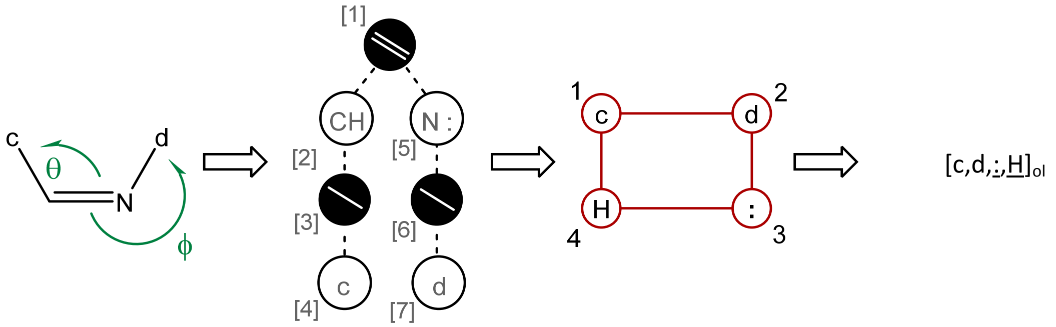


Figure 64 The stereo recognition pattern 'olefin4' representing olefin motif F (Figure 63).

The 6 olefin patterns were validated with the test cases listed in Figure 65 and the results presented in Table 9.

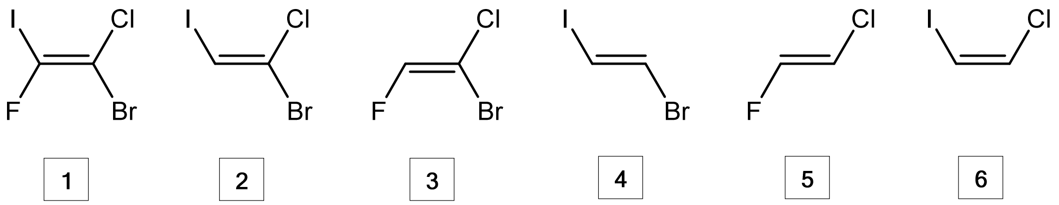


Figure 65 Olefin test cases. The internal atom numbering is: F = 1; Cl = 2; Br = 3; I = 4.

Test Molecule	Tested Motif	Matched Pattern	Generated Descriptor	Validated
1	A	‘olefin1’	[4 2 3 1]$_{ol}$	✓
2	B	‘olefin2a’	[4 2 3 H]$_{ol}$	✓
3	C	‘olefin2b’	[H 2 3 1]$_{ol}$	✓
4	D	‘olefin3a’	[4 H 3 H]$_{ol}$	✓
5	E	‘olefin3b’	[H 2 H 1]$_{ol}$	✓
6	F	‘olefin4’	[4 2 H H]$_{ol}$	✓

Table 9 Outcomes produced by the stereo-perception module. The test input molecules are listed in Figure 65 and the motifs in Figure 59.

Figure 67 to Figure 68 present the motifs used as the basis of the recognition patterns for allenes, bi-aryl atropisomers and coordination complexes. The patterns created for these motifs were validated in a similar manner to the tetrahedron and olefin patterns with a covering set of unit tests. The details of the results are not presented here but are available via running the unit test program ‘*test-stereo-finder*’.

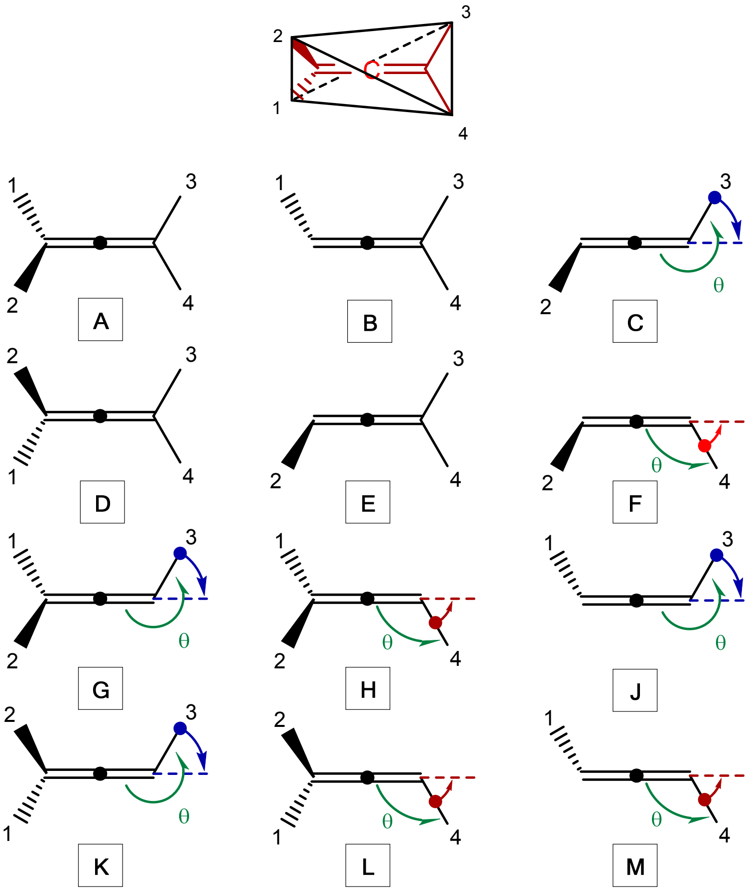


Figure 67 ‘Allene’ recognition motifs. The numbered atoms are the corresponding positions in the generated ‘al’ stereo descriptors as defined in the reference framework. The red and blue arrows and dashed lines mark the minimum and maximum extents of the ligand angle limits.

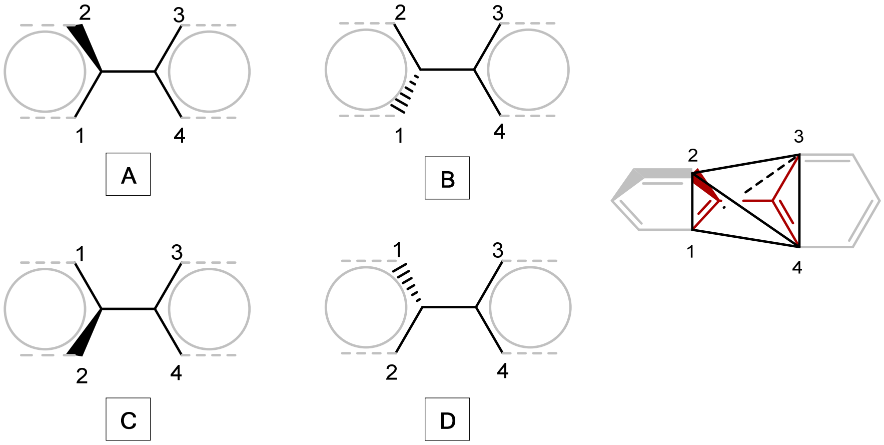


Figure 66 A selection of ‘atropisomer’ recognition motifs. The numbered atoms are the corresponding positions in the generated ‘atr’ stereo descriptors in indicated in the reference framework.

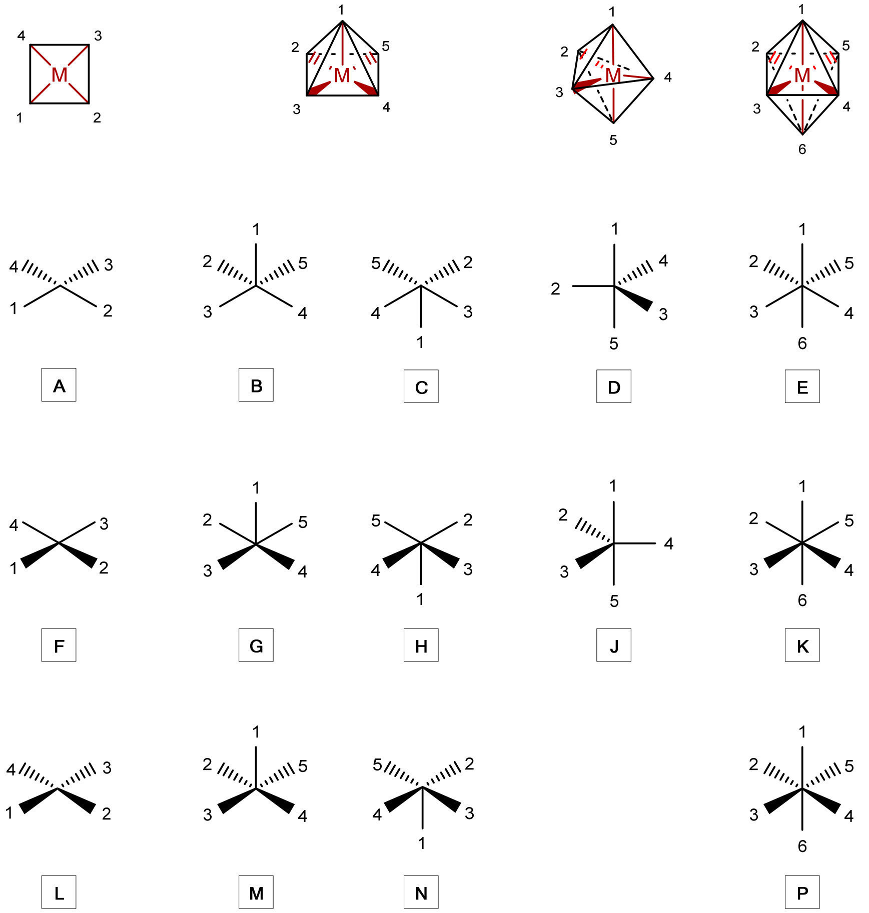


Figure 68 Coordination complex motifs. The numbered atoms are the positions in the generated 'spl', 'spy', 'tbp' and 'oct' stereo descriptors as indicated in the reference framework models corresponding to the geometries: square-planar; square-pyramid; trigonal bipyramid; octahedron.

Stereo Descriptor Matching

The stereo descriptor matching algorithm presented in earlier in this chapter was extensively validated using the '*test-stereo-match*' module in the unit test program '*test-stereo-finder*'. The tests were designed to completely cover the rotation and reflection permutations belonging to each framework. A self-consistency test was constructed that demonstrated that all hand written rotation permutations belonged to a cyclic group and that the reflection permutation was not a member. A series of unit tests were then written for each framework type to validate that 'same-coset', 'mirror-coset' and 'isomer-coset' relationships between stereo descriptors were valid.

est #	Descriptor 1	Descriptor 2	Coset Relationship	Validated
1	$[\ 1\ 2\ 3\ 4\ ]_{tet}$	$[\ 1\ 2\ 3\ 5\ ]_{tet}$	none	✓
2	$[\ 1\ 2\ 3\ 4\ ]_{tet}$	$[\ 2\ 1\ 3\ 4\ ]_{tet}$	mirror	✓
3	$[\ 1\ 2\ 3\ 4\ ]_{tet}$	$[\ 1\ 2\ 3\ 4\ ]_{tet}$	same	✓
4	$[\ 1\ 2\ 3\ 4\ ]_{tet}$	$[\ 1\ 3\ 4\ 2\ ]_{tet}$	same	✓
5	$[\ 1\ 2\ 3\ 4\ ]_{tet}$	$[\ 1\ 4\ 2\ 3\ ]_{tet}$	same	✓
6	$[\ 1\ 2\ 3\ 4\ ]_{tet}$	$[\ 2\ 1\ 4\ 3\ ]_{tet}$	same	✓
7	$[\ 1\ 2\ 3\ 4\ ]_{tet}$	$[\ 2\ 3\ 1\ 4\ ]_{tet}$	same	✓
8	$[\ 1\ 2\ 3\ 4\ ]_{tet}$	$[\ 2\ 4\ 3\ 1\ ]_{tet}$	same	✓
9	$[\ 1\ 2\ 3\ 4\ ]_{tet}$	$[\ 3\ 1\ 2\ 4\ ]_{tet}$	same	✓
10	$[\ 1\ 2\ 3\ 4\ ]_{tet}$	$[\ 3\ 2\ 4\ 1\ ]_{tet}$	same	✓
11	$[\ 1\ 2\ 3\ 4\ ]_{tet}$	$[\ 3\ 4\ 1\ 2\ ]_{tet}$	same	✓
12	$[\ 1\ 2\ 3\ 4\ ]_{tet}$	$[\ 4\ 1\ 3\ 2\ ]_{tet}$	same	✓
13	$[\ 1\ 2\ 3\ 4\ ]_{tet}$	$[\ 4\ 2\ 1\ 3\ ]_{tet}$	same	✓
14	$[\ 1\ 2\ 3\ 4\ ]_{tet}$	$[\ 4\ 3\ 2\ 1\ ]_{tet}$	same	✓
15	$[\ 1\ 2\ H\ 4\ ]_{tet}$	$[\ 2\ 4\ H\ 1\ ]_{tet}$	same	✓
16	$[\ 1\ 2\ H\ 4\ ]_{tet}$	$[\ 2\ H\ 4\ 1\ ]_{tet}$	mirror	✓
17	$[\ 1 : 3\ 4\ ]_{tet}$	$[\ : 3\ 1\ 4\ ]_{tet}$	same	✓
18	$[\ 1 : 3\ 4\ ]_{tet}$	$[\ 1\ 3 : 4\ ]_{tet}$	mirror	✓
19	$[\ 1\ 2\ 3\ 4\ ]_{tet}$	$[\ 1\ 2\ 3\ 4\ ]_{al}$	none	✓

Table 10 Validation tests for matching tetrahedral stereo descriptors.

Chapter 3

Symmetry and Isomorph Free Pattern Matching

Introduction

The extraction of stereo selective reaction examples to support a retrosynthetic rule requires that an appropriate query is derived from a transformation rule and compared to each record in a reaction database. The condition required for a reaction record to match the query is that a subgraph isomorphism exists between the query graph and the target graph, with the additional constraint that the stereochemical features also match. The representation of a reaction as a hyperstructure graph (*vide supra*) suggests that subgraph isomorphism algorithms can be readily adapted to solve the reaction search problem as reaction hyperstructures are represented as coloured (labelled) graphs. The adaptation of a constraints satisfaction problem (CSP) solver to solving subgraph isomorphism with stereochemical reaction constraints is described.

The exhaustive application of a reaction rule to transform a target molecule into all possible precursors frequently generates duplicate molecules due to symmetries in either the target molecule or the transform rule, or both. A novel isomorph free matching algorithm based on a CSP solver is described that avoids the generation of duplicates. A key step in isomorph free matching is the detection of stereochemically constrained permutation symmetries in both the query reaction rule and the target molecule graphs.

The validity of the CSP solver approach to isomorph free matching and stereochemical constrained symmetry perception is tested by the implementation of additional algorithms for the recognition of *meso* and (pseudo) C_2 symmetry, the stereotopicity of trigonal centres and the asymmetric/pseudo asymmetric classifications of stereogenic centres.

The Subgraph Isomorphism Definition

In graph theoretical terms a subgraph isomorphism exists between a query graph G_q and a target graph G_t if all the atoms of G_q can be mapped to a subset of the atoms of G_t such that the bonds in G_q are also completely mapped to a subset of the bonds in G_t. This constraint explicitly requires that if a bond exists between two atoms in G_q, then a bond must also exist between the mapped atoms in G_t. Additional constraints may require that if the atoms and bonds of G_q are coloured (have properties such as atomic number or bond order), then the mapped atoms and bonds in G_t are identically or equivalently coloured. [41]

Subgraph isomorphism belongs to a class of algorithmically equivalent problems that are labelled NP-complete.[172] That is there are no known solutions that in the worst case do not exhibit exponential rise in time behaviour for a linear increase in the size of the input. In the case of subgraph isomorphism the size of the input is the sum of the count of the atoms of G_q and G_t. (i.e. $|V(G_q)| + |V(G_t)|$ where V(G) is the set of atoms in graph G). The application of the most simplistic brute force solution of subgraph isomorphism requires testing every mapping of an atom in G_q with all the atoms in G_t before validating that the bond and colour constraints are satisfied. This requires generating and testing $|V(G_t)|! \,/\, (|V(G_t)| - |V(G_q)|)!$ mappings, thus demonstrating that even for small graphs a solution using this approach soon becomes impractical.

Fortunately there exist linear and low order polynomial algorithms that allow for the rapid discovery of target graphs that cannot possibly match a query graph. [173, 174, 175] Thus a screening phase is set up at the beginning of a search such that as much of the database as possible is eliminated before the need to apply a more costly subgraph isomorphism test. In general terms the database is pre-processed and matched against a library of graph fragments and an index is constructed. The query graph is similarly matched against the fragment index and database records that do not contain a fragment found in the query are immediately eliminated.

Subgraph Isomorphism Algorithms

Practical subgraph isomorphism algorithms have been devised that avoid a brute force approach by using constraints to prune out unproductive branches of a backtrack search.[176, 177, 178, 179] Thus after screening, the backtrack search can be solved in excellent time in the vast majority of cases. This is particularly true of graphs which exhibit low vertex connectivity and are vertex and edge coloured. Molecule graphs have a connectivity limited by atom valence and atom and bond 'colours' derived from atomic number, connectivity counts, charge, hydrogen counts and bond order. Reaction hyperstructures have marginally higher atom connectivity within the reaction site where atom connectivity is determined by valence combined with the number of made and/or broken bonds. The reaction hyperbonds on the other hand have more colours compared to regular molecules. This suggests that applying a subgraph isomorphism algorithm to a reaction hyperstructure graph may improve search times because the special features of the reaction site provide tighter constraints leading to better pruning heuristics.

A number of practical algorithms for molecule substructure searching have been published over the last 40 years. Research is still very much active [180] as new optimisations are sought to improve search times. The subgraph isomorphism problem remains NP-complete, but new algorithms aim to further reduce the space occupied by problem cases. The approaches can be divided into two general classes: backtracking algorithms; [177, 178, 179] and partitioning-relaxation algorithms. [181, 176, 182]

Backtracking algorithms improve upon the brute force approach by favouring a finer grained incremental generate and test approach. For example after a new pair of query and target atoms is selected and successfully bound, the next query atom chosen is one *adjacent* to a query atom that has already been bound. This heuristic ensures that the adjacency constraint can immediately be tested and unproductive branches in the search tree abandoned as early as possible (an example of a *pruning* heuristic). If an atom pair assignment is successful the algorithm steps forward and selects the next query atom and repeatedly chooses an untried target atom to attempt bind to it. If a matching target atom cannot be found the algorithm backtracks to the previous query atom and attempts to bind the next untried target atom. The algorithm is successful when all query atoms are bound. It is unsuccessful when backtracking has returned to the first query atom and there are no more untried target atoms.

The partitioning-relaxation algorithms first divide the atoms of the query and target into sets with equal or equivalent colours (based on atom type, connectivity degree, charge *etc.*) and then use an iterative procedure guided by atom adjacency to refine these sets into progressively smaller partitions until convergence is achieved and no further subdivisions can be made. Three outcomes are possible: a unique solution is found when each partition contains a single pair of query and target atoms; no solution is possible at all if at least one query atom has no target correspondents; and multiple solutions may exist if multiple target correspondents exist for at least one query atom. The latter case can be completely solved by employing a backtracking algorithm to make the final atom-atom assignments.

Stereochemical Searching

Little exists in the literature concerning matching stereochemical constraints for solving the subgraph isomorphism between a query substructure and a target molecule. It is known that commercial molecule search systems such as Isentris [183] and ChemFinder [184] amongst others are capable of substructure searches with stereochemical constraints, however the methods used in these systems remain unpublished.

Lipkus and Blower [185] have proposed a general matching method that can be used to apply relative stereochemical constraints to searches of the CAS registry file. The algorithm sets out to prove that one stereo isomer covered by the query is present in the set of possible stereo isomers covered by the target molecule representation. The method is applied as a post check after a substructure match and involves building a labelled bipartite multigraph [q] that represents the *relationships* between groups of stereogenic centres with known relative configuration. The graph vertices represent each *group* of relative stereocentres in both query and target molecule (Figure 69: graph 3).

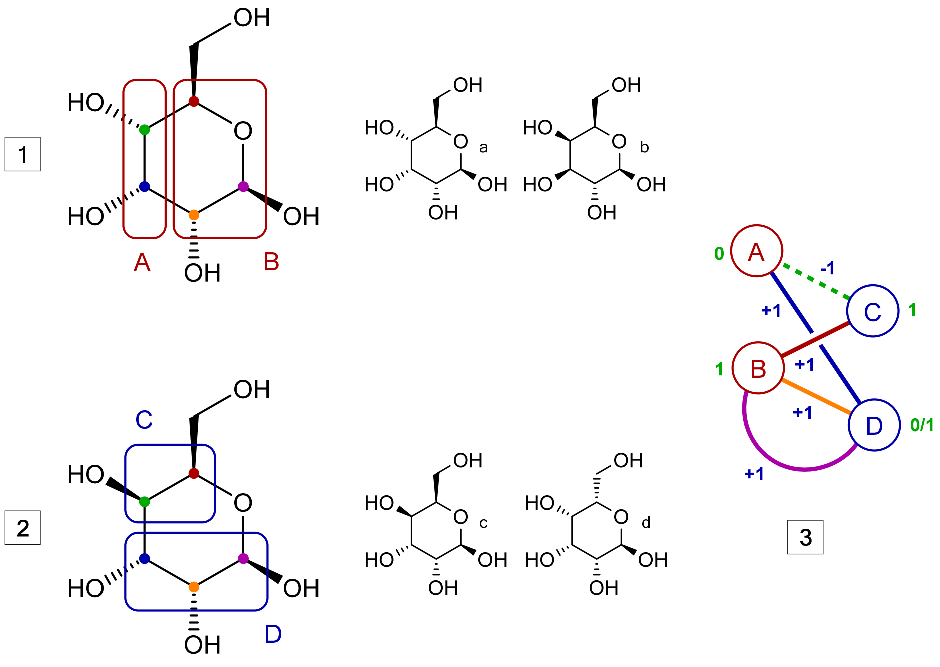


Figure 69 A graph representation of stereochemical constraints: 1) the query molecule with labelled groups of known relative stereochemistry representing the stereoisomer set (a) and (b); 2) the target molecule with labelled relative stereochemical groups representing the stereoisomer set (c) and (d); 3) the parity labelled bipartite graph representing the stereochemical constraints after mapping query 1 on 2. Labelling the vertices of the graph with '0' and '1' according to the edge parities results in a frustration at vertex D, proving that 1 and 2 do not cover a common stereoisomer.

Graph edges are added between the group vertices whenever there is a corresponding mapping between the query and target atom-based stereocentres in the group. The mapping is found in the subgraph isomorphism solution after finding a substructure match. An edge is labelled with

[q] A bipartite graph consists of two disjoint sets of vertices such that every edge joins a vertex in one set to a vertex in the other set. A multigraph permits multiple edges between the same pair of vertices.

the parity +1 if the mapping between the corresponding pair of stereocentres belongs to the *same* rotation coset. An edge is labelled with the parity -1 if the stereocentre pair belongs to a *mirror* coset. A pair of vertices may be joined by multiple edges if more than one pair of corresponding stereocentre exists in the same pair of stereo groups. However, if the multiple edges are not labelled with the same parities then the stereochemical constraints are not satisfied and no match is possible.

A match exists if a consistent binary labelling of the stereochemical graph vertices is possible (Figure 69: green numbers on graph 3). The procedure starts by choosing a vertex and arbitrarily labelling it ‘0’. Each connected edge is then traversed and if the edge parity is +1 the vertex attached to the other end is labelled ‘0’, otherwise it is labelled ‘1’. The procedure is repeated over each subsequent connected edge where parity of +1 retains the vertex label while a parity of -1 inverts the label by transforming ‘0’ to ‘1’ or ‘1’ to ‘0’. If the visited vertex is already labelled then that label must be consistent with the edge parity that has just been traversed. A mismatch is proven if vertex consistency is *frustrated* when two connected edges are in contradiction and attempt to induce both ‘1’ and ‘0’. It can be seen that frustration can only occur when an odd number of negative edge parities exist in a cycle.

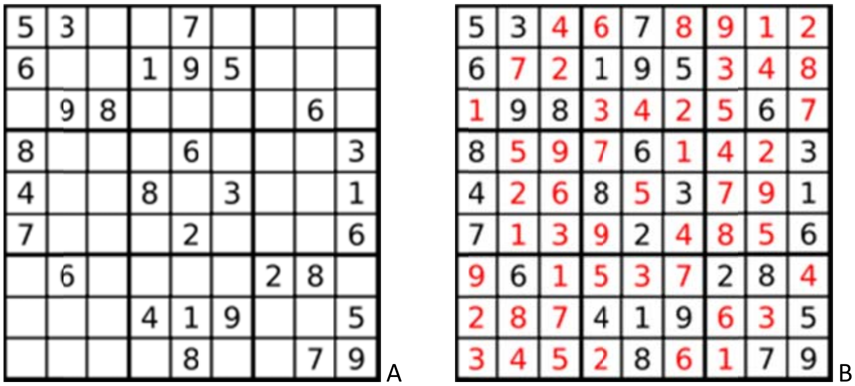
The Lipkus-Blower procedure has the advantage that a single query graph can be used to represent a set of distinct stereoisomers, saving multiple invocations of the subgraph isomorphism algorithm for each explicit stereoisomer. It has the disadvantage that if a stereochemical match is frustrated then the subgraph isomorphism matcher must be re-invoked to determine if alternative solutions exist. The matcher has to exhaust all possible mappings to determine that no solution exists. This situation can be improved if the stereochemical constraints procedure is *embedded* into the subgraph isomorphism algorithm such that the satisfaction of stereochemical constraints is continuously monitored as each query atom to target atom binding is made. This would ensure that a backtracking step is forced as *soon* a stereochemical constraint fails.

Constraint Satisfaction Problems

It has been recognised that (sub)graph isomorphism is a specific case of the more general constraints satisfaction problem (CSP).[186, 186] Ullman has published an extensive review of the CSP literature with guidance on the application of a CSP solver to the subgraph isomorphism problem. [180, 187] CSPs are an intensely researched area with applications in artificial intelligence, operations research and game playing. The popular Sudoku number puzzle [188] is an excellent example of a CSP. A graph isomorphism matcher implemented as a CSP solver is attractive as it is then possible to add stereochemical constraints directly to the problem representation. This would achieve the goal of interleaving stereochemical constraints satisfaction within a backtracking search with opportunities to add optimal ordering heuristics to improve search times.

Definition of a CSP

A CSP is formally defined as follows: [180] it is a triple <V, D, C> where V is a set of variables, D is a domain of values to be assigned to the variables and C is a set of constraints that limit the allowed assignments. Each constraint is itself a pair <T, R> where T is a n-tuple [r] of variables and R is an n-ary [s] relation on the domain D. An evaluation a variable *v* is a function *f* that maps from the set of variables V to the domains values in D (*i.e.* $f : V \rightarrow D$). An evaluation *f* satisfies a constraint $<(v_{1}, \ldots v_{n}), R>$ if $(f(v_1), \ldots f(v_n)) \in R$. A solution is an evaluation that satisfies *all* constraints.



5	3			7				
6			1	9	5			
	9	8					6	
8				6				3
4			8		3			1
7				2				6
	6					2	8	
			4	1	9			5
				8			7	9

A

5	3	4	6	7	8	9	1	2
6	7	2	1	9	5	3	4	8
1	9	8	3	4	2	5	6	7
8	5	9	7	6	1	4	2	3
4	2	6	8	5	3	7	9	1
7	1	3	9	2	4	8	5	6
9	6	1	5	3	7	2	8	4
2	8	7	4	1	9	6	3	5
3	4	5	2	8	6	1	7	9

B

Figure 70 The Sudoku puzzle is an example of a constraints satisfaction problem: A) initial problem state as a partial solution; B) a complete solution that satisfies all rows, columns and 3x3 block constraints.

This definition is best explained by considering the Sudoku puzzle (Figure 70). In the puzzle there are 81 variables arranged as a 9 x 9 matrix. Each variable can take a value from the finite domain {1, 2, 3, 4, 5, 6, 7, 8, 9}. The constraints are: each row (a tuple of variables) can only use each value once (an n-ary relation); the constraint described for each row is also applied to each column; the 9 x 9 matrix is blocked into 3 x 3 groups of grids, each of 3 x 3 variables with the constraint that the assigned domain values must only occur once in each block. Starting from a

[r] An n-tuple is an ordered list of n elements.

[s] An n-ary relationship is a relationship that maps between a subject and n values.

given partial solution, the game player must use either logic and/or a backtracking search to find a complete solution that satisfies all constraints.

Stereochemical Subgraph Isomorphism as a CSP

Subgraph isomorphism can be solved using a binary CSP solver.[180] The variables (V) of the CSP are represented by the query atoms while the domain values (D) of the CSP consist of the target atoms. Constraints in a binary CSP are represented by a constraints graph in which the variables are represented by vertices and edges represent pairs of variables that are subject to a binary constraint. Two variables are said to be *adjacent* if, and only if, they are subject to the same binary constraint. Hence we can define A_i to be the set of variables adjacent to variable V_i and C_{ij} is the constraint between variable V_i and $V_j \in A_i$.

When variable V_i is an atom, A_i is equivalent to the ith row vector in the atom adjacency matrix of the molecule graph (or reaction hyperstructure graph). Thus the graph adjacency matrix can be used as the atom/bond constraints graph in a subgraph isomorphism constraints satisfaction problem.

The subgraph isomorphism problem description can be extended to include stereocentres as additional variables and domain values to obtain an explicit mapping between stereocentres in the CSP solution. When the variable V_i is a stereocentre then A_i is equivalent to a row from the parity labelled adjacency matrix of the Lipkus-Blower stereochemical constraints graph (*vide supra*). These adjacency relationships make it possible to treat subgraph isomorphism and stereochemical matching in a unified manner within a CSP solver.

If a constraint between variables V_i and V_j exists, then in any solution *s* there must be a predicate function $P_{ij}(s_i, s_j)$ that evaluates to true. When V_i and V_j are atoms, the function $P_{ij}(s_i, s_j)$ is true when $V_j \in A_i$ (the atoms are bonded). When V_i and V_j are stereocentres the function $P_{ij}(s_i, s_j)$ is true only if the parity labelling of the stereo group vertices that contain V_i and V_j are not frustrated when applying the edge parity in A_i. The adjacency matrix row A in a stereochemical constraints graph is conveniently and compactly represented by two bit-vectors, one for positive edge parities, and the other for negative edge parities.

Domain Reduction

Domain reduction is the early removal of values that cannot be part of a solution. When part way towards a solution we may have *X* variables instantiated (that is query atoms, bonds or stereocentres assigned to target counterparts). At this point it is possible to identify domain values for the remaining variables that cannot be part of any solution and thus prune out parts of the remaining search space by reducing the number of variable/value combinations explored. The identification of non-solution values is based on the following inference:

- Given that a binary constraint between V_i and V_j in a solution *s* is satisfied when $P_{ij}(s_i, s_j)$ is true, then a value *y* in the domain D_i of V_i cannot belong to a solution unless there is another value $z \in D_j$ such that $P_{ij}(y, z)$ is also true.
- Under these conditions $z \in D_j$ is said to support $y \in D_i$.[189] Consequently a value $y \in D_i$ that is supported by at least one value in domain D_j is supported in constraint C_{ij}.
- A value $y \in D_i$ is defined as fully supported if, and only if, it is supported in every C_{ij} such that $V_j \in A_{i.}$
- Hence a value $y \in D_i$ that is not fully supported cannot belong to any solution *s* and can thus be removed from domain D_i.

Put succinctly, a value can only exist in a domain if *all* the *adjacent* constraint variables also have domain values *adjacent* to the considered value in the constraints graph.

The removal of a value from a domain may undermine support for variables in other domains. Consequently the removal procedure may be repeated until convergence is reached where all remaining values are fully supported and domain reduction is maximised. [190]

This *reduction-to-convergence* technique forms the basis of the Ullman algorithm, [177] used for solving subgraph isomorphism, and the arc consistency algorithms[t] (AC-3, AC-4 *etc.*) used in CSP solvers. [190, 191] Other reported approaches elect not to achieve convergence, [186] instead propagating the reduction by a fixed number of relaxations in what is known as a *look-ahead* domain reduction. This approach is sometimes necessary because most of the reduction gains occur in the first pass, subsequently achieving less and less gains in later passes at the cost of consuming increasing processing time.

In general a CSP solver invokes a domain reduction procedure of some sort immediately after each variable instantiation. If any variable's domain is emptied after domain reduction then a solution cannot be achieved with the current partial solution and the solver must backtrack to try a different variable instantiation and proceed from there. Domain reduction can also be performed before the backtracking search is entered. Here unsupported values are permanently removed from domains as they can never lead to a solution. If any domain is emptied in this pre-

[t] Arc consistency is referring to the relations represented by edges in the constraints graph.

processing phase then no solutions at all are possible and the search can be immediately curtailed.

Invariant Domain Reduction

In graph and subgraph isomorphism (i.e. exact structure and substructure search) a technique known as invariant domain reduction can be applied in the pre-processing phase.[180, 192] This is performed via the application of a unary constraint represented by the predicate P_i. $P_i(z)$ is true if, and only if, $z \in D_i$ satisfies the constraint. Hence the unary predicate P_i selects a subset of D_i.

In graph isomorphism an atom *invariant* is a set of properties such that a pair of atoms can be assigned only if the invariants of both atoms exactly correspond. For subgraph isomorphism this is relaxed to allow specific relationships between the invariant values. For example the connectivity degree of a query atom cannot exceed the degree of a target atom to qualify for domain membership but it can be less. Invariant properties include atomic number, bond order and the geometry class of stereocentres. McKay lists many more general graph invariants that can be used for atom partitioning such as path distances between atoms. [193] In the context of stereochemical reaction searching, invariant properties derived from the perception of changes in a reaction hyperstructure can be considered. This includes all perceived changes to atom, bond, ring and stereochemical reaction properties (*vide supra*).

Efficient invariant domain reduction is achieved by: finding all invariant property partitions of atoms, bonds and stereocentres within the query; finding all invariant partitions of atoms, bonds and stereocentres in the target; then selecting single representative pairs from each of these query and target partitions such that if a unary predicate taking the invariant values as arguments evaluates as true, then *all* the values in the target partition are added to the domains of *all* the variables in the corresponding query partition.

The independent invariant partitions in the query (or target) are found by comparing all query variables with each other (or all target values with each other) and partitioning them into common sets if, and only if, all invariant properties are equal. The same auto-partitioning procedure is performed for both a graph or subgraph isomorphism problem.

The *AllDifferent* Constraint

The *AllDifferent* constraint requires that every variable in the query (atom, bond and stereocentre) has a value in the solution that differs from the value of every other variable in the query. For example, a query atom cannot be simultaneously mapped to two different target atoms. This constraint is enforced after each variable instantiation by removing the assigned value from all domains. However the values must be restored whenever the CSP solver backtracks and unbinds the current instantiated variable. This restoration is managed by a push-pop stack data structure.

In principle before *AllDifferent* domain reduction is performed, copies of the original domains are pushed onto the top of a stack. Restoration entails removal of the domain copies from the top of the stack to replace the current domain sets.

In this authors work the implementation of these operations was made far more efficient by pushing and popping just the set of the current assigned values and using set difference operations on the original domain sets to modify a temporary copy of the domain sets (the edited domain sets). This provided a significant time and space optimisation.

Choosing an Order for Variable Instantiation

The order that variables are instantiated is important as optimised ordering frequently shortens the search by promoting early pruning of the largest unproductive branches in the search tree. [186]

Two general approaches are possible: [180] a static ordering is computed once in a pre-processing step using heuristics to estimate an optimal ordering based on the *initial* domain state; or a dynamic ordering needing frequent computation using heuristics to estimate an optimal *partial* ordering based on the *current* domain state. In the latter case it has be observed that gains made must be offset by the overheads of enforced constraint re-evaluation each time the variable order is changed. On the whole this extra work cancels out the hoped for gains. [186]

In practice an effective static ordering can be achieved by first selecting the variable with the smallest initial domain. [186] The next variable is chosen from the set of variables linked by constraints to those variables that have already been instantiated (i.e. those that *precede* it in the ordered variable list) *and* has the smallest initial domain. This is repeated until all variables are selected into the list. This approach tries to ensure that domain values are removed as quickly as possible to reduce the number of tested combinations (the “ASAP” principle). [194]

This latter approach was adopted for the stereochemical CSP solver implementation as the ordering heuristic ensured that all bond and stereocentre constraints are evaluated as early as possible when applied to molecule or reaction matching.

Solving Stereochemical Constraints

Stereochemical constraints are tested in the following manner. Each stereo constraint is marked for testing when all atoms referenced in the query stereo descriptor have been instantiated by the CSP solver. These evaluation points are marked in the ordered query list in the CSP pre-processing step. The query stereo descriptor is inspected and the corresponding (mapped) target atoms are then used to lookup the corresponding candidate target stereo descriptor (if it exists). If found, the target stereo descriptor is compared in detail to the query stereo descriptor to determine if they belong to the 'same-coset' or the 'mirror-coset' (see Chapter 3).

The CSP solver can be configured to allow absolute, relative, epimer or diastereomer stereochemical matching and the match state reports which of these conditions were met. Support for multiple relative stereo groups was not implemented meaning that the stereo constraints graph was simplified to holding just two nodes representing the chiral centres in the query and target graphs (Figure 71). When configured for relative matching the first query stereo descriptor match sets an edge parity of plus 1 in the stereo constraints graph if the match was the 'same-coset' value or as minus 1 if it was the 'mirror-coset' value. Subsequent stereo descriptor constraint tests must match with the *same* edge parity (F and G, Figure 71) otherwise the constraint fails and the search backtracks. When the CSP solver is configured for absolute stereochemical matches the only permitted edge parity is plus 1 (F, Figure 71). Epimer matching

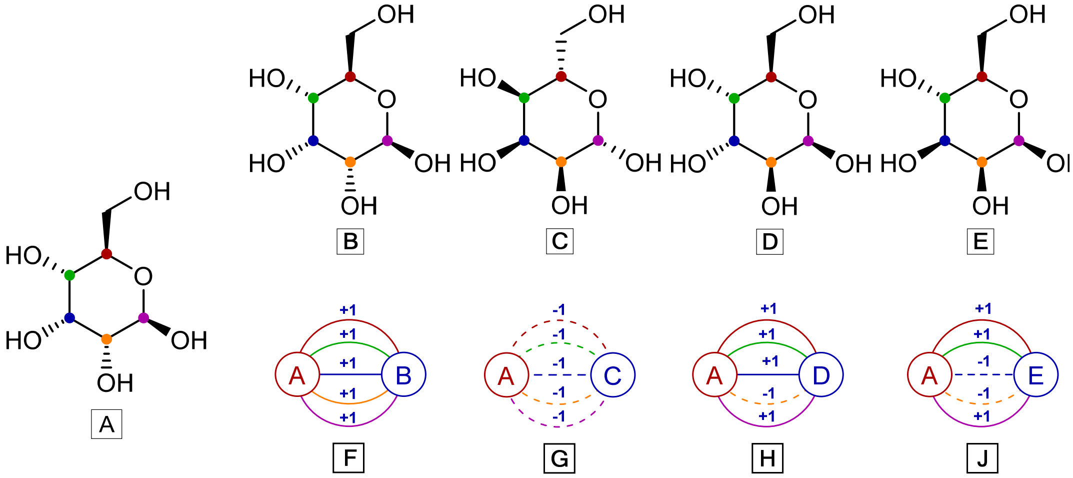


Figure 71 **Stereochemical matching. Query A matches molecules B and C but not D or E when the CSP is set for relative stereochemical matching. Query A only matches B when set for absolute stereochemical matching. Constraints graph F is a solution to an identical stereochemical match (all +1); G is a solution to an enantiomer match (all -1); H is a solution to an epimer match (one -1, more than zero +1); J is a solution to a diastereomer match.**

requires two or more edges in the stereo constraints graph but permits just one edge with a minus 1 parity (H, Figure 71) representing the inversion of one stereocentre. No constraints are set for diastereomer matches but the resulting match state is inspected to determine that it is not an absolute, relative or epimer match before reporting it as a diastereomer match.

Stereocentres that do not support a mirror coset (such as olefins or square planar complexes) do not take part in the constraints analysis as there is no parity interdependence with other stereo descriptors.

Implementing a Stereochemical CSP Solver

A specialised CSP solver has been designed to perform stereochemical exact structure and substructure searches on both molecules and reactions. Figure 72 presents the overall flow chart for the operation of the CSP solver (*vide infra*).

The solver is entered at step 1 to find the first solution. Query and target graphs representing molecules or reactions are bound to the solver. All subsequent operations assume these graphs remain available to access the atom/bond/stereo descriptor variables and domain values, the atom/atom and atom/stereo descriptor adjacency matrices and other supporting information required for implementing the unary predicates in the invariant domain reduction step.

At step 2, invariant domain reduction is performed to assign initial target domain values for all relevant atoms, bonds and stereocentres. Explicit hydrogen atoms and bonds connected to explicit hydrogen atoms are removed at this stage. Different implementations of this step are required depending on whether exact or substructure matches are being sought and if molecules or reactions are being compared.[u] This is followed by domain reduction of unsupported variables, with a choice to either reduce to convergence or perform a single pass reduction. If any domain is emptied at this point the routine exits.

The variables are now sorted into an instantiation order according to McGregor's optimal sorting algorithm[186] favouring an initial variable with the smallest domains followed by variables with the largest number of immediately solvable constraints and smallest domains.

[uu] The stereochemical CSP supports replaceable implementations for binding query and target graphs, and solving unary and binary/n-ary constraints. This allows a wide range of problem types to be configured and solved, such as assembling transforms rules into hierarchies, matching rules to target molecules and so on.

The final initialisation step requires that the sorted variable list is scanned from beginning to end to locate the earliest points that each stereochemical constraints test can be triggered. When all the atoms associated with a particular stereocentre have been encountered, a sub-list entry is inserted at that point to mark the constraint test point. Reaction hyperstructures will insert test points for both educt and product stereocentres.

At step 5, a tentative query atom to target atom assignment has been proposed. At this point atom adjacency constraints are validated such that a connecting bond must exist between all adjacent instantiated atoms and that these target bonds are selected from the corresponding bond domain.

Any stereochemical constraints present are tested at this point according to the Lipkus-Blower stereo constraints graph method. The inclusion of both educt and product stereocentres in a single stereo constraints graph ensures that constancy occurs both within the educt and product structures and *across* the reaction arrow.

If any of the constraints fail the algorithm returns to step 4 to try another atom from the target domain set. If all candidate atoms have been exhausted in the domain the algorithm proceeds to steps 8 and 9 to backtrack to the previous query atom and then continues to test the next untried candidate target atom.

If the constraints tests pass at step 5 the algorithm proceeds to step 6 to perform the following steps: record the query to target atom binding in a map; save state on a stack to support any later backtrack steps; apply the *AllDifferent* constraint to remove all assigned values from the unassigned domains; and performs a one pass look-ahead domain reduction by removing unsupported values.

The detection of an empty domain set after domain reduction immediately enforces backtracking at step 8 as no solution is now possible in the current branch of the search tree. Otherwise the next variable is selected at step 7 to be solved by a repetition of the mentioned steps.

The algorithm exits under two conditions: all variables are assigned to domain values so a solution exists and is stored as a map (step 23); or no variables are assigned and no further solutions are possible (step 14).

The CSP solver design supports re-entry to find alternative solutions at step 10. After the initial solution has been found the final state of the solver is preserved on the stack. The re-entry step 11 pops the stack to unbind the last query to target atom assignment to initiate a new search for the next solution. The last query variable is now tested against the next untried value in the

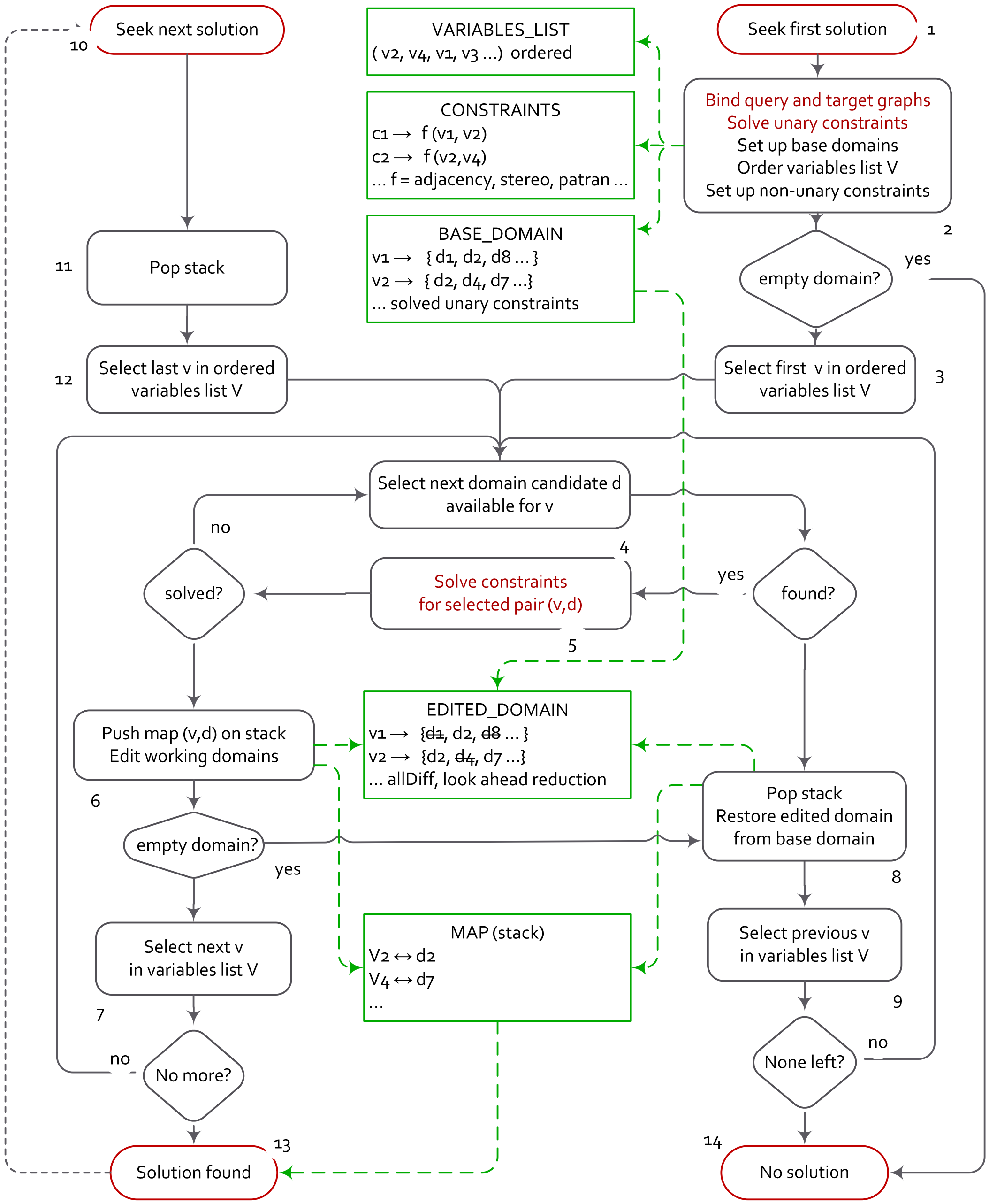

Seek next solution
10
VARIABLES_LIST
(v2, v4, v1, v3 ...) ordered
Seek first solution
1
CONSTRAINTS
c1 → f (v1, v2)
c2 → f (v2,v4)
... f = adjacency, stereo, patran ...
Bind query and target graphs
Solve unary constraints
Set up base domains
Order variables list V
Set up non-unary constraints
2
BASE_DOMAIN
v1 → { d1, d2, d8 ... }
v2 → { d2, d4, d7 ...}
... solved unary constraints
11
Pop stack
empty domain?
yes
12
Select last v in ordered variables list V
Select first v in ordered variables list V
3
Select next domain candidate d available for v
no
4
solved?
Solve constraints for selected pair (v,d)
yes
found?
5
Push map (v,d) on stack
Edit working domains
EDITED_DOMAIN
v1 → {~~d1~~, d2, ~~d8~~ ... }
v2 → {d2, ~~d4~~, d7 ...}
... allDiff, look ahead reduction
6
Pop stack
Restore edited domain from base domain
empty domain?
yes
8
Select next v in variables list V
MAP (stack)
V2 ↔ d2
V4 ↔ d7
...
Select previous v in variables list V
7
9
no
No more?
None left?
no
13
14
Solution found
No solution

Figure 72 A specialised CSP solver optimised for subgraph isomorphism.

corresponding domain by re-entering the main search loop at step 4. The algorithm can be repeatedly re-entered to collect all solutions until the *no-solution* state is reached.

Matching Implicit Hydrogen Atoms

The development of a structural language[v] to represent stereo selective transform rules necessitated the use of generic R_1, R_2, … atoms to clearly define stereochemical relationships across the reaction arrow when the ligands are anonymous. R_n atoms must support matching implicit (non-drawn) hydrogen atoms. This requirement alters the behaviour of the CSP solver with respect to emptied domains as in certain circumstances this does not signal that a search should backtrack or be abandoned.

CSP solvers can be designed to be flexible, partially relaxing constraints and allowing solutions not to comply with all constraints. [195, 196] A quality measure can be introduced that is a function of how many of the constraints are satisfied up to an externally selected limit. However it is not necessary to use this approach to support implicit hydrogen matches. The implemented solution introduces a new 'hydrogen count' constraint. All R_n atoms in the query have a coded value for implicit hydrogen added to their domains sets during the CSP solver set up (step 2). This code is the last value selected from any domain when assigning values to variables. When step 4 selects the implicit hydrogen code, the hydrogen count constraint is then applied in step 5. This requires that the target atom that is mapped to the query atom *adjacent* to the R_n atom must have at least one remaining unbound hydrogen atom available. If the constraint predicate evaluates as true (the adjacent hydrogen count is greater than zero) then the implicit hydrogen code is assigned to the query atom in the solution map and an additional note is made on the stack that the available hydrogen count on the adjacent atom is reduced by one. The stack is used so that backtracking can undo the hydrogen count reduction if the implicit hydrogen assignment is later un-bound.

There are further consequences for supporting implicit hydrogen:

- Domain reduction of unsupported variables cannot be performed on query atoms that support implicit hydrogen atoms, as it is no longer possible to infer if the variable is unsupported.
- An empty domain does not indicate that a solution is impossible. A special note is made of all query atoms that support implicit hydrogen atoms so that they are eliminated from the empty domain tests.

[v] This is described in detail in the next chapter.

- All variables (atoms) in the ordered query list that support implicit hydrogen must be *preceded* by its adjacent atom as these are paired with it in the hydrogen count constraints lists. This ensures that the hydrogen count constraint can be immediately solved as the adjacent target atom holding the available hydrogen count is known. This also implies that any query atom supporting implicit hydrogen cannot be the first instantiated variable.

Solving Symmetry Problems in Pattern Matching

> "*The efficient use of symmetry operations is one of the most difficult problems in a backtrack search. It is both a blessing and a curse --- a blessing because there is always hope of further optimization and a curse because it is the source of many programming errors*". Professor C. W. H. Lam.[W]

A number of problems in retrosynthetic analysis, such as the generation of precursors and the perception of functional groups, require that a query pattern is exhaustively matched to a molecule graph. If the target or query graph contains symmetries then the solutions will contain redundancies which need subsequent detection and removal.

To avoid this problem from the outset a novel isomorph rejection algorithm has been developed based on a CSP solver. This efficiently finds the automorphism generator set of any stereochemical graph and uses this to avoid duplicate matches before they are located. Execution times and comparisons to alternative automorphism algorithms are presented using sample molecule graphs and “challenge” graphs selected from the literature.

The isomorph rejection algorithm is used as the basis of an efficient perception algorithm that finds permutation symmetries in molecule and reaction graphs. The use of the symmetry perception algorithm for assigning *non-asymmetric*, *pseudo-asymmetric* and *asymmetric* stereocentre labels is presented. A method for the efficient classification of *meso* compounds, *C_2 symmetric* and *pseudo C_2 symmetric* compounds is described. Finally a novel procedure is used to assign stereotopicity labels to trigonal centres using the perceived symmetry classes.

Introduction

The motivation to implement a method for determining stereochemically constrained topological symmetry is two-fold. First it provides a means to efficiently eliminate duplicate precursor molecules when applying a retrosynthetic transform. This is a problem known as

[W] From a presentation on the problem of the “projective plane of order 10” accessed at http://www.cecm.sfu.ca/organics/papers/lam/paper/html/node4-an.shtml isomorph free generation [197] and is important when either (or both) the reaction rule or target molecule exhibit symmetries. Second, it provides a robust method for identifying identical appendages and hence solves a necessary step in the classification of symmetric, asymmetric and pseudo-asymmetric stereo centres.

Symmetry properties of molecules are important for the interpretation of IR spectra or X-ray diffraction patterns amongst a number of molecular phenomena. This type of geometrical symmetry requires information on the placement of atoms in three dimensional space with the application of formal group theory to determine the point or space group of the molecule. [198] A different type of symmetry that has relevance in chemistry is topological graph symmetry based on graph connectivity and provides a description of the permutation groups of the graph representing the atoms and bonds. Topological symmetry has applications in predicting the number and intensities of peaks in ^{1}H-NMR and ^{13}C-NMR [199] as well as solving problems in structure elucidation, [200] retrosynthetic analysis and compound retrieval from databases.

Topological Symmetry

Graph symmetry is represented by its automorphism group. A concise mathematical treatment is presented by Balasubramanian[201] and the terms and concepts are briefly summarised as follows. An automorphism of a graph is an isomorphism (bijective mapping) of the graph onto itself. Each automorphism is a permutation of the graph vertices that preserves bond adjacency (and bond and atom colouring) and the set of all these permutations is known as the automorphism group (denoted by **Γ**). Automorphism *partitioning* on the other hand is the division of the graph atoms into orbits (a set) such that there exist automorphisms within the automorphism group that map each atom member of the orbit to only the other orbit members. As a consequence the orbits in the automorphism partition must be disjoint and the union of all the orbits is the set of all atoms.

The properties of, and differences between, the automorphism group and the automorphism partition are as follows. Every symmetry operation on a graph is represented by an automorphism and conversely each automorphism represents a discrete symmetry operation. Hence the total symmetry of the graph is represented by the automorphism group. In the worst case the order of the automorphism group is N!, [x] where N is the number of atoms. The automorphism partition on the other hand represents the orbit in the graph of each atom under the operation of the automorphism group. An orbit is thus the set of atoms that a particular

[x] This is true in fully connected or complete graphs. In asymmetric graphs the order is 1.

atom can be moved to by members of the automorphism group. Atoms in the same orbit cannot be distinguished by any graph property, and are in every way topologically identical. The orbits are commonly referred to as equivalence (or symmetry) classes. In the limiting case the number of orbits is the size of the graph N. In some applications such as the detection of identical appendages in stereochemical naming or predicting the number of distinct peaks in a ^{13}C NMR, it is sufficient to know only the automorphism partition. [202]

Related to atom orbits are stabilisers. An atom stabiliser is the sub group of the automorphism group that leaves the atom unchanged (fixed in position). Stabilisers can be chained to fix a set of atoms. Chaining is a recursive process where the next atom in the set is stabilised within the stabiliser group of the previous atom, starting the chain at the first atom and the automorphism group. The definitions of an orbit, stabiliser and stabiliser chain are formalised as:

$$Orbit_v\,(\boldsymbol{\Gamma}) = \{\, g\,(v) \mid g \in \boldsymbol{\Gamma} \,\} \qquad eq\ 1.$$

$$Stabiliser_\Gamma\,(v) = \{\, g \in \Gamma \mid g\,(v) = v \,\} \qquad eq\ 2.$$

$$Stabiliser_\Gamma\,(u,v,\ \ldots) = \{\, g \in Stabiliser_\Gamma\,(v,\ \ldots) \mid g\,(u) = u \,\} \qquad eq\ 3.$$

$\boldsymbol{\Gamma}$ is the automorphism group; g is an automorphism; u, v are graph vertices (atoms).

Table 11 illustrates the automorphism group and partition concepts when applied to a representative graph and it introduces the concept of the automorphism generator set.[203] The generator set is an irreducible subset of automorphisms whose combinations under the group operation generate the complete automorphism group. The generator set listed in Table 11 is represented in cyclic form and operates by expressing a permutation as a series of circular rearrangements of groups of elements to transform one automorphism into another. For example, the operation (1 2)(4 5)(7 8) on the identity group [1 2 3 4 5 6 7 8] generates [2 1 3 5 4 6 8 7] and the composition of generator (1 2)(4 5)(7 8) with generator (1 7)(2 8)(3 6) operating on [1 2 3 4 5 6 8 7] generates [7 8 6 5 4 3 2 1] and so on.

A useful property of the generator set is the maximum number of generators needed to form the automorphism group never exceeds (N – 1) where N is the number atoms in the graph. [204] It is known that finding the ideal *smallest* generating set is computationally NP-hard[y], but finding a

[y] Non-deterministic polynomial-time hard problems are a class of problem in computational complexity theory for which it is suspected that no polynomial time solution exists, but this is not yet proven.

reasonable small one can be computationally less demanding, requiring only a polynomial time algorithm. [204]

The derivation of the automorphism partition directly from a generator set is performed by computing the unions of all intersecting cyclic cells extracted from the generating set. Working with the example in Table 11, the cyclic cells of the generators are extracted into the collection (1 2), (4 5), (7 8), (1 7), (2 8), (3 6), (1 2), then rearranged and clustered into 3 groups of mutually intersecting cells **[**(1 2) (1 2) (2 8) (7 8) (1 7)**], [**(3 6)**], [**(4 5)**]** and reduced into the disjoint partition sets {1 2 7 8}, {3 6}, {4 5}.

Perceiving Topological Symmetry

Two classes of algorithms designed to recognise topological graph symmetry have been described in the chemistry literature. They differ in both algorithm complexity, accuracy and the information discovered.

The first group are based on atom partitioning and iterative labelling procedures based on relaxation methods. These generally run in polynomial time, [205] but only discover the automorphism partition. Within this group three approaches have been tried: calculation of extended connectivity, [170, 68, 79, 206, 207] of which methods based in the Morgan algorithm [66] have been extensively studied; computation of the higher powers of the adjacency matrix; [85, 208] and the determination of the eigenvalues of the adjacency matrix. [209] It has been recognised that this class of algorithm is approximate [210, 82] and many types of graph (some simple) are not correctly

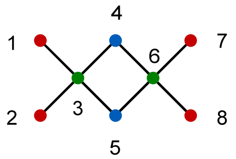
					
Automorphism					
Group (Γ)				Generator Set	Partitions
[1 2 3 4 5 6 7 8]	[1 2 3 4 5 6 8 7]	[8 7 6 5 4 3 2 1]	[8 7 6 5 4 3 1 2]	(1 2) (4 5) (7 8)	● {1 2 7 8}
[2 1 3 4 5 6 7 8]	[2 1 3 4 5 6 8 7]	[7 8 6 5 4 3 2 1]	[7 8 6 5 4 3 1 2]	(1 7) (2 8) (3 6)	● {3 6}
[1 2 3 5 4 6 7 8]	[1 2 3 5 4 6 8 7]	[8 7 6 4 5 3 2 1]	[8 7 6 4 5 3 1 2]	(1 2)	● {4 5}
[2 1 3 5 4 6 7 8]	[2 1 3 5 4 6 8 7]	[7 8 6 4 5 3 2 1]	[7 8 6 4 5 3 1 2]		

Table 11 Automorphism group, generators and partitions of a representative graph.

partitioned due to the presence of *isospectral* points. [211, 212] Isospectral points are atoms that ultimately receive the same label but are not symmetrically equivalent. Two atoms receiving different labels are never topologically equivalent but the converse cannot be proven without testing them against the automorphism group.

Razinger *et al*, amongst others, have provided realistic counter examples [202] that demonstrate algorithm failures in both the Morgan [66] and the related SEMA [68] algorithms. The cause of the problem has been identified by Read and Corneil [210] as being the inability to find a suitable graph invariant[z] to induce a good enough initial atom partition. The search for graph invariants that can solve the automorphism problem in polynomial time continues to ensnare and occupy chemists. [205, 80, 81, 213, 214] This is despite Read and Corneil proving that any complete solution is non-polynomial and equivalent to performing a graph isomorphism.[210]

The second group of algorithms is based on using a backtrack search to find either the full automorphism group or alternatively a generator set. The simplest approach uses a brute force search that systematically generates each automorphism and so directly generates the automorphism group. An example that illustrates the combinatorial problem of this approach within realistic chemical space is the relatively simple graph SS 22 (Figure 77: *vide infra*) which has 54 vertices and $(3!)^{12}.(2!)^{6}.6.2$ = 1,671,768,834,048 automorphisms. Assuming 10,000 automorphisms can be generated per second [aa] this needs over 5 years to compute and would exhaust storage space long before the procedure completes.

McKay's graph isomorphism algorithm (NAUTY) was the first reported to generate an automorphism generator set, albeit as a side effect of canonical graph naming. [215, 193, 216] The original algorithm has subsequently been greatly improved to overcome especially difficult graphs [217] and handle very large sparse symmetry graphs in the SAUCY [218], BLISS [219] and NISHE [220] programs. The NAUTY and BLISS algorithms are primarily designed to solve graph isomorphism while the SAUCY algorithm focuses exclusively on discovering automorphism generators and is reported to process sparse graphs of 1 million vertices in less than 1 second. These algorithms are used in engineering applications such as very large integrated circuit design and layout.

One (largely ignored) algorithm published in the chemistry literature by Figueras [86] describes a backtrack search to discover the automorphism partition. A careful analysis of the algorithm reveals that it *implicitly* finds a small automorphism generator set. This is not preserved but the

[z] A graph invariant is a property (or set of properties) that does not depend on how the graph atoms are numbered.

[aa] This is a realistic figure when run on a 3GHz machine with 4Gb memory.

information gained is converted directly into the automorphism partition. The reported results indicate that the algorithm has potential to perform extremely well. This algorithm will be described in more detail in a following section as it provided a stepping stone towards the development of a new algorithm that efficiently finds symmetry in stereochemical graphs.

Symmetry Breaking and Isomorph Rejection

Once graph symmetry has been identified it can be used to eliminate redundant solutions in exhaustive subgraph isomorphism testing (Figure 73).

A number of general isomorph rejection algorithms have been reported and are applicable to backtrack searching. [197, 221, 222] Over the last 10 years the role of isomorph rejection and symmetry breaking has become a major topic within constraints satisfaction problem solving. [223, 224, 225] Two problems have been addressed. First, to prevent a search wastefully exploring sub trees that are symmetric to one that failed to yield a solution, and second, to prevent duplicate solutions being discovered.

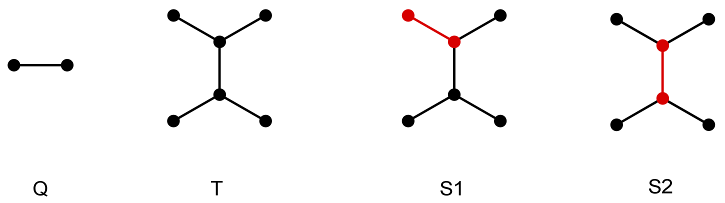
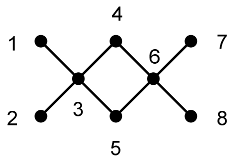


Figure 73 Isomorph rejection reduces the possible mappings of graph Q to T from 10 redundant solutions to just 2 unique solutions S1 and S2.

Two approaches known as Symmetry Breaking During Search (SBDS) [226] and Symmetry Breaking by Dominance Detection (SBDD) [227] have been applied to reduce the computational burden of the search when the problem is dominated by symmetry. The procedures assume that the automorphism group (or the generator set) of the problem representation is known in advance. Gent *et al.* have provided a detailed description of the theory behind SBDS. [226]



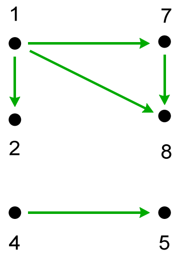
Fixed values	Stabiliser Groups	Orbits of stabilisers	Orbit Graph [219]
	Γ (see Table 11)	{ 1 2 7 8 }{ 3 6 }{ 4 5 }	
1	[1 2 3 4 5 6 7 8] [1 2 3 5 4 6 7 8] [1 2 3 4 5 6 8 7] [1 2 3 5 4 6 8 7]	{ 1 }{ 2 }{ 3 }{ 4 5 }{ 6 }{ 7 8 }	
1 2	[1 2 3 4 5 6 7 8] [1 2 3 5 4 6 7 8] [1 2 3 4 5 6 8 7] [1 2 3 5 4 6 8 7]	{ 1 }{ 2 }{ 3 }{ 4 5 }{ 6 }{ 7 8 }	
1 2 3	[1 2 3 4 5 6 7 8] [1 2 3 5 4 6 7 8] [1 2 3 4 5 6 8 7] [1 2 3 5 4 6 8 7]	{ 1 }{ 2 }{ 3 }{ 4 5 }{ 6 }{ 7 8 }	
1 2 3 4	[1 2 3 4 5 6 7 8] [1 2 3 4 5 6 8 7]	{ 1 }{ 2 }{ 3 }{ 4 }{ 5 }{ 6 }{ 7 8 }	
1 2 3 4 5	[1 2 3 4 5 6 7 8] [1 2 3 4 5 6 8 7]	{ 1 }{ 2 }{ 3 }{ 4 }{ 5 }{ 6 }{ 7 8 }	
1 2 3 4 5 6	[1 2 3 4 5 6 7 8] [1 2 3 4 5 6 8 7]	{ 1 }{ 2 }{ 3 }{ 4 }{ 5 }{ 6 }{ 7 8 }	
1 2 3 4 5 6 7	[1 2 3 4 5 6 7 8]	{ 1 }{ 2 }{ 3 }{ 4 }{ 5 }{ 6 }{ 7 }{ 8 }	

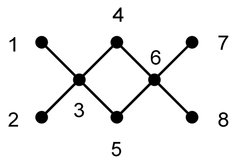
Table 12 The selection and action of stabiliser groups and orbits using a representative graph. Green values are the last fixed vertex. Red numbers are stabilised vertices. Blue numbers are eliminated symmetrical choice points. The stabilising actions are coded as an orbit graph where directed edges from vertices mark eliminated permutation choices.

The SBDS and SBDD methods require the application of orbits and stabilisers to determine the existence of symmetric choice points in a CSP solving algorithm. For example when the solver attempts to assign a value to a variable and the assignment subsequently fails then it is reasonable to exclude testing the equivalent values, as assignments using them must also fail. The variable and value equivalents at any stage of the search are determined from the orbits of stabiliser groups fixed by the currently assigned variables and values.

Table 12 sets out the processing steps for an example graph representing problem symmetries to illustrate the SBDS procedure described by Gent. The discussion that follows assumes these are value symmetries, but would equally apply to variable symmetries.

When the CSP solver is started no value is selected (*i.e.* none is fixed by assignment to a variable) hence the value equivalences are the orbits of the full automorphism group. For convenience it is assumed that the order the values are assigned to the problem variables is simply 1, 2, 3 … 8 and this is used to follow how the symmetries break down under partial assignment.

After the first value 1 is fixed the automorphism group is reduced to the stabiliser subgroup that fixes 1[bb]. The value equivalents are now represented by the orbits of the stabiliser group for 1, being {1} {2} {3} {4 5} {6} {7 8}. As further values become fixed a stabiliser chain develops occasionally reducing the size of the equivalence sets. It should be noted that placing the variables in a different order results in the creation of different stabiliser chains, but the final outcome is always equivalent.



	1	2	3	4	5	6	7	8
1	0	2	1	2	2	3	4	4
2	2	0	1	2	2	3	4	4
3	1	1	0	1	1	2	3	3
4	2	2	1	0	2	1	2	2
5	2	2	1	2	0	1	2	2
6	3	3	2	1	1	0	1	1
7	4	4	3	2	2	1	0	2
8	4	4	3	2	2	1	2	0

Distance matrix

Fixed vertex	Orbit splits on unfixed vertices
1	{ 1 (0), 2 (2), 7 (4), 8 (4) } { 3 (1), 6 (3) } { 4 (2), 5 (2) }
2	{ 2 (0) } { 7 (4), 8 (4) } { 3 (1) } { 6 (3) } { 4 (2), 5 (2) }
3	{ 7 (3), 8 (3) } { 3 (0) } { 6 (2) } { 4 (1), 5 (1) }
4	{ 7 (2), 8 (2) } { 6 (1) } { 4 (0), 5 (2) }
5	{ 7 (2), 8 (2) } { 6 (1) } { 5 (0) }
6	{ 7 (1), 8 (1) } { 6 (0) }
7	{ 7 (0), 8 (2) }
8	{ 8 (0) }
	Key: { *vertex* (*distance from last fixed vertex*), … }

Table 13 **Using the distance matrix to calculate a stabiliser chain for the representative graph. Vertices are fixed in the sequence 1,2,… 8. Underlined values identify the orbits that are split when members have non-equal distances from the last fixed vertex. Blue values identify eliminated symmetry choice points. Green values are the last fixed vertices from which distances are measured.**

[bb] Typically solved by calling functions from the GAP library - http://www.gap-system.org/ .

The next consideration is to establish the symmetrical choice points from the orbits derived from the stabiliser chain. When a search fails and a backtracking step is taken, a trial value is unbound (unfixed) from a variable and the last orbit in the chain is discarded. The unbound value is noted and if any equivalents of that value exist in the current orbit, those equivalent values are eliminated from the search. These points are marked in blue in Table 12.

The identified symmetric choice points are transformed into an orbit graph [228] where directed edges connecting variables identify the redundant choice points. Table 12 shows that the orbit graph for the value fixing sequence 1, 2, 3, 4, 5, 6, 7, 8 establishes that the following partial sequences would lead to redundant symmetric results; 2 …, 7 …, 8 …, 1,2,3,5 …, and 1,2,3,4,5,6,8, … . Solutions beginning with these sequences can be eliminated.

The use of an orbit graph can be substituted by using the graph distance matrix [229] to compute an orbit chain *during* the backtrack search. The distance matrix *D* of a labelled graph *G(V, E)* is a symmetric square matrix of dimensions N = |V|. The elements *d(u,v)* of the matrix are defined as:

d (u, u) = 0

d (u, v) = {N if no path exists between u, v; otherwise the shortest path between u and v}

Note that *N* is used as a convenience value when *u* and *v* are not connected as the longest possible path in any connected graph is *N – 1*. When *u,v* are connected the distance is the number of edges in the shortest path. This matrix is easily calculated in polynomial time. [230, 231]

Table 13 includes the distance matrix of the example graph as an aid to following the stepwise procedure to form the orbit chain. Starting with the automorphism partition of the graph, we fix the first vertex *v* when it is assigned to a variable. Each of the vertex members *u* of the partition are now labelled with the distance *d(u,v)* from the fixed vertex. Singleton orbits in the partition are ignored in this procedure as they cannot be split further. All distance labelled members of an orbit are now subdivided into new orbits according to equivalent distance labels. The distance labelling and orbit splitting procedure is repeated when each vertex is fixed in the search.

The Figueras Automorphism Partition Algorithm

Figueras devised an apparently efficient algorithm that was specifically designed to precisely solve the graph automorphism partitioning problem in reasonable time while avoiding the issues found with relaxation methods. [86] The algorithm incorporates a backtrack search. The steps of the algorithm are shown in Listing 7 (*vide infra*).

The essential data structures manipulated by the algorithm are a pair of square bit matrices DOMAIN and EQUIV, each dimensioned by the number of atoms. DOMAIN is the set of initial candidate atoms for each atom. When the algorithm completes EQUIV will redundantly record the equivalent atoms for each atom and thus represents the automorphism partition.

The procedure relies on two standard functions, a version of the Morgan algorithm (named "*Morgan"*) that is used to induce an initial partition for each atom in the DOMAIN matrix (line 1) and a simple backtracking search algorithm (named "*Automorphism"*) that can choose atoms from DOMAIN to either generate an automorphism map or indicate that an automorphism is absent (line 6). Initialisation is completed by setting the initial entries of the matrix EQUIV such that each atom is initially equivalent only to itself.

The algorithm proceeds to enter a loop that tries the following: a pair of atoms *a, k* are selected in turn under the condition that *k* is not already known to be equivalent to *a* (line 5); and, if no more non-equivalent atom candidates exist for *a,* then update *a* to be the next atom (line 3). Next, temporarily force *k* to be the only domain candidate for *a* (the remaining candidates are saved to be restored later; line 5) and see if there is an automorphism that exists when *a* ↔ *k* (line 6). [cc] Any automorphism that is found is preserved in MAP.

```
 1.  FOR ALL: a ∈ ATOMS (G) DOMAIN (a) ← Morgan (a)           ( initialisation – G is the graph)
 2.  FOR ALL: a ∈ ATOMS (G) EQUIV (a) ← {a}                 ( initialisation –  equivalent  to self)

 3.  FOR ALL: a ∈ ATOMS (G)                                  ( visit each atom in the graph)
 4.    WHILE: DOMAIN (a) ≠ EQUIV (a)
 5.       k ←  ∈ (DOMAIN (a) – EQUIV (a)),  temp ← DOMAIN (a), DOMAIN (a) ←  {k}
 6.       MATCH, MAP  ← Automorphism (G, DOMAIN)                  ( a ↔ k forced in domain)
 7.       IF: MATCH is true                                ( automorphism exists for a ↔ k)
 8.          FOR ALL: m ∈ ATOMS (G)
 9.              n  ← MAP (m)
10.              IF: n  ∉ EQUIV (m)
11.                 EQUIV (m) ← EQUIV (m) ∪ {n}                            ( add n to set)
12.                 FOR ALL: e ∈ EQUIV (m)
13.                    EQUIV (e) ← EQUIV (e) ∪ EQUIV (m)
14.       ELSE:                                            ( automorphism absent for a ↔ k)
15.          DOMAIN (a) ← DOMAIN (a) – {k}                            ( remove k from set)
16.          FOR ALL: m ∈ ATOMS (G)
17.              IF: m ∈  EQUIV (a)
18.                 DOMAIN (m) ← DOMAIN (m) – {k}                         ( remove k from set)
19.    DOMAIN (a) ← temp
```

Listing 7 The Figueras Automorphism Partition algorithm.

[cc] ↔ is used to mean "equivalent to".

If an automorphism exists (line 7) then the returned map is scanned (lines 8 – 13) and for each atom $\underline{m}$, its EQUIV entry is updated with the atom $\underline{n}$ that was mapped to $\underline{m}$ (line 11). Now that it is known that $\underline{m} \leftrightarrow \underline{n}$, update all known equivalents of $\underline{m}$ (including $\underline{n}$) with the equivalents of $\underline{m}$ (line 13) to act on the equivalence relation that if $x \leftrightarrow y$ and $y \leftrightarrow z$ then $x \leftrightarrow z$. The outcome is the partial symmetry partition, stored in EQUIV, is updated with all information gained by the newly discovered automorphism.

If the equivalence mapping $\underline{a} \leftrightarrow \underline{k}$ does not produce an automorphism then eliminate $\underline{k}$ from the domain of $\underline{a}$ (line 15) and from all domains of atoms known to be equivalent to $\underline{a}$ (lines 16 – 18). Finally, restore the saved domain of $\underline{a}$ (line 19) to allow the next qualifying $\underline{a}$, $\underline{k}$ pair to be selected. The algorithm halts when all qualifying $\underline{a}$, $\underline{k}$ pairs have been tried.

The essential feature of this algorithm is that most of the possible $\underline{a}$, $\underline{k}$ pairs that lead to a redundant automorphism are pruned away as incremental partial symmetry is discovered. Information deposited in EQUIV is used to edit increasingly larger sections of DOMAIN so this accelerates the reduction of automorphism tests. It appears that most, but possibly not all, automorphisms that are compositions of earlier discovered automorphisms are eliminated by this procedure and this is the origin of the good performance quoted by Figueras.

The Figueras algorithm was implemented as an initial candidate component for the construction of an isomorph rejection algorithm. This reference implementation also provided a performance baseline for improvements. The experimental implementation made the following changes to that described by Figueras. The *Morgan* and *Automorphism* functions where directly replaced by the stereochemical CSP solver. The arc consistency test, look-ahead editing and predicate evaluating procedures operated by the CSP solver served to set the reduced domains. In addition the handling of stereochemical constraints, already implemented by the CSP solver, could be used to advantage to discover symmetries in stereodefined molecules. Direct access to the CSP solver value domains allowed the Figueras algorithm to control CSP solver domain editing when an automorphism did not exist for $\underline{a}$, $\underline{k}$. Results from running this implementation are presented in the results section.

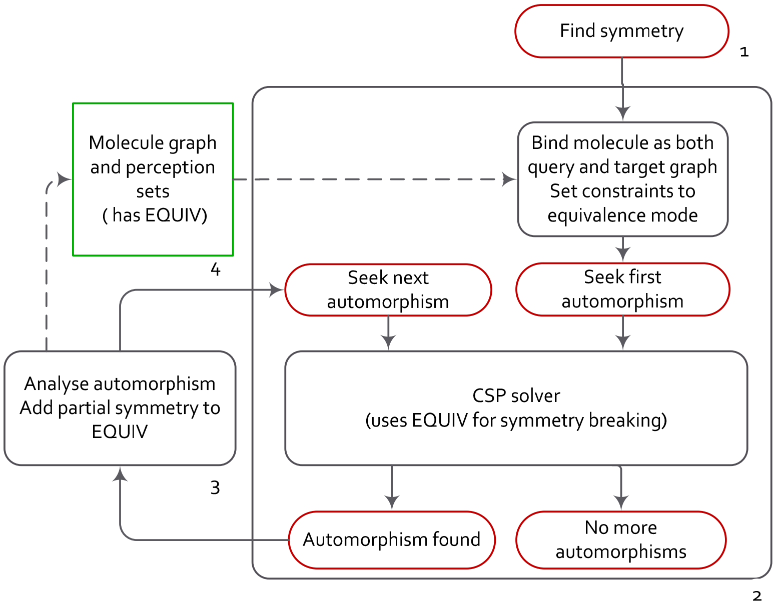
A Novel Isomorph Rejection Algorithm

An isomorph rejection algorithm must work with the *a priori* knowledge of the symmetries of the two input graphs Q and T [dd]. Conversely the automorphism generator developed by Figueras starts out with no knowledge of the input graph symmetry but as partial symmetries are found

[dd] Q is a query molecule, reaction or structural pattern, T is a target molecule or reaction.

through automorphisms, the algorithm improves on its ability to reject subsequent redundant automorphisms. This implies that the selection of an automorphism generator set is just a special case of the isomorph rejection problem. This observation indicated that both symmetry perception and isomorph rejection could be implemented using the same stereochemical CSP solver. This idea was the impetus to devise a new approach.

When solving a Q → T isomorph problem, the CSP solver can invoke two new instances of itself in an automorph configuration [ee] to discover Q and T symmetries. Used in combination with an isomorph rejection procedure embedded within the CSP solver, it provides a means to generate isomorph free solutions. The symmetries of the input graphs are needed only if at least one Q → T solution is found. If a Q → T match was not forthcoming then no knowledge of Q or T symmetry is required as there are no isomorphs to reject.


Find symmetry
1
Bind molecule as both query and target graph
Set constraints to equivalence mode
Molecule graph and perception sets (has EQUIV)
4
Seek next automorphism
Seek first automorphism
Analyse automorphism
Add partial symmetry to EQUIV
CSP solver
(uses EQUIV for symmetry breaking)
3
Automorphism found
No more automorphisms
2

Flow chart 1 **The symmetry perception routine.**

[ee] By configuration it means the selection of the Q and T graph type in conjunction with appropriate implementations of the unary and binary constraint solvers required to map Q → T (an isomorph problem) and Q → Q or T → T (an automorph problem). The constraint solvers will differ depending on whether Q is a molecule, reaction, or a structural pattern or if it is solving a graph automorphism (equivalent structure or pattern match), graph isomorphism (exact structure match) or subgraph isomorphism (substructure match).

The isomorph rejection algorithm is separated into two parts, a symmetry perception module that uses the CSP solver to find automorphisms in Q and T and a symmetry breaking module embedded into the CSP solver to control backtracking and edit out redundant branches of the search tree.

Flow chart 1 outlines the structure of the symmetry perception module. A CSP solver is set to discover automorphisms (box 2) by setting the same molecule to act as the query and the target. The constraints are set for solving atom, bond and stereocentre syntax *equivalence*. For example special atom types such as X (halogen) have candidates atoms selected *only* from other X atoms and not specific atom types such as F, Cl, Br or I which are otherwise inferred as members of X. This ensures that query pattern symmetry is treated in the same manner as molecule or reaction symmetry[ff]. This equivalence constraint is also applied to PATRAN expressions (*vide supra*) such that the expressions themselves must lexically match at the syntax level. The CSP solver is called to find a series of automorphisms, each of which is analysed by the perception executive (box 3). This executive incrementally adds perceived symmetry information to an EQUIV matrix (box 4) stored with the molecule that is undergoing analysis. Internally the CSP solver uses the symmetry breaking or isomorphism rejection procedures to eliminate redundant

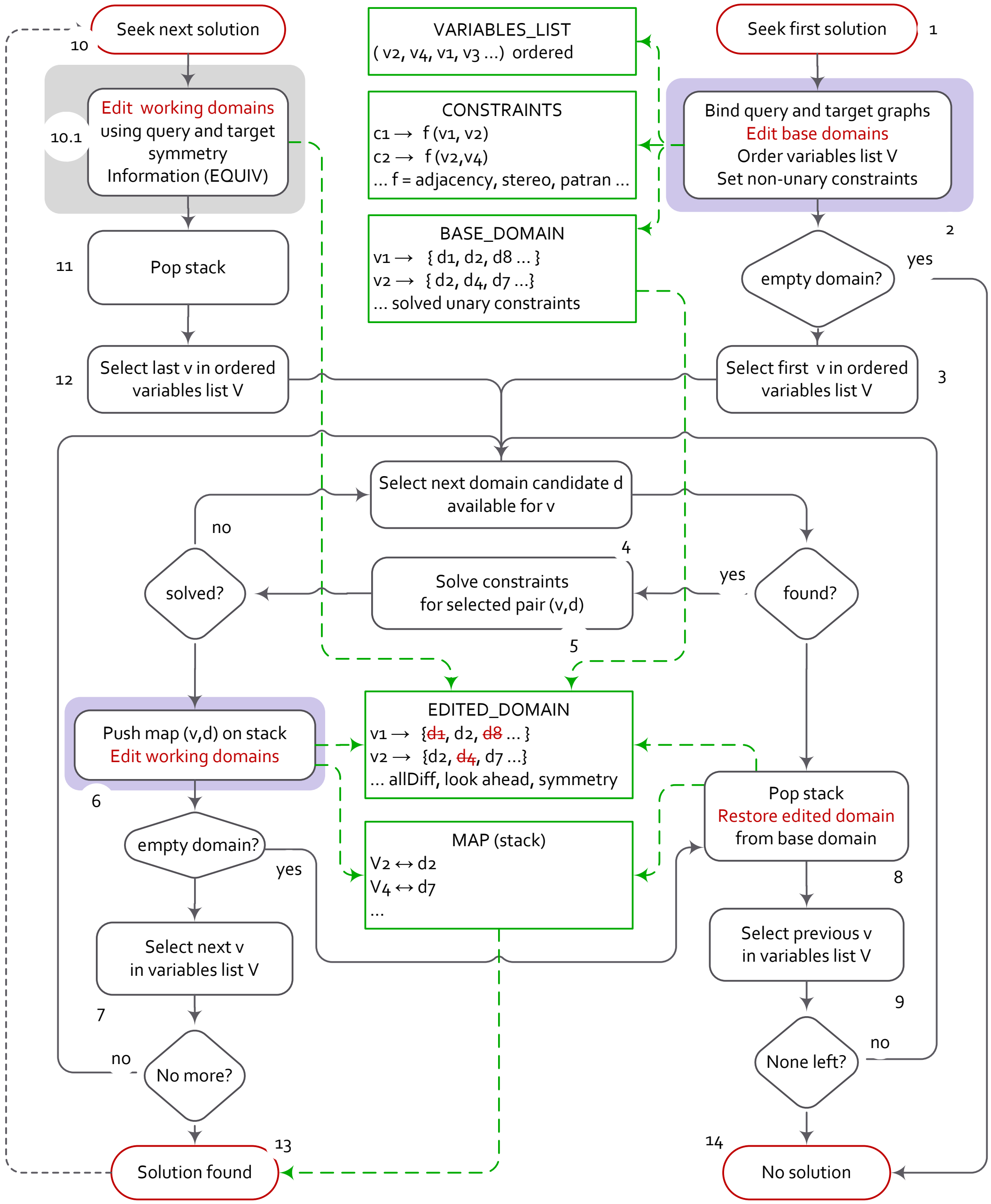
```
FOR ALL: $a \in$ ATOMS (G)
    EQUIV (a) $\leftarrow \varnothing$                                     ( initialisation )

MATCH $\leftarrow$ true
WHILE: MATCH is true
    (MATCH, MAP) $\leftarrow$ Automorphism (G, EQUIV)        ( a CSP solver set for automorphism)

    IF: MATCH is true                                         ( automorphism found )
        atoms $\leftarrow$ ATOMS (G)
        FOR ALL: $m \in$ atoms
            n $\leftarrow$ MAP (m)
            WHILE: $n \neq m$                                   (trace the orbit cycle)
                EQUIV (m) $\leftarrow$ EQUIV (m) $\cup$ EQUIV (n) $\cup$ { n }
                atoms $\leftarrow$ atoms – { n }      (no need to visit n in future as it has been processed)
                n $\leftarrow$ MAP (n)

            EQUIV (m) $\leftarrow$ EQUIV (m) – { m }          (the equivalents of m do not include itself)

            FOR ALL: $e \in$ EQUIV (m)          (all equivalents of m are mutually equivalent with
                                                                              each other)
                EQUIV (e) $\leftarrow$ EQUIV (e) $\cup$ EQUIV (m) $\cup$ { m } – { e }
```

Listing 8 The Analyse Automorphm procedure for incrementally finding atom equivalents (orbits) from an automorphism generator set.

[ff] This is the symmetry of the pattern syntax and not its semantics.

automorphisms. These procedures accesses the EQUIV symmetry information bound to the CSP solver at problem initialisation.

Listing 8 shows the procedure detailing the operation of the symmetry perception algorithm presented in Flow chart 1. The automorphism analysis section (box 3) is used to update EQUIV


Seek next solution
10
Edit working domains using query and target symmetry Information (EQUIV)
10.1
Pop stack
11
Select last v in ordered variables list V
12
VARIABLES_LIST
(v2, v4, v1, v3 ...) ordered
CONSTRAINTS
c1 ⟶ f (v1, v2)
c2 ⟶ f (v2,v4)
... f = adjacency, stereo, patran ...
BASE_DOMAIN
v1 ⟶ { d1, d2, d8 ... }
v2 ⟶ { d2, d4, d7 ...}
... solved unary constraints
Seek first solution
1
Bind query and target graphs
Edit base domains
Order variables list V
Set non-unary constraints
2
empty domain?
yes
Select first v in ordered variables list V
3
Select next domain candidate d available for v
no
solved?
4
Solve constraints for selected pair (v,d)
5
yes
found?
EDITED_DOMAIN
v1 ⟶ {~~d1~~, d2, ~~d8~~ ... }
v2 ⟶ {d2, ~~d4~~, d7 ...}
... allDiff, look ahead, symmetry
Push map (v,d) on stack
Edit working domains
6
empty domain?
yes
MAP (stack)
V2 ↔ d2
V4 ↔ d7
...
Pop stack
Restore edited domain from base domain
8
Select next v in variables list V
7
Select previous v in variables list V
9
no
No more?
None left?
no
13
Solution found
14
No solution

Flow chart 2 The location of symmetry breaking modification made to the CSP solver: forward breaking (lilac); reverse breaking (grey).

and is executed by lines 7 – 16. The analysis is similar to that described by Figueras but has been improved by directly tracing the full orbit of each atom from the automorphism map (MAP). For example if we have the solution map, 1 → 4, 2 → 1, 3 → 2, 4 → 3, we can transform this to a set of closed orbits by following each thread through the map and connecting the ends. For example the map contains the single thread 1 → 4, 4 → 3, 3 → 2, 2 → 1 which yields the orbit (1 4 3 2).

The post-match analysis proceeds in two phases: finding the new equivalents of an atom from the automorphism map by the thread tracing method (lines 9 – 14); propagating the known equivalents of atom m to each (new) equivalent of m as all equivalents sharing an orbit must be in mutual agreement (lines 15 – 16).

The symmetry pruning components are implemented within the CSP solver module. The areas requiring modification are identified with lilac shading for forward symmetry breaking and grey shading for reverse symmetry breaking in the CSP solver flow chart (Flow chart 2).

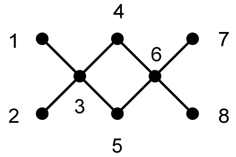
Two complementary groups of modifications where made: symmetry breaking logic that acts during forward tracking steps (Flow chart 2, boxes 2 and 6); and symmetry breaking logic that acts only on backtracking steps when seeking secondary solutions (Flow chart 2, box 10.1). The former optimises searches by avoiding redundant exploration within symmetrical parts of the target graph that are proven to lead to failures. The latter accounts for symmetry in both the query and target graphs and avoids redundant solutions and implements full isomorph rejection.

If forward symmetry breaking is enabled the value domain of the first query atom (BASE_DOMAIN) is permanently edited to preserve just one representative target atom of each symmetry class during initialisation (Flow chart 2, box 2). This is permissible because at this point <u>no</u> query-target atom pair has been fixed. Once the search is entered a fresh stabiliser chain must be calculated on each forward tracking step for every query atom bar the first (Flow chart 2, box 6). This procedure uses the distance splitting algorithm (*vide supra*) and the edited equivalence classes are used to remove symmetric choice points from the working domain matrix (EDITED_DOMAIN). Both these domain editing steps require *a priori* knowledge of the symmetry of the query and target graphs so symmetry perception must be performed before the first solution is sought.

```
FOR ALL: a ∈ ATOMS (G)                                              ( G is the query graph)
    m ← MAP (a)                                                 ( MAP is the last CSP solution)
    EDITED_DOMAIN (a)  ← EDITED_DOMAIN (a) – EQUIV (m) – { m }          ( target symmetry)
    FOR ALL: e ∈ EQUIV (a)                                               ( query symmetry)
        EDITED_DOMAIN (a) ← EDITED_DOMAIN (a) –  { MAP (e) }
```

Listing 9 The Symmetry pruning procedure used to reject isomorph solutions.

Reverse symmetry breaking is invoked only after the CSP solver is re-entered to find secondary



	1	2	3	4	5	6	7	8	← Vertices
Step	{1 2 7 8} ↓	{1 2 7 8} ↓	{3 6} ↓	{4 5} ↓	{4 5} ↓	{3 6} ↓	{1 2 7 8} ↓	{1 2 7 8} ↓	← Base Domains
A	{1 2 7 8} • →	{~~1~~ 2 7 8} →	{3 6} →	{4 5} →	{4 5} →	{~~3~~ 6} →	{~~1 2~~ 7 8} →	{~~1 2 7~~ 8} → •	← Edited Domains
	1	2	3	4	5	6	7	8	

B							{~~1 2 7~~ 8}
							←
							{~~1 2 7~~ 8}
							→
								{~~1 2 7 8~~}
								← •
								{~~1 2~~ 7 ~~8~~}
								→ •
									Orbits found ↓

	1	2	3	4	5	6	8	7	{7 8}

C				{4 5}
				←
				{4 5}
				→
					{~~4 5~~}
					←
					{4 ~~5~~}
					→
						{~~3 6~~}
						←
						{~~3~~ 6}
						→
							{~~1 2 7 8~~}
							←
							{~~1 2~~ 7 8}
							→
								{~~1 2 7 8~~}
								← •
								{~~1 2 7~~ 8}
								→ •	

	1	2	3	5	4	6	7	8	{4 5} {7 8}

D	{~~1~~ 2 7 8}
	←
	{~~1~~ 2 7 8}
	→
		{~~1 2 7 8~~}
		←
		{1 ~~2~~ 7 8}
		→
			{~~3 6~~}
			←
			{3 6}
			→
				{~~4 5~~}
				←
				{4 5}
				→
					{~~4 5~~}
					←
					{4 5}
					→
						{~~3 6~~}
						←
						{~~3~~ 6}
						→
							{~~1 2 7 8~~}
							←
							{~~1 2~~ 7 8}
							→
								{~~1 2 7 8~~}
								← •
								{~~1 2 7~~ 8}
								→ •	

	2	1	3	4	5	6	7	8	{1 2} {4 5} {7 8}

E	{~~1 2~~ 7 8}
	←
	{~~1 2~~ 7 8}
	→
		{~~1 2 7 8~~}
		←
		{~~1 2 7~~ 8}
		→
			{~~3 6~~}
			←
			{~~3~~ 6}
			→
				{~~4 5~~}
				←
				{4 5}
				→
					{~~4 5~~}
					←
					{4 5}
					→
						{~~3 6~~}
						←
						{3 ~~6~~}
						→
							{~~1 2 7 8~~}
							←
							{1 2 ~~7 8~~}
							→
								{~~1 2 7 8~~}
								← •
								{~~1~~ 2 ~~7 8~~}
								→ •	

	7	8	6	4	5	3	1	2	{1 2 7 8} {3 6} {4 5}
F	{~~1 2 7 8~~} • ←	{~~1 2 7 8~~} ←	{~~3 6~~} ←	{~~4 5~~} ←	{~~4 5~~} ←	{~~3 6~~} ←	{~~1 2 7 8~~} ←	{~~1 2 7 8~~} ← •	

Key: • → start, → forward step, ← backtrack, → • solution found, ← • restart search, • ← finished

{1 2}	Base Domain. On each (→) copied to Edited Domain then updated
1 2	Domain values available
1	Domain value picked
~~1 2~~	Values struck out by the allDifferent constraint on (→). Revoked by (←)
~~1 2~~	Values struck out by Symmetry constraints on (← •). Revoked by next (→)
~~1 2~~	Values struck out by Adjacency constraints
1 2 3	A solution

Figure 74 A trace showing the generation an automorphism generator set.

solutions (Flow chart 2, box 10.1). The entire working domain matrix (EDITED_DOMAIN) containing the candidate target atoms is pruned in one overall step. The symmetry equivalents of the target atom mapped to each query atom are removed from the query atom domain (Listing 9 line 3), followed by removing the target atoms mapped to the symmetry equivalents of the query atom (Listing 9 lines 4 – 6). This accounts for symmetries present in both query and target graphs.

The domain edit using partial or full symmetries is permissible as all query atoms are currently fixed at re-entry. The search algorithm can now backtrack as far as it needs to find the first non-symmetric choice. Once a non-symmetric choice is made and the CSP forward tracks it is no longer permissible to retain the symmetry edits as a new stabiliser chain is now in charge. The working domain (EDITED_DOMAIN) is reset to a copy of the base domain and reedited to apply all active constraints.

The operation of the isomorph rejection procedure is best understood by following a trace. Figure 74 demonstrates the construction of a small automorphism generator set for a representative graph.

The base domain is partitioned using the vertex degree to group atoms with 1, 2 and 4 non-hydrogen connections. Thereafter a base domain is used to load the corresponding edited domain each time a forward step is made. This is followed by removal of all currently assigned atoms (the allDiff constraint).

The first automorphism solution found (row A) is in always the identity automorphism. This gains no new information so on re-entry (row B) no symmetry editing occurs. The back track step continues until the next available choice (7 → 8) is made and the forward steps lead to the first non-trivial solution. The automorphism found reveals that 7 and 8 are equivalent. This symmetry information is used on the next re-entry (row C) to force the search back to the choice point (4 → 5) by the temporary removal of symmetrical target atoms which are reinstated after the (4 → 5) assignment. This process continues gathering new partial symmetries until all back track choices are eliminated by symmetry pruning and the search ends (row F). The order of the found generator set is 4, which is significantly smaller than the automorphism group (16) but a little larger than the smallest basis generator set (3).

The isomorph rejection algorithm can be combined with symmetry perception in two configurations. Figure 75 illustrates these possibilities with a procedure designed to convert a target molecule to a set of precursor molecules by the application of a reaction rule. Configuration A is capable of optimising symmetric search paths when attempting to find the first retron match. However the cost of finding rule and target molecule symmetries may dominate the time taken to find the first solution. Configuration B overcomes this cost penalty by delaying symmetry perception until the second and subsequent attempts to find a unique retron match. At this point symmetry perception is mandatory if isomorph rejection is required.

The choice between these two configurations in terms of overall efficiency can only be determined by experimentation. Configuration A may be practical when the perception costs are amortised, either because reaction rule symmetries are cached for later reuse or when many rules are applied to each target molecule. It may be beneficial to process all reaction rules off-line and find and record the automorphism generator sets with the rules.

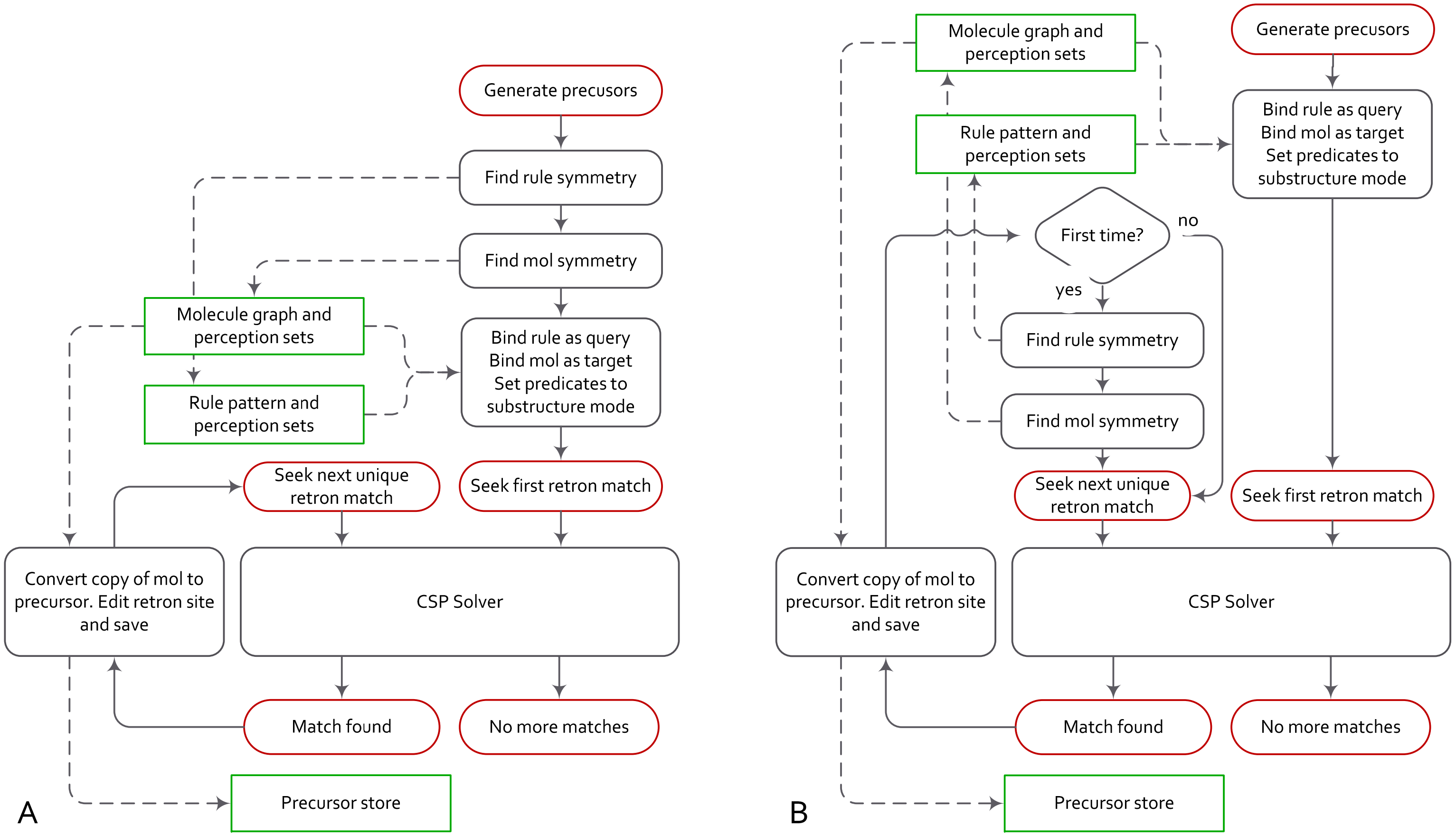

Generate precusors
Find rule symmetry
Find mol symmetry
Molecule graph and perception sets
Rule pattern and perception sets
Bind rule as query Bind mol as target Set predicates to substructure mode
Seek next unique retron match
Seek first retron match
Convert copy of mol to precursor. Edit retron site and save
CSP Solver
Match found
No more matches
Precursor store
A
Molecule graph and perception sets
Generate precusors
Rule pattern and perception sets
Bind rule as query Bind mol as target Set predicates to substructure mode
First time?
no
yes
Find rule symmetry
Find mol symmetry
Seek next unique retron match
Seek first retron match
Convert copy of mol to precursor. Edit retron site and save
CSP Solver
Match found
No more matches
Precursor store
B

Figure 75 **Two CSP Solver configurations for performing non-redundant precursor generation in retrosynthetic analysis:**
A - Supporting all matches with eager symmetry perception; B - Supporting only subsequent matches with lazy symmetry perception.

Results and Discussion

All experiments were run on an Intel Core2 Duo machine running at 2.83 GHz with 3.37 Gb of available RAM under the Windows XP (SP3) operating system. Timings were performed with the Windows high precision timer functions *QueryPerformanceCounter* and *QueryPerformanceFrequency* which allow timed intervals to be measured at a resolution close to 1 nanosecond with the CPU operating at 2.83 GHz. All timed intervals were reproducible on multiple runs to within 100 nanoseconds.

The FG series of 28 test molecules and graphs (Figure 76) are those selected by Figueras [86] and have in turn been collected from a variety of papers that used them to test various symmetry determining algorithms. A proportion of these where devised as counter examples that exposed failing cases for the Morgan and related algorithms. Others are classes of graph which are known to be generally difficult to process or have features such as isospectral points. The SS series (Figure 77) was devised to observe the behaviour of the symmetry finding algorithm as graphs are systematically grown with added atoms and increasing symmetry reaching up to very large automorphism groups (SS5 – SS24). The series is completed with some cyclic regular graphs (SS25 - SS29), wheel graphs (SS30 - SS34) and some interesting mesh graphs (MK1, AR1). Not all graphs are chemically realistic and are added to test the algorithm beyond the boundaries in which it is expected to operate.

The Morgan algorithm used is a custom version designed to generate graph hash keys and does not preserve any symmetry information. It is included for comparative timing purposes and acts as a representative for the relaxation class of algorithm. As backtrack algorithms are expected to have poorer performance than the relaxation algorithms, it serves to benchmark a baseline for the experiments.

All implemented algorithms were run in an environment that supports lazy evaluation of rings and stereochemistry. As a consequence the timings include the time to perform ring perception when cycles are present.

The following pre-computed values are included with the results (Table 14, Table 15); *n* – the number of atoms in the graph; *Classes* – the size of the automorphism partition; *Order* – the size of the automorphism group.

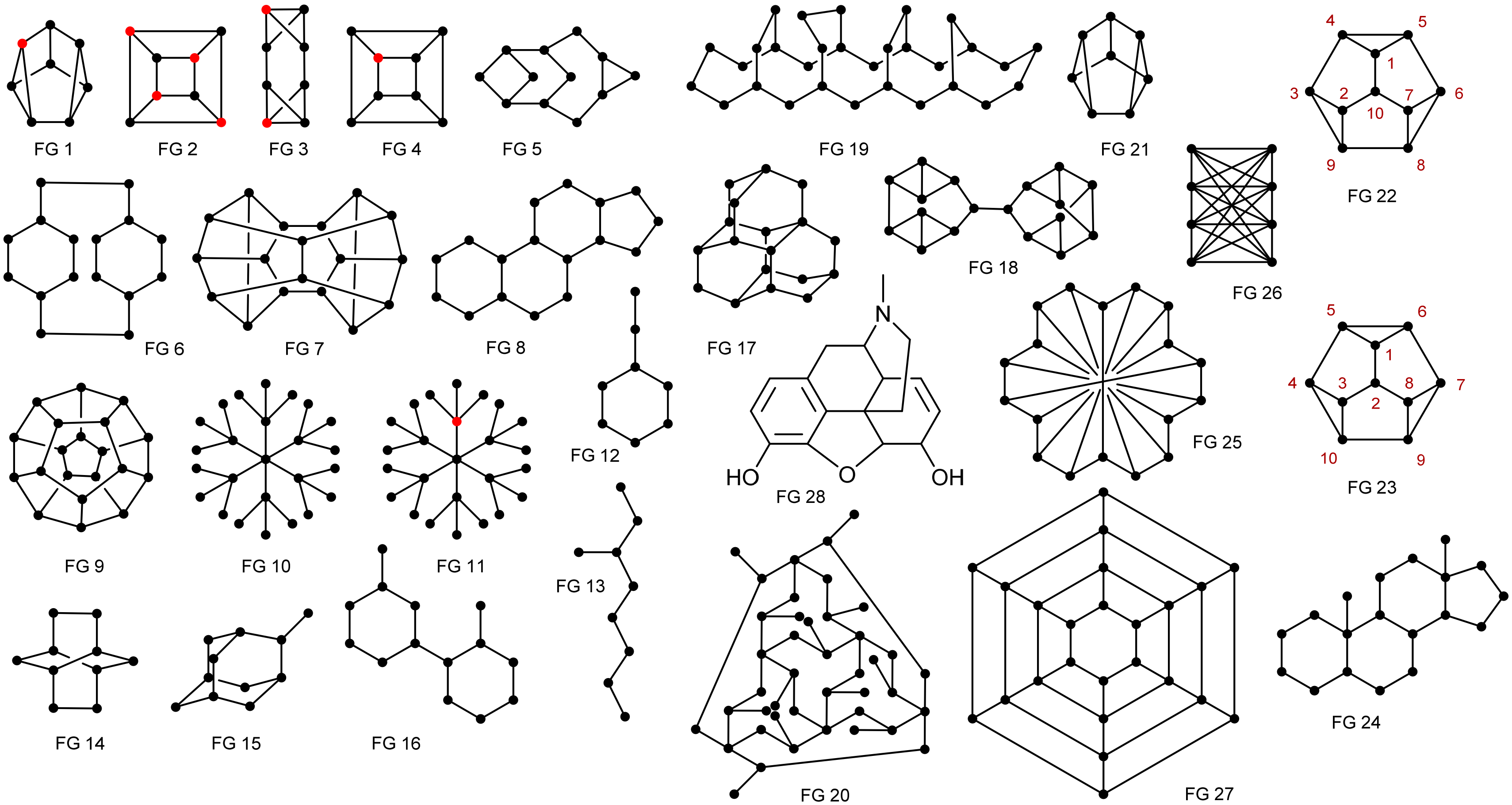


Figure 76 Challenge graphs FG1 – FG28 are designed to test symmetry perception.

Structure	n	Classes	Order	Generator Size		CPU time (ms)					Performance ratio	
				Figueras	IR	Morgan algorithm	Figueras algorithm	Isomorph Rejection (IR) algorithm			Morgan / B	Figueras / B
								A	B	C		
FG 1	8	8	1	1	1	1.4	2.3	2.0	2.0	0.7	0.70	1.15
FG 2	8	2	24	4	4	1.2	0.7	2.6	1.9	0.8	0.63	0.37
FG 3	8	3	4	3	3	1.4	2.7	2.1	2.1	0.8	0.67	1.29
FG 4	8	4	6	3	3	1.2	2.9	2.0	1.8	0.8	0.67	1.61
FG 5	12	7	4	3	3	2.0	9.2	2.8	2.8	0.9	0.71	3.29
FG 6	16	3	16	5	5	3.3	17.3	4.9	4.4	1.2	0.75	3.29
FG 7	18	6	4	3	3	14.0	15.0	15.7	15.6	1.7	0.90	0.96
FG 8	17	17	1	1	1	2.0	19.4	3.1	3.1	1.2	0.65	6.26
FG 9	20	1	120	7	4	10.3	19.0	26.1	12.3	1.9	0.84	0.65
FG 10	25	3	33592320	18	18	0.3	4.7	760569.2	2.5	2.3	0.12	1.88
FG 11	25	5	5598720	-	17	0.3	7.7	116974.2	2.3	2.2	0.13	3.35
FG 12	8	6	2	2	2	0.1	1.5	0.6	0.6	0.5	0.17	2.50
FG 13	9	9	1	1	1	0.1	0.3	0.5	0.5	0.5	0.20	0.60
FG 14	10	3	4	3	3	1.4	1.1	2.1	2.1	0.8	0.67	0.52
FG 15	11	6	4	3	3	1.3	5.5	2.1	2.1	0.9	0.62	2.61
FG 16	14	14	1	1	1	1.1	0.8	1.8	1.8	0.8	0.61	0.44
FG 17	18	8	4	4	3	4.4	2.7	5.9	5.7	1.5	0.77	0.47
FG 18	18	8	32	5	6	4.6	11.9	7.6	6.2	1.9	0.74	1.92
FG 19	30	16	2	2	2	7.9	3.1	9.5	9.5	1.8	0.83	0.34
FG 20	42	14	3	3	2	12.0	175.7	15.4	14.9	4.2	0.81	11.79
FG 21	8	3	4	3	3	1.4	3.1	2.6	2.6	1.4	0.53	1.19
FG 22	10	3	6	3	3	2.6	6.3	3.2	3.0	1.0	0.87	2.10
FG 23	10	3	6	3	3	2.2	7.8	3.2	3.1	1.0	0.71	2.52
FG 24	19	19	1	1	1	2.1	2.4	3.1	3.1	1.0	0.68	0.77
FG 25	18	1	36	6	4	6.7	13.2	23.8	20.4	1.7	0.33	0.65
FG 26	8	2	8	4	3	12.2	4.8	13.6	13.3	1.0	0.92	0.36
FG 27	24	2	24	7	4	49.6	59.0	54.1	52.2	2.4	0.95	1.13
FG 28	21	21	1	1	1	2.6	1.5	3.8	3.8	1.2	0.68	0.39

Table 14 The performance of various symmetry perception algorithms on a set of 'challenge' graphs FG1 to FG28. n is the number of atoms; IR is the Isomorph Rejection algorithm; IR configuration A generates the full automorphism group and *includes* ring perception time; IR configuration B finds a small automorphism generator set and *includes* the ring perception time; IR configuration C finds a small automorphism generator set and *excludes* the ring perception time.

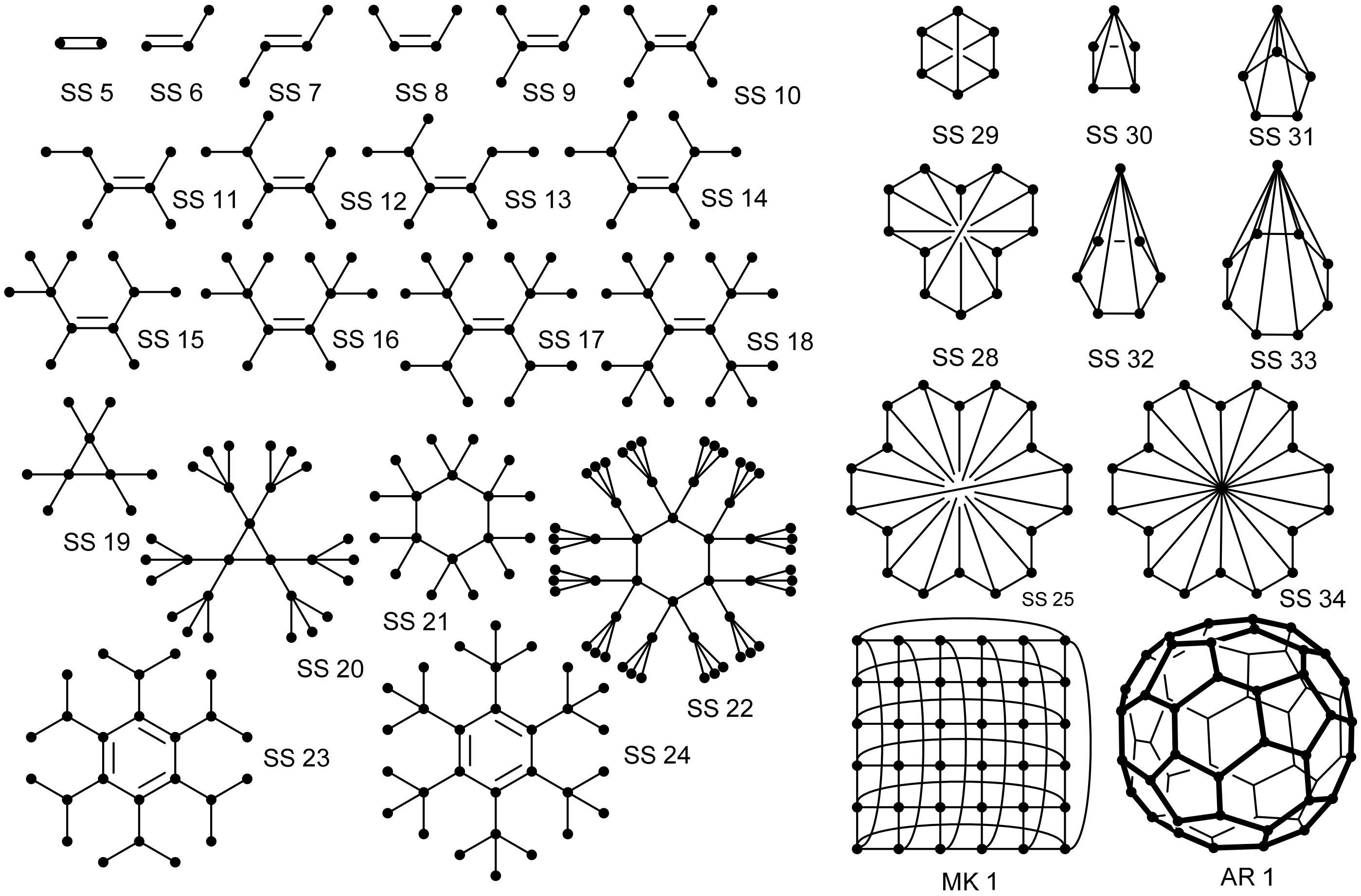


Figure 77 A series of graphs leading to very large automorphism groups (SS5 – SS24). Regular graphs; wheel and Moebius graphs (SS25 – SS34); mesh graphs (MK1,AR1).

Structure	n	Classes	Order	Generator Size	CPU time (ms)				Performance ratio
				IR algorithm	Morgan algorithm	Isomorph Rejection (IR) algorithm			Morgan / B
						A	B	C	
SS 5	2	1	2	2	< 0.1	0.3	0.3	0.2	<0.33
SS 6	3	3	1	1	0.1	0.3	0.3	0.3	0.33
SS 7	4	2	2	2	0.1	0.4	0.4	0.3	0.25
SS 8	4	2	2	2	0.1	0.4	0.4	0.3	0.25
SS 9	5	5	1	1	0.1	0.4	0.4	0.4	0.25
SS 10	6	2	4	3	0.1	0.8	0.6	0.6	0.17
SS 11	7	7	1	1	0.1	0.5	0.5	0.5	0.20
SS 12	8	7	2	2	0.1	0.7	0.7	0.6	0.14
SS 13	9	8	2	2	0.1	0.6	0.6	0.6	0.17
SS 14	10	4	8	4	0.2	0.7	0.7	0.6	0.29
SS 15	11	8	12	4	0.3	0.8	0.7	0.7	0.43
SS 16	12	4	72	6	0.2	1.7	0.9	0.8	0.22
SS 17	16	5	288	8	0.2	4.9	1.2	1.1	0.17
SS 18	18	3	5184	11	0.2	86.1	1.6	1.6	0.13
SS 19	9	2	48	6	0.2	1.3	0.7	0.6	0.29
SS 20	27	3	2239488	18	0.4	49621.6	2.7	2.5	0.15
SS 21	18	2	768	9	0.3	17.7	1.4	1.2	0.21
SS 22	54	3	1671768834048	33	0.8	n/a	7.3	6.9	0.11
SS 23	24	3	768	9	0.4	20.0	1.7	1.5	0.24
SS 24	30	3	559872	15	0.5	13932.3	2.4	2.2	0.21
SS 25	18	1	36	4	18.7	29.1	10.5	3.9	1.78
SS 28	12	1	24	4	4.9	7.0	5.8	1.2	0.84
SS 29	6	1	72	5	1.0	3.0	1.6	0.6	0.63
SS 30	5	2	8	3	0.6	1.2	1.1	0.6	0.54
SS 31	6	2	10	3	0.9	1.7	1.5	0.7	0.60
SS 32	7	2	12	3	1.2	2.1	1.8	0.7	0.67
SS 33	9	2	16	4	1.8	3.8	2.6	1.0	0.69
SS 34	19	2	36	4	6.7	23.8	20.4	1.7	0.33
MK 1	25	1	200	5	250.0	181.3	83.6	6.4	2.99
AR 1	60	1	120	4	77.7	553.9	255.8	3.4	0.30

Table 15 The performance of symmetry perception algorithms on an acyclic graph series with increasing symmetry, plus regular, wheel and mesh graphs. n is the number of atoms; IR is the Isomorph Rejection algorithm; IR configuration A generates the full automorphism group and *includes* ring perception time; IR configuration B finds a small automorphism generator set and *includes* the ring perception time; IR configuration C finds a small automorphism generator set and *excludes* the ring perception time.

The Isomorph Rejection algorithm was tested in 3 different configurations with results shown in columns A, B and C. Column A lists the times when the rejection mode is disabled and the full automorphism group is calculated. In all cases, bar SS22, the measured order agrees exactly with the theoretical value calculated from the discrete symmetry elements. The running time for SS22 was estimated to exceed at least 1 year so the result could not be confirmed. Column B lists times measured under the same running conditions as the Morgan and Figueras algorithms. This includes ring perception which is computed by lazy (on-demand) evaluation. Column C lists times measured without the contribution from ring perception.

In general the processing times of all algorithms are very good. As expected the Morgan class of algorithm can be up to an order of magnitude faster when processing small acyclic graphs (SS5 – SS10). However the differences diminish significantly when the number of cycles and total numbers of atoms increases. On average the Morgan algorithm ran twice as fast as the isomorph rejection algorithm when measured over the full set of 58 graphs.

Ring perception times dominate symmetry perception in ring dense graphs (FG7/9/20/25/27, SS25/34, MK1, AR1). The conclusion is that symmetry perception in cyclic graphs contributes only a small proportion of the overall time when compared to other tasks such as ring and aromatic bond perception. When ring perception is discounted the range of times measured for the isomorph rejection algorithm on the sample set is 0.2 to 6.9 ms with a mean time of 1.4 ms *(*column B*)*.

Overall the performance of the Isomorph Rejection algorithm was better than the Figueras algorithm even when discounting FG20 where the Figueras method was 12 times slower. Figueras reported that his algorithm was sensitive to the ordering of the atoms in the input graphs. [86] Examples of this sensitivity is found the related graphs FG8 and FG24 which exhibit extreme variation in times and FG22 and FG23 which are two orderings of otherwise identical graphs. This sensitivity is not apparent in the Isomorph Rejection algorithm as the CSP solver automatically reorders the input graph using a range of optimising heuristics (*vide supra*). The size of the generated automorphism group occasionally differs between the two algorithms but is always smaller than the size of the graph. The differences can be explained by the order which each algorithm processes the atoms.

Validating Stereochemical Matching

A series of experiments were conducted to validate the use of a CSP solver to accurately solve a range of stereochemical problems using symmetry perception with stereochemical constraints. The first experiment involved the design of an algorithm to recognise asymmetric, pseudo-asymmetric and non-asymmetric stereogenic centres. The second experiment builds on this algorithm to perceive and label *meso* compounds, C_2 symmetric compounds and pseudo C_2 symmetric compounds. The final experiment was the development of an algorithm to recognise homotopic, enantiotopic and diastereotopic faces of a variety of selected sp^2 atoms and sp^3 atoms with two lone pairs.

Assigning Non-asymmetric, Pseudo-asymmetric and Asymmetric Centres

The identification of stereo asymmetry utilises a CSP solver configured such that all selected matches of the molecule graph to itself contain at least one stationary[a] inverted parity (mirror-coset) match between stereocentres. This condition identifies stereogenic centres which have

Inputs: *G is a graph representing a molecule which may have stereogenic centres*

Operation: *AUTOMORPHISM is a function using a CSP solver to solve stereochemical constraints set for identical, enantiomer, epimer and diastereomer automorphism matches. STEREOS(G) is a function returning the set of non-planar stereogenic centres in G. MAP is a bijective mapping of G onto itself and is a CSP solution augmented by TYPE which identifies how groups of stereogenic centres map to each other. S is a set of stereogenic centres; s, s', p are stereogenic centres; t is the match parity between two stereogenic centres ($\oplus$ for same image or $\ominus$ for mirror image).*

Outputs: *ASYM is the set of stereogenic centres with non-identical substituents. NON_ASYM is the set of stereogenic centres with at least one pair of identical substituents. PSEUDO_ASYM is the set of stereogenic centres with a pair of mirror image substituents.*

```
MATCH ← true
WHILE: MATCH is true
  (MATCH, TYPE, MAP) ← AUTOMORPHISM (G, EQUIV)     ( a CSP solver set for automorphism )
  IF: TYPE is ENANTIOMER_MATCH                       ( mirrored stereochemical automorphism found )
      c ← 0;
      FOR ALL: $s \in$ STEREOS (G)
          c ← c + 1                                  ( count total number of  inverted stereocentres )
          IF: s = MAP (s)
            $S \leftarrow S \cup \{ s \}$                ( collect stationary inverted stereocentres )
     IF: c = 1                                       ( only a single stationary inverted stereocentre )
         NON_ASYM ← NON_ASYM ∪ S                     ( has identical appendages )
     ELSE:
        PSEUDO_ASYM ← PSEUDO_ASYM ∪ S                ( has mirror image appendages )
  ELSE IF: TYPE is EPIMER_MATCH                      ( single inverted stereocentre automorphism found )
      FOR ALL: $s \in$ STEREOS (G)
          (p, t) ← MAP (s)
          IF: $s = p \wedge t = \ominus$                ( found inverted stationary stereogenic centre? )
            NON_ASYM ← NON_ASYM ∪ { s }              ( store non-asymmetric centre )
  ELSE IF: TYPE is DIASTEREOMER_MATCH                ( multiple inverted stereocentre solution found )
     c ← 0
     FOR ALL: $s \in$ STEREOS (G)
         (p, t) ← MAP (s)
         IF: $s = p \wedge t = \ominus$                 ( count inverted stationary stereogenic centres )
           c ← c + 1, s' ← s                         ( note which stereocentre was stationary )
    IF: c = 1                                        ( only one inverted stationary centre found? )
        PSEUDO_ASYM ← PSEUDO_ASYM ∪ { s' }           ( store pseudo-asymmetric centre )
PSEUDO_ASYM ← PSEUDO_ASYM – NON_ASYM                 ( correct the pseudo-asymmetric set )
ASYM ←STEREOS (G) – NON_ASYM – PSEUDO_ASYM           ( all unassigned centres are asymmetric )
```

Listing 10 An algorithm for recognising asymmetric, pseudo-asymmetric and non-asymmetric stereogenic centres.

[a] A mapping of an object to itself is a stationary match.

identical or mirror image substituents. Each generated solution is first categorised according to the numbers of retained and inverted parity matches at each stereocentre. Those containing only inverted parity matches are labelled an ‘enantiomer’ match; those with one inverted and one or more retained parity matches are labelled an ‘epimer’ match; those with two or more inverted and one or more retained parity matches are a labelled a ‘diastereomer’ match.

The algorithm required to make the stereocentre assignments is show in Listing 10. The main

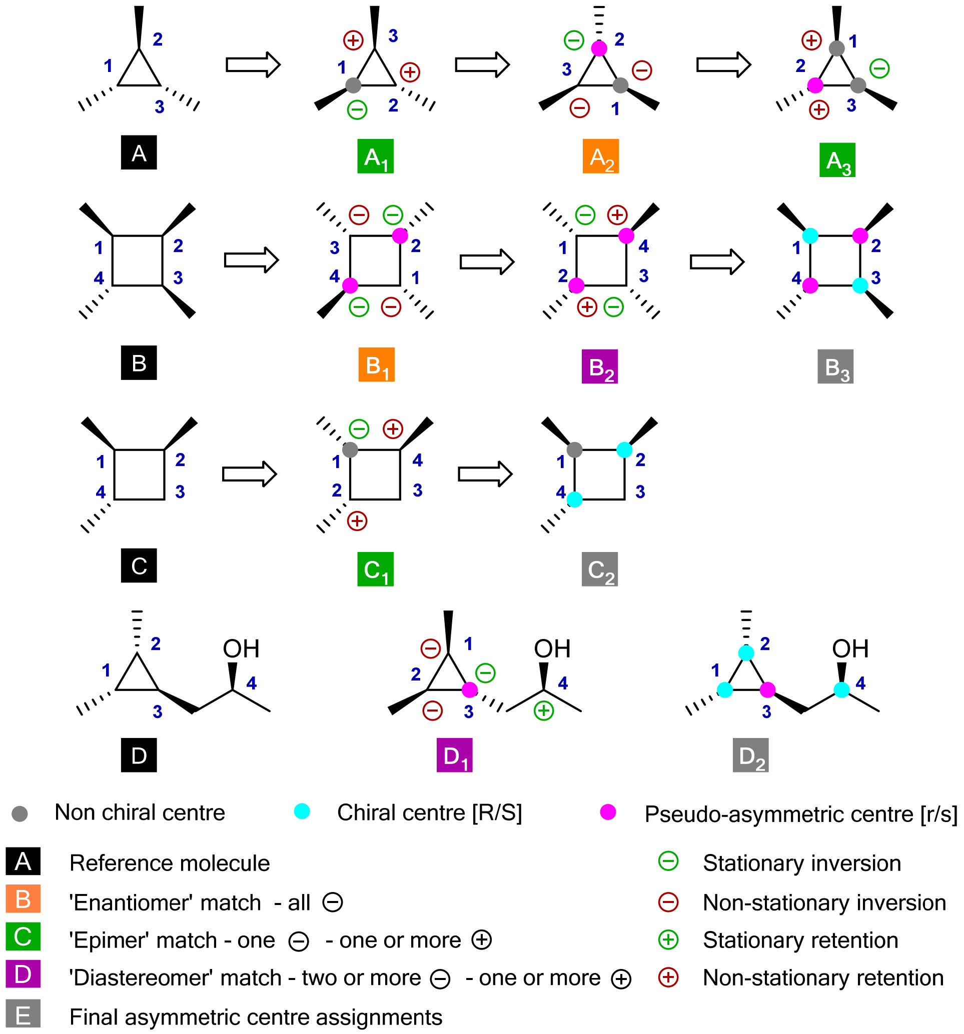

A
A$_1$
A$_2$
A$_3$
B
B$_1$
B$_2$
B$_3$
C
C$_1$
C$_2$
D
D$_1$
D$_2$
OH
1
2
3
4
● Non chiral centre
● Chiral centre [R/S]
● Pseudo-asymmetric centre [r/s]
A Reference molecule
B 'Enantiomer' match - all ⊖
C 'Epimer' match - one ⊖ - one or more ⊕
D 'Diastereomer' match - two or more ⊖ - one or more ⊕
E Final asymmetric centre assignments
⊖ Stationary inversion
⊖ Non-stationary inversion
⊕ Stationary retention
⊕ Non-stationary retention

Figure 78 The stepwise perception of asymmetric, pseudo-asymmetric and non-asymmetric stereogenic centres. CSP solutions containing stationary inverted stereogenic matches are used to systematically discover all non-asymmetric and pseudo-asymmetric centres. Unassigned centres at the end of the procedure are labelled asymmetric.

loop (lines 1 – 3) performs a series of automorphism matches on graph G, followed by an analysis that depends on the type of stereochemical match. ‘Enantiomer’ solutions (line 4) are scanned for a fixed stereogenic centre (line 8). If the molecule contains only one centre (line 10) then a pair of substituents must be identical and G has a non-asymmetric centre. If more than one stereogenic centre is present then the stationary centre has mirror image substituents and a pseudo-asymmetric centre has been located (*c.f.* A_2, B_1 in Figure 78). 'Epimer' solutions (line 14) contain only one inverted stereocentre (*c.f.* A_1, A_3, C_1 in Figure 78). If the inverted centre is stationary (line 17) then at least two substituents (which may contain stereocentres) are identical and the centre is non-asymmetric. 'Diastereomer' solutions (line 19) with one stationary inverted centre (line 25) identify centres with mirror image substituents and are thus pseudo-asymmetric centres (*c.f.* D_1 in Figure 78). If more than one stationary inverted centre is present then they are all asymmetric centres. This situation arises because the stationary centre shared by the substituents of another stationary centre is frustrated[b] and consequently the arms of the cyclic substituent 'pair' cannot be identical (*c.f.* B_1 in Figure 78).

A miss-assignment of pseudo-asymmetric centres occurs when a plane of symmetry exists between a pair of non-asymmetric centres. The same centres are also recognised as non-asymmetric via 'epimer' solutions that swap the identical substituents. This miss-assignment is corrected by giving priority to the non-asymmetric evaluation (line 27).

The symmetry pruning optimisations performed by the CSP solver prevent the unwanted exhaustive generation of all enantiomer, epimer and diastereomer automorphisms. This means that not all assignments can be deduced from the initial set of generated automorphisms. This deficiency is solved by introducing a post processing step (not shown in Listing 10) that propagates the non-asymmetry and pseudo-asymmetry information from these stereogenic centres to their rotational symmetry equivalent stereogenic centres.

The main procedure concludes by computing the set of asymmetric centres by deducting the proven non-asymmetric and pseudo-asymmetric from the set of non-planar stereogenic centres (line 28).

The algorithm was tested on a set of molecule representations and the results are presented in Figure 79 and Figure 81 (*vide infra*). All stereogenic centre classifications produced by an implementation of the algorithm are in agreement with expectations.

[b] In this case the same centre has to be simultaneously in two different configurations which is impossible.

Recognising *Meso*/Centro-symmetry, C_2 and Pseudo C_2 Symmetry in Molecules

Recognition of *meso* and C_2 molecular symmetry is supported by using a stereochemical semi-automorphism [c] procedure to locate rotational axes, mirror planes and swappable substituents.

Inputs: *G is a graph representing a molecule which may have stereogenic centres*

Operation: *AUTOMORPHISM is a function using a CSP to solve stereochemical constraints set for identical, enantiomer, epimer and diastereomer automorphism matches. ATOMS(G) is a function returning the set of atoms in G. MAP is a bijective mapping of G onto itself and is a CSP solution augmented by TYPE which identifies how groups of stereogenic centres map to each other. m, n are atoms. On first use all sets are empty.*

Outputs: *EQUIV(A) is the set of atoms symmetrical with atom A by rotation. IN_MIRROR(A) is the number of reflection planes passing through atom A. ACROSS_MIRROR(A) is the set of atoms that are mirror images of atom A. SWAP_EQUIV(A) is the set of atoms equivalent to atom A when swapping a pair of identical substituents.*

```
MATCH ← true
WHILE: MATCH is true
   (MATCH, TYPE, MAP) ← AUTOMORPHISM (G, EQUIV)        ( a CSP solver set for automorphism )
   IF: TYPE is IDENTICAL_MATCH                         (absolute stereochemical automorphism found )
      …
      ( see Listing 8 for finding rotationally equivalent atoms – stored in EQUIV )

   ELSE IF: TYPE is ENANTIOMER_MATCH                   ( mirrored stereochemical automorphism found )
       FOR ALL: m ∈ ATOMS (G)
          n ← MAP (m)
          IF: n = m                                                        ( is atom stationary? )
             IN_MIRROR (m) ← IN_MIRROR (m) + 1                ( count reflection planes cutting atom )
          ELSE:                                                                  ( reflected atom )
             ACROSS_MIRROR (m) ← ACROSS_MIRROR (m) ∪ { n }                 ( store reflected pairs )

   ELSE IF: TYPE is EPIMER_MATCH                       ( single inverted stereocentre automorphism found )
       FOR ALL: m ∈ ATOMS (G)
          n ← MAP (m)
          IF: n ≠ m                                                               ( swapped atom? )
             SWAP_EQUIV (m) ← SWAP_EQUIV (m) ∪ { n }                         ( store swapped pairs )
```

Listing 11 An algorithm for locating reflection planes and swappable substituents in a molecule containing stereogenic centres.

These features are detected and stored as sets of atom properties as shown in Listing 11.

Information about rotational axes in the molecule is detected by the stereochemically identical solutions emitted by the CSP solver. The orbits of stereogenic centres around an axis are stored in the bit-set matrix EQUIV. The procedure is identical to that described earlier in this chapter for

[c] True automorphism requires that all stereochemical parities match. Semi-automorphism allows the parities to miss-match in some constrained way to reveal underlying symmetry elements.

recording atom orbits (line 5). The addition of stereochemical constraints promotes what are otherwise permutation symmetry orbits to rotational symmetry orbits. [d]

Mirror planes are revealed by the 'enantiomer' solutions (line 6). Stationary atoms in the solution lie in a mirror plane (line 9) while non-stationary atoms are mapped to their reflections across the mirror plane. The algorithm records how many reflection planes in which each atom lies in the IN_MIRROR vector (line 10) and the reflected atom pairs are stored in the ACROSS_MIRROR bit-set vector (line 12). The method is unable to determine the degree of the

Inputs: *G is a graph representing a molecule. ACROSS_MIRROR, EQUIV, SWAP_EQUIV are sets of atoms with symmetry properties (see Listing 11). NON_ASYM, PSEUDO_ASYM and ASYM are sets of stereogenic centres (see Listing 10).*

Operation: *STEREOS(G) is a function returning the set of stereogenic centres in G. ROOT_ATOM(S) returns the atom at the stereogenic centre S. SN_STEREOS are stereogenic centres that have reflection symmetries; C2_STEREOS are those with C_2 symmetry; PC2_STEREOS are those in a pseudo C_2 environment. On first use all sets are empty.*

Outputs: *IS_MESO is true for meso and centro-symmetric molecules; IS_C2 is true for C_2 symmetric molecules; IS_PSEUDO_C2 is true for pseudo C_2 symmetric molecules; IS_CHIRAL is true for chiral molecules, false for achiral molecules.*

```
1.   FOR ALL: s ∈ STEREOS (G)
       ( look for symmetry relationships between asymmetric centres )
2.     a ← ROOT_ATOM (s)
3.     IF: s ∉ NON_ASYM
4.         IF: |ACROSS_MIRROR (a)| > 0                          ( in an S_n environment ? )
5.             SN_STEREOS ← SN_STEREOS ∪ { s }
6.         IF: |EQUIV (a)| ∈ {1, 3, 5, ...}        ( + self means it is in a 2n orbit ... implies a C_2 axis )
7.             C2_STEREOS ← C2_STEREOS ∪ { s }
8.         IF: |SWAP_EQUIV (a)| > 0                      (symmetrical about a non-chiral centre ? )
9.             PC2_STEREOS ←PC2_STEREOS ∪ { s }

       ( special case of non-asymmetric centres in both meso and pseudo-C_2 environments )
10.    ELSE IF: |ACROSS_MIRROR (a)| > 0 ∧ | SWAP_EQUIV (a)| > 0 ∧
                 |ACROSS_MIRROR (a) ∩ SWAP_EQUIV (a)| = 0  ( symmetry partners must be different )
11.        SN_STEREOS ← SN_STEREOS ∪ { s }
12.        PC2_STEREOS ←PC2_STEREOS ∪ { s }

     ( determine molecule symmetry classifications )
13.  IS_MESO       ← |SN_STEREOS| > 0
14.  IS_C2         ← |C2_STEREOS| > 0
15.  IS_PSEUDO_C2 ← |PC2_STEREOS| > 0
16.  IS_CHIRAL     ← |ASYM| > 0 ∧ |SN_STEREOS| = 0
```

Listing 12 An algorithm for recognising *meso*, centro, C_2 and pseudo C_2 symmetry in molecules.

[d] An exception occurs when orthogonal C_2 and C_n axes are present as the deduced orbit is the product of the individual orbits of the atom about each C axis.

S_n rotoreflection axis and consequently the distinction between *meso* and centro-symmetric molecules is not made.

The 'epimer' solutions (line 13) allow just one stereogenic unit at a time to be inverted to reveal

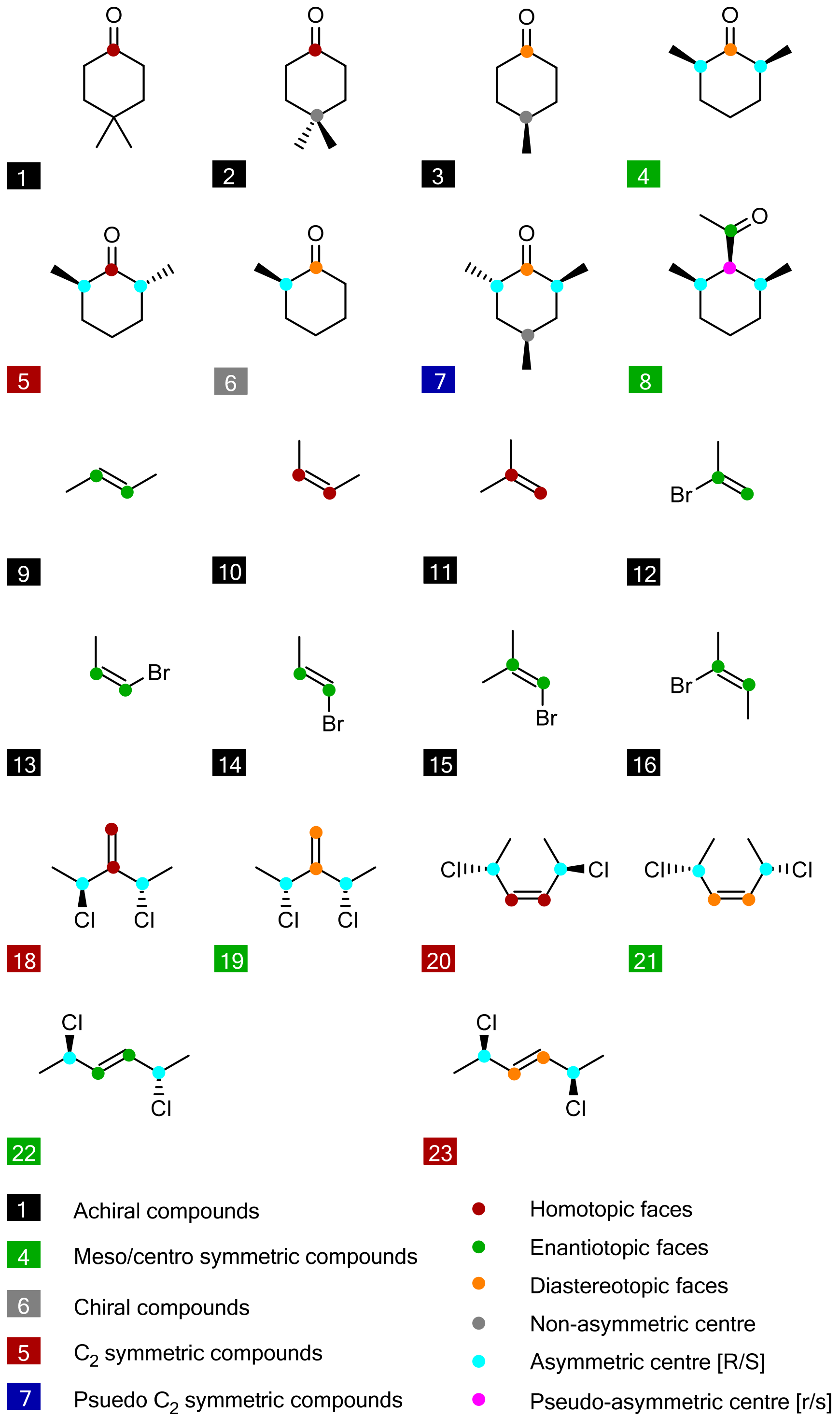


Figure 79 Classification results for stereogenic centre asymmetry and sp^2/sp^3 stereotopicity.

the locations of *stereochemically identical* substituents. The equivalence is recorded by scanning the solution and recording only the non-stationary atom pairings in the SWAPPED_EQUIV bit-set vector (lines 16-17). The maximum number of epimer matches will not exceed the number of stereogenic centres in the molecule as only one centre at a time will be inverted.

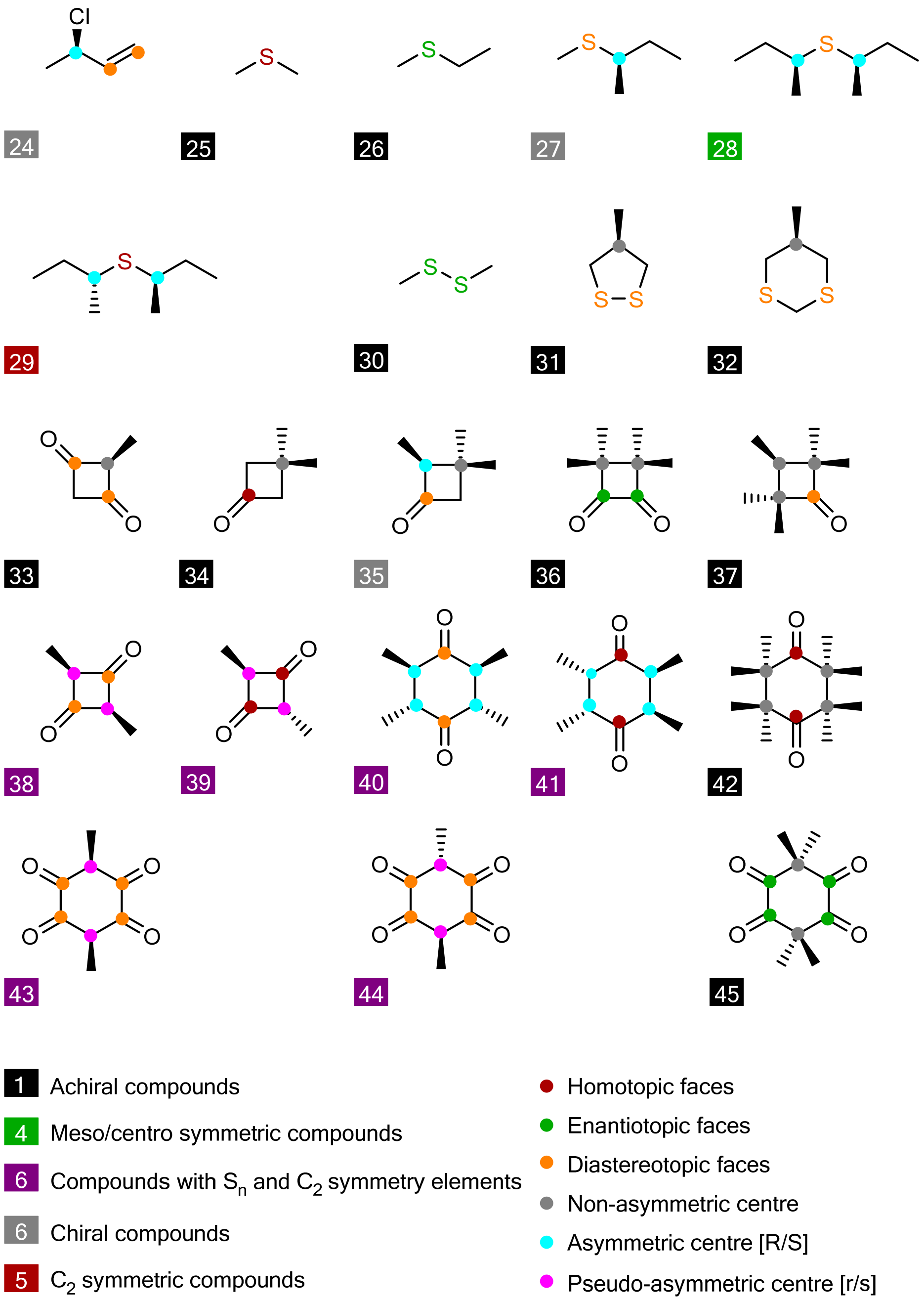

Cl
S
24
25
26
27
28
29
30
31
32
33
34
35
36
37
38
39
40
41
42
43
44
45
O
1 Achiral compounds
4 Meso/centro symmetric compounds
6 Compounds with S_n and C_2 symmetry elements
6 Chiral compounds
5 C_2 symmetric compounds
● Homotopic faces
● Enantiotopic faces
● Diastereotopic faces
● Non-asymmetric centre
● Asymmetric centre [R/S]
● Pseudo-asymmetric centre [r/s]

Figure 79 (Cont.) Classification results for stereogenic centre asymmetry and sp^2/sp^3 stereotopicity.

Listing 12 details an algorithm used to recognise achiral molecules including the subset of *meso*/centrosymmetric molecules. Chiral molecules and the subset of C_2 symmetric and pseudo

C_2 symmetric molecules are also recognised. Molecules with multiple symmetry elements can exhibit both *meso* and C_2 symmetry or *meso* and pseudo C_2 symmetry. Overall these latter molecules are achiral as they contain the *meso* plane of symmetry.

The chiral/achiral classification procedure (Listing 12: *vide supra*) scans all asymmetric and pseudo-asymmetric centres looking for specific symmetrical relationships with other stereogenic centres (lines 1-3). The stereogenic atom (line 2) is used as a proxy to check if the stereogenic centre has a mirror image (line 4), or is symmetrical with another stereogenic centre via an even n-fold rotation axis (line 6), or is swappable without loss of equivalence about a non-asymmetric stereogenic centre (line 8). These conditions establish if the stereogenic centre has a *meso*, C_2 or pseudo C_2 relationship to other stereogenic centres and the existence of such stereogenic centre relationships is sufficient to classify the molecule accordingly (lines 13-15) and thus determine if the molecule is chiral or achiral. The final step in the algorithm is the determination that a molecule is chiral if it has asymmetric centres but is not a *meso*/centrosymmetric compound (line 16)

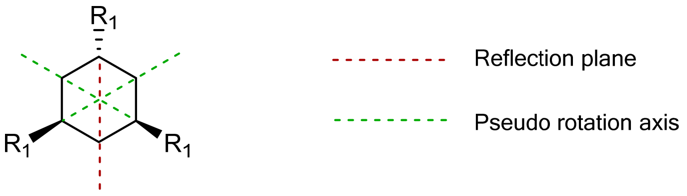


Figure 80 Symmetrically substituted (1r)-1,3,5-cyclohexanes exhibit both *meso* and pseudo C_2 symmetry.

A special case exists with tri-substituted (1r)-1,2,3-cyclopropanes and (1r)-1,3,5-cyclohexanes containing three identical substituents as these compounds have two opposed pseudo-C_2 axes and one reflection plane (Figure 80). The stereogenic centres lying on the pseudo C_2 axes are necessarily non-asymmetric and consequently are excluded from the above pairwise relationship analysis. A special test was added to find pairs of non-asymmetric centres that are in both environments and that the partner stereogenic centres for the two relationships are different (line 10). If the condition is met the stereogenic centre is assigned dual *meso*/pseudo C_2 relationships (lines 11-12).

.

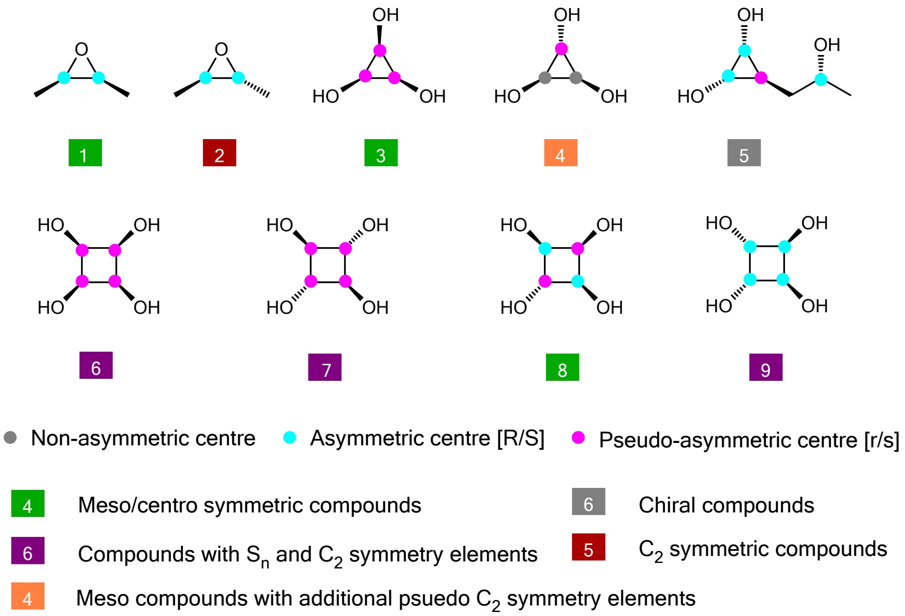

O
1
O
2
OH
HO
OH
3
OH
HO
OH
4
OH
OH
HO
5
HO
OH
HO
OH
6
HO
OH
HO
OH
7
HO
OH
HO
OH
8
HO
OH
HO
OH
9
● Non-asymmetric centre
● Asymmetric centre [R/S]
● Pseudo-asymmetric centre [r/s]
4 Meso/centro symmetric compounds
6 Chiral compounds
6 Compounds with S_n and C_2 symmetry elements
5 C_2 symmetric compounds
4 Meso compounds with additional psuedo C_2 symmetry elements

Figure 81 **Results for the perception of chiral/achiral classification of molecules containing multiple symmetry elements.**

Recognition of and Classification of Stereotopic Faces

Figure 82 presents the algorithm for assigning the stereotopicity of a trigonal face in a molecule. [232]

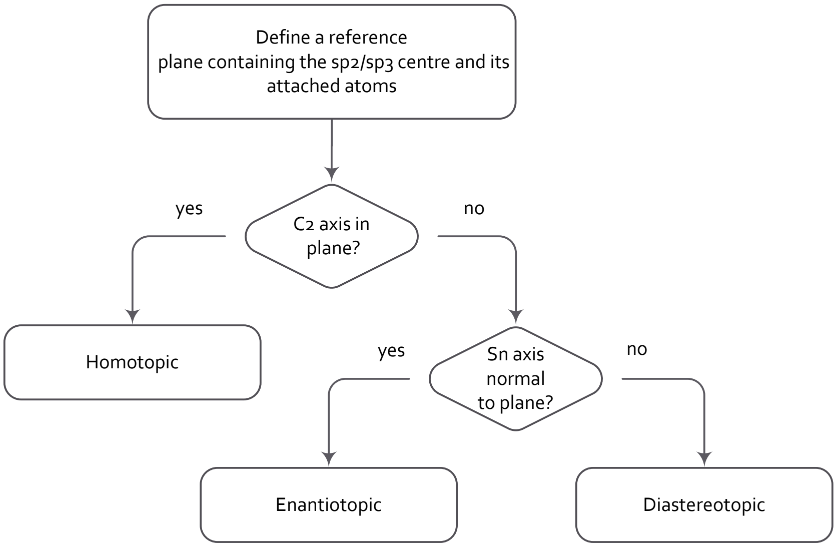


Figure 82 **An algorithm for classifying a stereotopic face.**

For the purposes of the experiment the atoms selected for analysis are all sp^2 hybridised atoms with two or more non-hydrogen substituents excepting those that are in aromatic rings. The lone pairs of di-substituted sp^3 hybridised sulphur atoms are stereotopic and are included in the

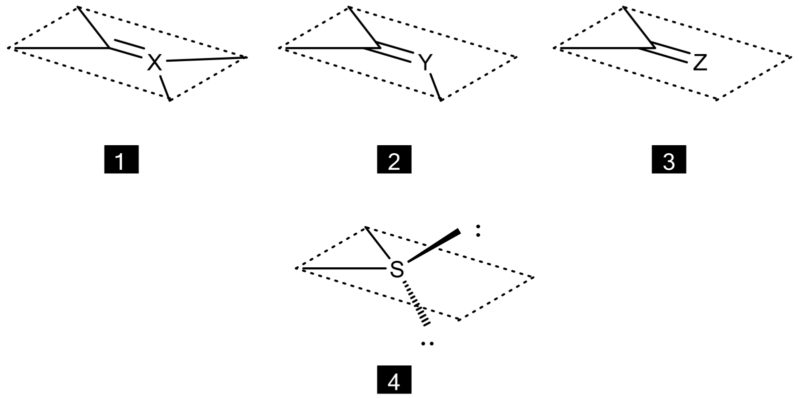
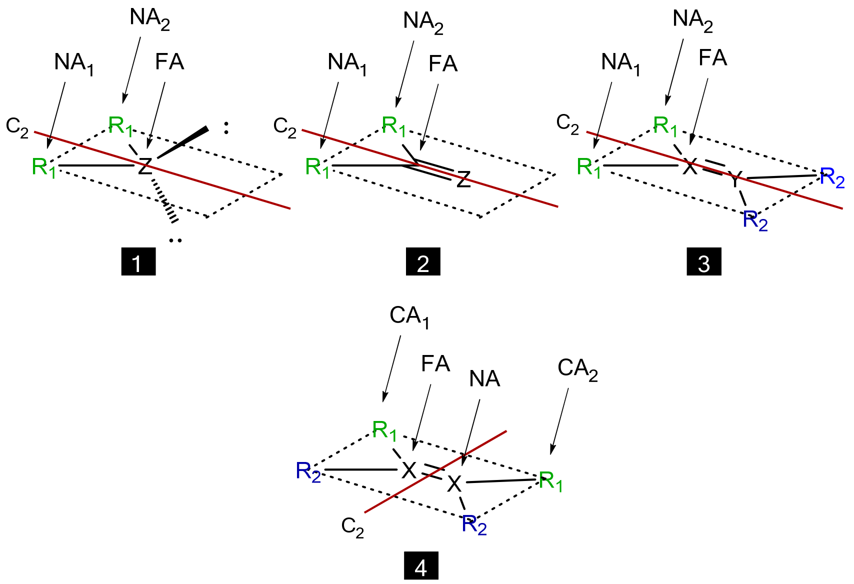


Figure 84 Reference planes for a selection of stereotopic faces. X = C, N^+; Y = N, S^+; Z = O, S.



$R_1 + R_1$, $R_2 + R_2$: substituents pairs symmetrical by rotation

FA - Face atom; NA, NA$_1$, NA$_2$ - neighbour atoms; CA$_1$, CA$_2$ - cis atoms

Figure 83 C2 rotation axes lying in the stereotopic reference plane.

selection (Figure 84). Each type of stereotopic face is defined by taking the immediate neighbour atoms (NA) of the selected sp^2/sp^3 face atom (FA) and using these to determine how C_2 and S_n axes are orientated to the face. The existence of a C_2 axis in the reference plane (Figure 84) is satisfied by establishing that any one of two conditions is true:

- First, two of the neighbour atoms NA_1 and NA_2 of face atom FA are symmetrically equivalent ($NA_1 \in$ EQUIV (NA_2)) (Figure 84: 1, 2 and 3).
- Second, if the stereotopic face atom FA is symmetrical to a neighbour atom NA (NA $\in$ EQUIV (FA)) then we have a symmetrical C=C bond between the NA and FA atoms. An in-plane C_2 axis passing through the C=C bond exists if any pair of *cis* atom neighbours CA_1 and CA_2 are also symmetrically equivalent ($CA_1 \in$ EQUIV (CA_2)) (Figure 84: 4).

The existence of a S_n axis normal to the reference plane (Figure 85) is satisfied by establishing that any one of three conditions is true:

- First, the face atom FA is in a mirror plane and *no* neighbour atoms lie across a mirror plane from any other neighbour atom (IN_MIRROR (FA) $> 0 \wedge$ ($\forall$ NA_1, NA_2; $NA_1 \notin$ ACROSS_MIRROR (NA_2))) (Figure 85: 1, 2). The second part of the test eliminates situations where an S_1 axis lies in the stereotopic reference plane (Figure 85: 4) but will allow stereotopic faces located on substituents of pseudo-asymmetry centres as these faces must lie on a mirror plane by definition (Figure 85: 1).
- Second, if the stereotopic face atom FA is symmetric by inversion of a neighbour atom NA (NA $\in$ ACROSS_MIRROR (FA)), then a S_2 axis passing through the C=C bond normal to the reference plane exists if any pair of *trans* atom neighbours TA_1 and TA_2 are symmetric by inversion ($TA_1 \in$ ACROSS_MIRROR (TA_2)) (Figure 85: 3).
- Third, if no asymmetric or pseudo-asymmetric centres are present in the molecule, it can be assumed that a conformation exists where a reflection plane must pass through the stereotopic reference plane. The latter test is necessary because information is only placed in the IN_MIRROR vector and ACROSS_MIRROR sets by symmetry perception when asymmetric and/or pseudo-asymmetric centres are present.

The stereotopic classification algorithm was implemented and tested on a set of molecule representations selected to cover the cases published by Kaloustian *et al* [232] and the results are included in Figure 79 (*vide supra*). All stereotopic classifications agree with expectations and the assignments published by Kaloustian *et al*.

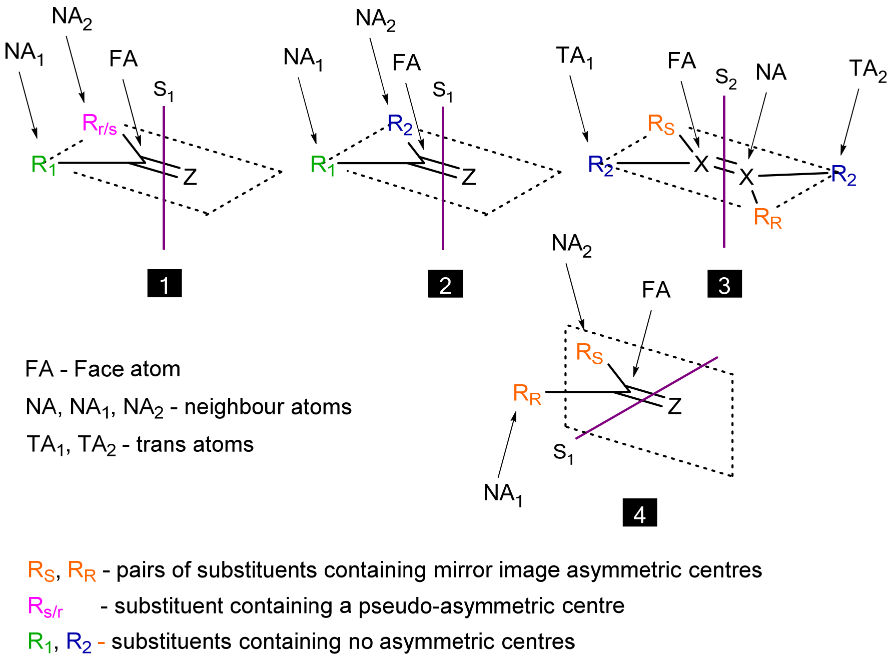

NA$_2$
NA$_1$
FA
S$_1$
R$_{r/s}$
R$_1$
Z
1
NA$_2$
NA$_1$
FA
S$_1$
R$_2$
R$_1$
Z
2
TA$_1$
FA
S$_2$
NA
TA$_2$
R$_S$
R$_2$
X
X
R$_2$
R$_R$
3
NA$_2$
FA
R$_S$
R$_R$
Z
S$_1$
NA$_1$
4
FA - Face atom
NA, NA$_1$, NA$_2$ - neighbour atoms
TA$_1$, TA$_2$ - trans atoms
R$_S$, R$_R$ - pairs of substituents containing mirror image asymmetric centres
R$_{s/r}$ - substituent containing a pseudo-asymmetric centre
R$_1$, R$_2$ - substituents containing no asymmetric centres

Figure 85 S_n rotoreflection axes that are orthogonal to the stereotopic reference plane (1 - 3). S_n axes lying in the stereotopic reference plane (4).

Conclusions

This chapter has described the implementation of a generally efficient constraints satisfaction problem solver capable of solving a wide variety of stereochemical graph isomorphism and subgraph isomorphism problems. Stereochemical constraints are solved within the CSP solver using the efficient ASAP constraints ordering principle.

The same CSP solver was used in a concise implementation of an isomorph rejection algorithm. The Figueras algorithm has been reviewed and tested and used as the impetus to devise a new algorithm that supports stereochemical isomorph rejection in a substructure match. The new algorithm has been used to exhaustively match any type of substructure query to a target molecule <u>*without*</u> producing duplicate solutions. The primary application of this algorithm will be to support the generation of duplicate free precursor molecules during the execution of a reaction rule during retrosynthetic analysis. This is described in Chapter 6.

The isomorph rejection algorithm is also used for topological symmetry perception for both molecules and reaction rules. Stereochemistry is naturally handled as the automorph matching function is implemented by calling the existing CSP solver that is able to solve stereochemical constraints. The correctness of the symmetry perception algorithm was evaluated by implementing and testing a series of algorithms that detect *meso* and C_2 symmetric molecules, asymmetric and non-asymmetric stereogenic centres and for determining the topicity of sp^2 and selected sp^3 centres.

The performance times across a range of simple and difficult graphs was excellent and offers a practical alternative to the approximate methods based on the less versatile Morgan or SEMA algorithms.

Chapter 4
Stereochemical Transform Rules

Introduction

This chapter introduces a new approach for defining transformation rules represented in the form of annotated reaction drawings created using a standard chemical drawing package. This allows the structural and stereocentre relationships to be represented in the natural structural language of chemistry. An annotation tool, available in MarvinSketch,[jj] is used to add functional constraints to atoms and bonds using an extended subset of the PATRAN language [55,56,42] to enable detailed retron descriptions to be defined.

The computer representation of stereospecific and stereoselective transform rules is of foremost importance to this research project. The priority for the transformation rule language is to accurately represent and process the stereochemical constraints and stereocentre relationships specified in the retron substructure. The presence of the retron in a target molecule is necessary to trigger the application of the transform to convert the target to a precursor molecule. An absence of appropriate structural, functional and stereochemical constraints in the retron would otherwise lead to poor results by presenting the user with unrealistic reaction sequences lacking appropriate activating, deactivating, directing or leaving groups; violating stereospecific mechanisms; or ignoring necessary stereoselective controls.

A new implementation of a functional group (FG) perception package is described that utilises the same graphical rule approach to define recognition patterns. These recognition patterns also incorporate PATRAN constraints which are evaluated by a CSP solver configured to match the substructure patterns to a target molecule. Performance optimisation *via* the use of a simple AND/OR screen system and a prebuilt pattern hierarchy is described.

The use of named property definitions attached to specific pattern atoms or bonds enables the recognition of electron-withdrawing or donating centres or potential nucleofuge and electrofuge leaving groups amongst other electronic features. The functional groups and atom/bond

[jj] MarvinSketch is supplied by ChemAxon -
http://www.chemaxon.com/products/marvin/marvinsketch/

properties can be referenced by name *via* PATRAN clauses to add necessary electronic constraints in a reaction rule.

The remodelled PATRAN language is described with particular emphasis on new features that have been introduced to support necessary constraints in stereoselective reaction rules.

A Stereochemical Retron Description Language

Background

Chapter 1 discussed the use of PATRAN in the LHASA program to define the keying retron substructure in a retrosynthesis transform. PATRAN exploited a set of structural constraint statements to precisely define the retron substructure environment. However it lacked any meaningful method to directly define stereogenic centres. However, manipulation of stereochemistry during the generation of the precursor molecule was supported in subsequent CHMTRN statements (e.g. RETAIN/INVERT AT ATOM*1).

Long *et al* developed an alternative 2D text-based retron language for the LHASA program using a text-based multi-line layout to represent a reaction scheme (*see Chapter 1*).[233] This system also introduced generic electronic constraints such as W atoms to represent electron-withdrawing groups. It dispensed with the CHMTRN structural manipulation statements by allowing the user to represent both educt and product fragments in the scheme from which the structural manipulation operations where automatically generated. However the direct representation of stereochemistry was not implemented even though the rule layout closely resembled a Fischer projection.

A new structural drawing-based retron language has been developed for this project. It retains the use of the constraint statements of the PATRAN language for its ability to describe detailed unary and n-ary constraints on atoms and bonds. It also adopts the approach of the LHASA 2D retron language to automate the generation of structural manipulation operations by specifying atom-atom mapped educt and product substructures but uniquely adds the automatic manipulation of stereochemistry to its capabilities.

The Graphical Reaction Rule Notation

Figure 86 illustrates the general principles of the graphical reaction rules. Atom and bond frameworks are drawn in the expected manner familiar to all chemists. Stereo “wedged” and “dashed” bond symbols are added to designate absolute or relative stereochemistry in the retron substructure and the generated precursor substructure.

Rule A (Figure 86) defines any substitution with inversion at a tetrahedral carbon atom[kk]. The rule requires that all corresponding educt and product atoms are atom-atom mapped so that a hyperstructure representation of the reaction (*vide supra*) can be constructed by the retrosynthesis module after the import of the MOLFILE data from the drawing application. Perception routines that analyse the hyperstructure deduce that: the educt C-R_4 bond is broken; the product C-R_5 bond is formed; and the stereogenic centre is inverted with substitution (*vide supra*). This information is used to construct the hyper-substructure and the constraints for a search query derived from the rule. The query is used to search for matching example reactions using a batch application built around a CSP solver configured to match a hyper-substructure query graph representing a reaction rule (containing the retron) to a hyperstructure graph representing a target reaction.

R atoms match any type of atom including non-explicit hydrogen atoms implied by the valence of the attached atom. Numbered R atoms of the form R_n add the constraint that atoms with identical n values must be symmetrically equivalent if and only if the R_n atoms are immediately attached to a *common* atom. However any pair of differently numbered R_n atoms is not required

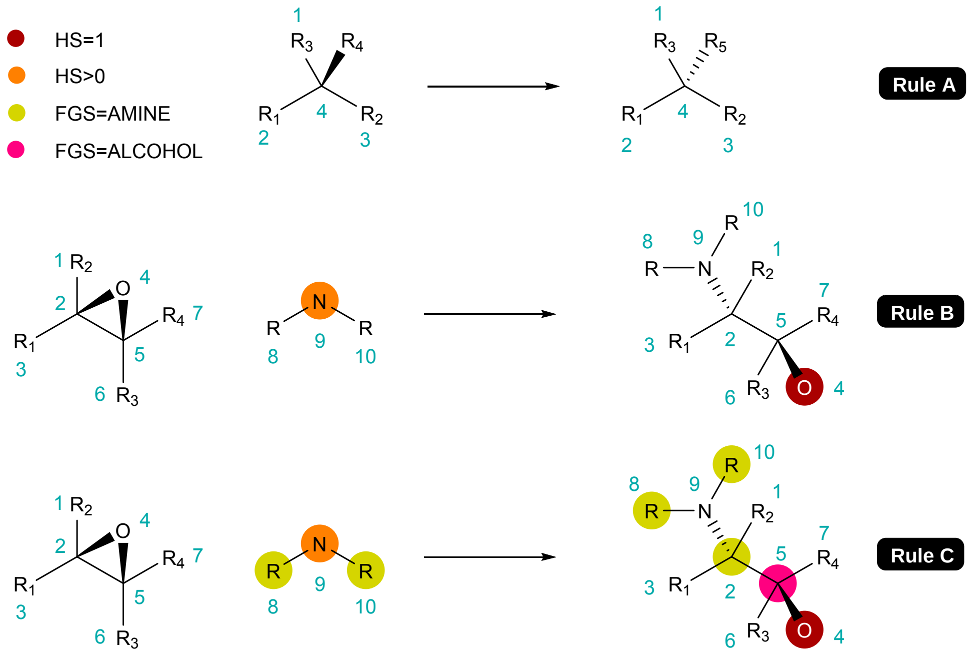


Figure 86 Graphical reaction rules illustrating the use atom-atom maps, stereochemical notations and PATRAN constraint annotations.

[kk] Note that this is an illustrative rule. It is not a useful retrosynthetic rule.

to be non-equivalent. This approach allows chiral and non-chiral centres to be specified in a natural manner in the drawing.

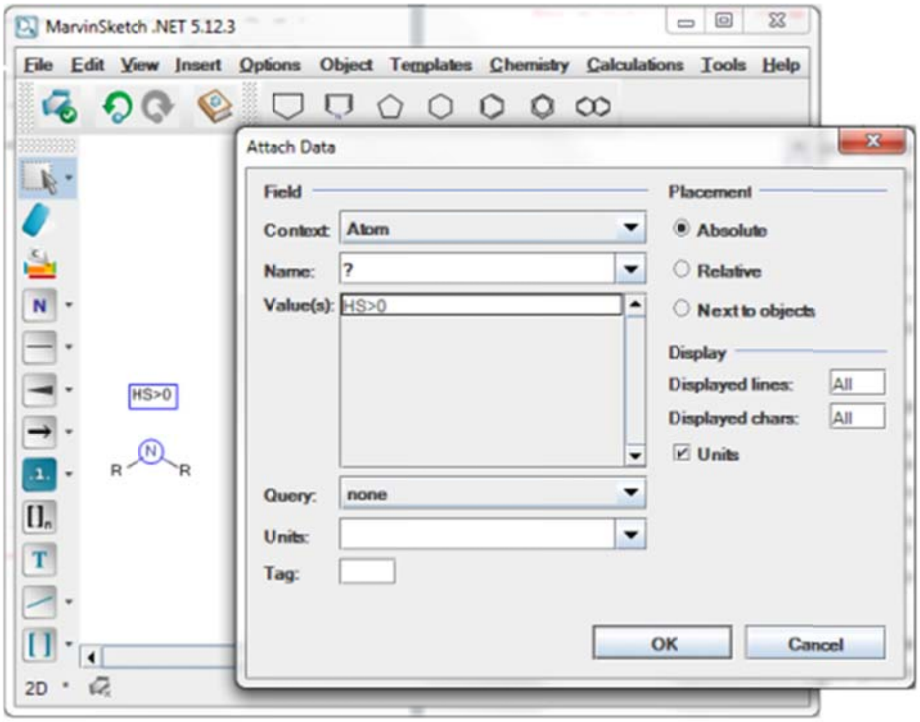
Atom Symbol	Interpretation
*	Any atom type including implied hydrogen
R	Any atom type including implied hydrogen (alias of *)
R_1 … R_n	Any atom type including implied hydrogen. The atom must be symmetrically equivalent to other identical R_n atoms attached to the same atom.
RS	A small atom. Currently only matches a hydrogen atom.
RM	A medium sized atom. Currently matches any non-aromatic carbon atom with 1 or 2 heavy atom attachments. (*e.g.* Me, Et *etc*.).
RL	A large atom. Currently matches: any atom in an aromatic cycle; any silicon atom; any tertiary or quaternary carbon atom. (*e.g.* Ph, SiR_3, iPr, tBu *etc*.).
A	Any non-hydrogen atom
Q	Any heteroatom (non-carbon and non-hydrogen)
X	Any one of F, Cl, Br, I
CX	Any one of Cl, Br, I
BX	Any one of Br, I
[N,O, S]	An inclusive list of specified elements
Not [N,O,S]	An exclusive list of specified elements
EWG	Electron-withdrawing atoms. This matches atoms perceived as electron-withdrawing centres as defined by functional group recognition patterns (*vide infra*).
EDG	Electron-donating atoms. This matches atoms perceived as electron-donating centres as defined by functional group recognition patterns (*vide infra*).
M	Any metal atom (electropositive atoms)
Z	Any non-metal atom with donor lone pairs (excludes boron and halogens)
NF	Nucleofuge leaving group. This matches atoms perceived as leaving centres taking away two electrons. These are defined by the functional group recognition patterns (*vide infra*).
EF	Electrofuge leaving group. This matches atoms perceived as leaving centres leaving electrons behind. These are defined by the functional group recognition patterns (*vide infra*).
AR	Any atom in an aromatic ring

Table 16 Atom types with interpretations that are supported in the graphical reaction rules.

Bond Symbol	Interpretation
— ▬ ┅	Single bond with or without a stereochemical up/down designations.
= ⨯	Double bond. Crossed double bonds disable stereochemical constraints (*e.g.* will match either *cis* or *trans* isomers)
≡	Triple bond
∷∷	Aromatic bond
╌═╌	Single or double bond
‒ - - - -	Single or aromatic bond
═ - - - -	Double or aromatic bond
········	Any bond

Table 17 Bond types with interpretations that are supported in the graphical reaction rules.

Table 16 and Table 17 list the full scope of the atom and bond types supported for drawing the rule atom and bond frameworks. Atoms types *, A, Q, X, M are standard query types in the


MarvinSketch .NET 5.12.3
File Edit View Insert Options Object Templates Chemistry Calculations Tools Help
HS>0
N
R
R
2D

Attach Data

Field

Context: Atom

Name: ?

Value(s): HS>0

Query: none

Units:

Tag:

Placement

● Absolute

○ Relative

○ Next to objects

Display

Displayed lines: All

Displayed chars: All

☑ Units

OK Cancel

Figure 87 Using the MarvinSketch chemical drawing editor “Attach Data” tool to attach the PATRAN expression “HS>0” to a selected atom.

MarvinSketch chemical drawing application and can be added directly via the Advanced Periodic Table tool. R_1, R_2 … R_n atoms are added via the R Group tool. The regular Periodic Table tool is used to create inclusive or exclusive element lists. The remaining atom types (R, RS, RM, RL, CX, BX, EWG, EDG, Z, NF, EF, AR) are non-standard types and can only be input via the Atom Alias tool. The perception of EWG, EDG, NF and EF in target molecules by the functional group perception module is described later in this chapter.

Rule B (Figure 86) represents a more practical rule that is designed to select example reactions in which suitable nitrogen nucleophiles bearing at least one hydrogen atom stereospecifically open an epoxide via an S_N2 mechanism. The stereochemistry of the carbon attacked by the nitrogen nucleophile is explicitly inverted in the rule while the stereochemistry of the resultant alcohol carbon is retained. This rule introduces the use of PATRAN constraints (*e.g.* hydrogen count expressions HS>0 and HS=1) to limit the extended atom and bond environments. Figure 87 illustrates the method used to attach a PATRAN expression to an atom (or a bond) using the MarvinSketch "attach data to object" tool. The field name used for PATRAN constraints is "?".[ll] The illustrated constraint limits the number of hydrogen atoms attached to the nitrogen atom.

Rule C adds a tighter set of functional group constraints that ensure that the product is precisely confined to examples that produce only 1,2-amino alcohols by adding the 'FGS=AMINE' and 'FGS=ALCOHOL' expressions to the relevant atoms. The scope and explanation of the supported PATRAN expressions is detailed in the next section.

The PATRAN Constraints Language

The original PATRAN language used in LHASA is a linear context-free grammar designed to describe a complete atom and bond substructure including chains, branches and rings. It supports a rich set of constraints designed to precisely describe the local substructure environment and a limited extended environment such as for describing partial embedment in rings or functional groups. It has been used in a number of applications ranging from the description of keying retrons in retrosynthesis rules,[55] substructure and superstructure screens to speed up chemical database searching,[56] functional group perception,[56] aromatic ring perception[59] and production rules in causal networks used for estimating synthetic accessibility.[42, 40]

[ll] Other field names used are: "*" to define functional group names; "@" to define atom or bond properties; "!" to set retrosynthetic goals (*vide infra*).

The previous section described how the atom/bond substructure framework is now handled by using standard chemical drawings. This removes the need for much of the PATRAN syntax concerned with atom and bond connectivity. Consequently the only parts of PATRAN that are retained are the atom and bond feature sets. The feature sets have been extended to support additional constraints expressions to provide finer control of the substructure environment for both reaction rules and functional group recognition patterns (*vide infra*).

Table 18 provides a complete formal syntax definition in the extended Backus-Naur production rule format (EBNF) for the version of PATRAN used in this project. The Backus-Naur symbols used are: “→“ is the production operator meaning “*is defined to be*”; “(*token1* \| *token2*)” is the ‘or’ operator representing a list of alternative productions; “< *tokens* >*” is the Kleene-star operator which indicates that the contained language tokens can be repeated zero or more times; and quoted words (in green) are language terminals and are the literal characters of the defined language. In Table 18 the root production rules are ATOMS_FEATURES and BOND_FEATURES.

ATOM_FEATURES	→	ATOM_FEATURE <OUTER_AND ATOM_FEATURE>*
BOND_FEATURES	→	BOND_FEATURE <OUTER_AND BOND_FEATURE>*
ATOM_FEATURE	→	(HYDROGENS | HALOGENS | HETEROATOMS | NON_AROMATIC_HETS | ELECTRON_WITHDRAWING_GROUPS | ELECTRON_DONATING_GROUPS | LONE_PAIRS | CHARGE | RING_SIZES | CONNECTIONS | SAME_RING | DIFF_RING | SAME_SYM | DIFF_SYM | ARYL | MULTIPLE_BOND |HYBRIDISATION | STEREO_CENTRE | STEREO_TOPICITY | INCLUDED_FUNCTIONAL_GROUPS | EXCLUDED_FUNCTIONAL_GROUPS | INCLUDED_PROPERTIES | EXCLUDED_PROPERTIES)
BOND_FEATURE	→	(RING_SIZES | SAME_RING | DIFF_RING | INCLUDED_FUNCTIONAL_GROUPS | EXCLUDED_FUNCTIONAL_GROUPS | INCLUDED_PROPERTIES | EXCLUDED_PROPERTIES)
HYDROGEN	→	“**HS**” (EQ_EXPRESSION | LT_EXPRESSION | GT_EXPRESSION)
HALOGENS	→	“**HALS**” (EQ_EXPRESSION | LT_EXPRESSION | GT_EXPRESSION)
HETEROATOMS	→	“**HETS**” (EQ_EXPRESSION | LT_EXPRESSION | GT_EXPRESSION)
NON_AROMATIC_HETS	→	“**NAHETS**” (EQ_EXPRESSION | LT_EXPRESSION | GT_EXPRESSION)
ELECTRON_WITHDRAWING_GROUPS	→	“**EWGS**” (EQ_EXPRESSION | LT_EXPRESSION | GT_EXPRESSION)
ELECTRON_DONATING_GROUPS	→	“**EDGS**” (EQ_EXPRESSION | LT_EXPRESSION | GT_EXPRESSION)
LONE_PAIRS	→	“**EPS**” (EQ_EXPRESSION | LT_EXPRESSION | GT_EXPRESSION)
CONNECTIONS	→	“**CONS**” (EQ_EXPRESSION | LT_EXPRESSION | GT_EXPRESSION)
CHARGE	→	“**CHARGE**” EQ (CHARGE_VALUE | TRISTATE_VALUE | LIST_VALUE)
RING_SIZES	→	“**RINGS**” EQ RINGS_VALUE
SAME_RING	→	“**SAMERING**” EQ GROUP_VALUE
DIFF_RING	→	“**DIFFRING**” EQ GROUP_VALUE
SAME_SYM	→	“**SAME**” EQ GROUP_VALUE
DIFF_SYM	→	“**DIFF**” EQ GROUP_VALUE

ARYL	→	“**ARYL**” EQ TRISTATE_VALUE
MULTIPLE_BOND	→	“**MBOND**” EQ TRISTATE_VALUE
HYBRIDISATION	→	“**SPS**” EQ POS_LIST_VALUE
STEREO_CENTRE	→	“**STEREO**” EQ TRISTATE_VALUE
STEREO_TOPICITY	→	“**TOPICITY**” EQ STEREO_TOPICITY_VALUE
INCLUDED_FUNCTIONAL_GROUPS	→	“**FGS**” EQ FG_ VALUE <OR FG_VALUE>*
EXCLUDED_FUNCTIONAL_GROUPS	→	“**FGNOT**” EQ FG_ VALUE <INNER_AND FG_ VALUE >*
INCLUDED_PROPERTIES	→	“**PROP**” EQ PROP_ VALUE <OR PROP_ VALUE >*
EXCLUDED_PROPERTIES	→	“**PROPNOT**” EQ PROP_ VALUE <INNER_AND FG_ VALUE >*
EQ_EXPRESSION	→	EQ POS_LIST_VALUE
LT_EXPRESSION	→	LT POS_VALUE
GT_EXPRESSION	→	GT POS_VALUE
GROUP_VALUE	→	POS_VALUE
CHARGE_VALUE	→	(“**ANION**” | “**CATION**” | “**NEUTRAL**”)
STEREO_TOPICITY_VALUE	→	(“**HOMO**” | “**ENANTIO**” | “**DIASTEREO**”)
TRISTATE_VALUE	→	(“**YES**” | “**NO**” | “**ANY**”)
RINGS_VALUE	→	RING_VALUE <OR RING_VALUE>*
RING_VALUE	→	POS_VALUE <INNER_AND POS_VALUE>*
FG_ VALUE	→	*List of tokens defined by functional group perception patterns e.g. “AMINE”, “ALCOHOL”*
PROP_ VALUE	→	*List of tokens defined by functional group perception patterns e.g. ”EWG”, “EDG”, “NF”, ”EF”*
LIST_VALUE	→	VALUE <OR VALUE>*
POS_LIST_VALUE	→	POS_VALUE <OR POS_VALUE>*
VALUE	→	POS_VALUE | NEG_VALUE
NEG_VALUE	→	“**–**“ POS_VALUE
POS_VALUE	→	DIGIT <DIGIT>*
DIGIT	→	(“**0**” | “**1**” | “**2**” | “**3**”| “**4**” | “**5**” | “**6**” | “**7**” | “**8**” | “**9**”)
EQ	→	“**=**”
LT	→	“**<**”
GT	→	“**>**”
OUTER_AND	→	“**;**”
INNER_AND	→	“**+**”
OR	→	“**,**”

Table 18 The formal EBNF definition of the PATRAN atom and bond constraint expression syntax.

The features supported break down into the following categories: counts of features (*e.g.* numbers of attached hydrogen, heteroatoms or halogens *etc.*); membership or non-membership of a category (*e.g.* ring sizes, functional groups, aromatic rings, stereotopicity); property values or value ranges (e.g. charges, hybridisation); binary or n-ary relationships (*e.g.* two or more atoms or bonds are in the same ring or are equivalent by symmetry).

The illustrative PATRAN constraints statement “HS=0,1;HETS=1,2” means the associated atom has either zero *or* one attached hydrogen atom *and* one *or* two attached hetero atoms. The “;” character is the associative ‘and’ operator binding the adjacent expressions which must both evaluate as true. The “,” character in value lists is the associative ‘or’ operator binding adjacent values, one of which in the chain must evaluate as true. The counted properties, such as HS, always include features explicitly drawn in the substructure as well as those found occupying the free valence sites of the substructure. The primary use of counted properties is to limit the choices of what can match the unoccupied free bonding sites of a substructure atom.

The new EWGS and EDGS counters (attached electron-withdrawing/donating groups) were introduced to allow non-electrophilic and/or non-nucleophilic C=C bonds to be easily defined in reaction rules by setting these counts to zero. The NAHETS (non-aromatic heteroatoms) feature was added to solve a recurring problem when using the HETS counter in the existing ARChem functional group patterns. This problem resulted in the failure to recognise certain functional groups attached to aromatic heterocycles when the group was bonded alpha to an aromatic heteroatom. The NAHETS expression avoids counting the aromatic heteroatom when constraining the number of heteroatoms bonded to functional group locant atoms[mm].

The atom/bond features SAMERING, DIFFRING, SAME and DIFF are n-ary predicates meaning that any number of atoms or bonds can be declared as (not) belonging to the same ring or they are symmetrically (non-)equivalent. The value of these features is a unique group number used to associate the atoms or bonds. Thus all atoms labelled with the expression SAMERING=1 must belong to the same ring for the substructure pattern to match.

The atom TOPICITY clause can be applied to any appropriate trigonal atom or certain tetrahedral heteroatoms bearing two sets of lone pairs. It is used in rules to select only enantioselective desymmetrisation reaction examples or to eliminate examples with C_2 symmetry.

The special atoms types EWG, EDG, NF and EF are automatically translated during query preparation into A (non-hydrogen) atoms with the associated PATRAN constraints expressions ‘PROP=EWG’, ‘PROP=EDG’, ‘PROP=NF’ and ‘PROP=EF’. Likewise multiple Rn atoms connected to a common atom are translated into plain R atoms with the associated expression ‘SAME=n’ to constrain a set of grouped R_n atoms to be symmetrically equivalent. The use of R atoms allows a pair of implicit and/or explicit hydrogen atoms to satisfy the constraint.

[mm] Locants are atoms or bonds bearing the functional group or property name.

The PATRAN statements are processed by the candidate selector and constraints validator modules of the substructure CSP solver. All unary expressions (HS, HALS, HETS, NAHETS, EPS, CONS, CHARGE, RINGS, ARYL, MBOND, SPS and STEREO) are processed in the CSP candidate selector to allow any resultant empty candidate atom/bond sets to immediately curtail the search as early as possible to record a non-match. The n-ary expressions (SAMERING, DIFFRING, SAME, DIFF) are handled in the constraints validator and invoked during the backtracking search loop (*vide supra*) as these expressions need to be re-evaluated each time a query to target atom binding changes.

The unary functional group, property and stereotopicity expressions (EWGS, EDGS, FGS, FGNOT, PROP, PROPNOT and TOPICITY) are configured to be optionally processed in either the CSP candidate selector or deferred to the constraints validator. The former forces the eager[nn] perception of functional groups, while the latter ensures lazy evaluation and avoids FG perception entirely if FG constraints are absent or are never evaluated. Experimentation determined that lazy evaluation was often superior due to the likelihood that most search paths are frequently curtailed by the failure to match a ‘cheaper’ property. Functional group perception is a particularly expensive operation and consequently was a focus for some optimisation work.

Functional Group Perception

This section describes the application of the isomorph rejection algorithm described in the previous chapter to improve and enhance the existing ARChem functional group perception module. Symmetry information is used to reduce the number of discrete matches needed to recognise all defined functional groups in a molecule. In addition a simple and robust screening system is described which significantly speeds up the perception process.

Functional groups have been used in retrosynthesis planning programs to select reaction transforms[100, 13] and plan functional group protection sequences [234] in the LHASA and SECS [98] systems and for cataloguing tolerated functional groups occurring in example reactions in the ARChem system. [40]

[nn] “Eager” evaluation computes results in advance with the expectation the results will be eventually be useful. “Lazy” evaluation is only invoked at the point the result is actually needed.

Functional Group Recognition

A limited number of approaches for the computer recognition of functional groups have been reported in the literature. An implementation reported for the LHASA system [90] used a precompiled binary search tree derived from hand created data tables. A restricted set of atom types were selected as starting points in the molecule and a systematic search was carried out over attached atoms and bonds until a functional group was recognised or the search tree was exhausted. The search progressed by answering a series of yes/no questions about atom and bond properties. This approach proved to be difficult to optimise and maintain particularly when the need arose to add new functional groups. [235] The CAMEO reaction prediction system described a similar approach to this method. [18]

```
...STARTP
...N(-C[FGNOT=AZIRIDINE])(-C[FGNOT=AZIRIDINE])-C[FGNOT=AZIRIDINE]
...{1,2,4,3}{1,3,2,4}{1,3,4,2}{1,4,2,3}{1,4,3,2}
...STARTP
...N(-C)-C[FGS=AZIRIDINE]-C[FGS=AZIRIDINE]-@1
...{1,2,4,3}
...ENDP
```

Figure 88 An excerpt from the LHASA pattern set for recognising tertiary amines illustrating the use of automorphism groups to eliminate duplicate matches. The aziridine recognition patterns must precede these tertiary amine patterns.

Bersohn [236] described a slightly different approach that searched for basic functional groups and then noted important relationships between these groups to create functional group combinations. For example carbon-carbon double bonds may be paired with other groups to produce vinylic and allylic variations of the functional group.

The difficultly in maintaining decision trees prompted an alternative approach in LHASA based on atom-by-atom pattern matching. [235] This pattern based approach is also used in the CAESA [237] and ARChem systems. [40] Each recognition pattern is written in the PATRAN language[238] which allows functional group constraints to be added to any pattern *via* the FGS and FGNOT expressions. Patterns that define a functional group must therefore precede any pattern that refers to that functional group. This ordering was maintained by hand in the pattern source file. New functional group patterns could be easily added and existing ones readily edited. The pattern writer was also expected to add the automorphism group in list form as a pattern adjunct to ensure that redundant matches due to symmetry would be recognised and the duplicates removed in a post-match filtering step (Figure 88).

Enhanced Functional Group Perception

A number of useful representational and performance improvements have been added to an implementation of the ARChem functional group perception module. These are: the necessary ordering of the patterns has been automated; a detailed AND/OR screening algorithm has been created to speed up functional group perception by maximising the elimination of non-matching patterns before a detailed atom-by-atom search is conducted; the patterns are drawn as annotated chemical fragment diagrams which are easier to visually interpret and validate; the automatic recognition of symmetry is used to eliminate match redundancies without the need to manually compose a set of automorphism group directives.

The 204 functional groups currently used by ARChem were transcribed from the original linear PATRAN text into drawn molecular fragment diagrams annotated with PATRAN constraints, functional group definitions and chemical property definitions. The transcription process involved adjusting symmetrical functional groups so all symmetrically equivalent atoms and bonds had identical constraints and definitions assigned to them. A number of structural and constraints errors present in the original linear patterns where made apparent when drawn in diagram form and these were corrected and validated. HETS expressions where replaced with NAHETS expressions where appropriate to solve the ‘functional group attached α to aromatic heteroatom’ issue (*vide supra*). Fifty four additional functional groups patterns were added by the author to expand the scope of the system to recognise a wider range of functionality especially with respect to the recognition of specific leaving groups and electron-withdrawing/donating groups.

Figure 89 illustrates the new formulation of the functional group patterns with a selection of examples. The patterns were drawn using the MarvinSketch program and the PATRAN based

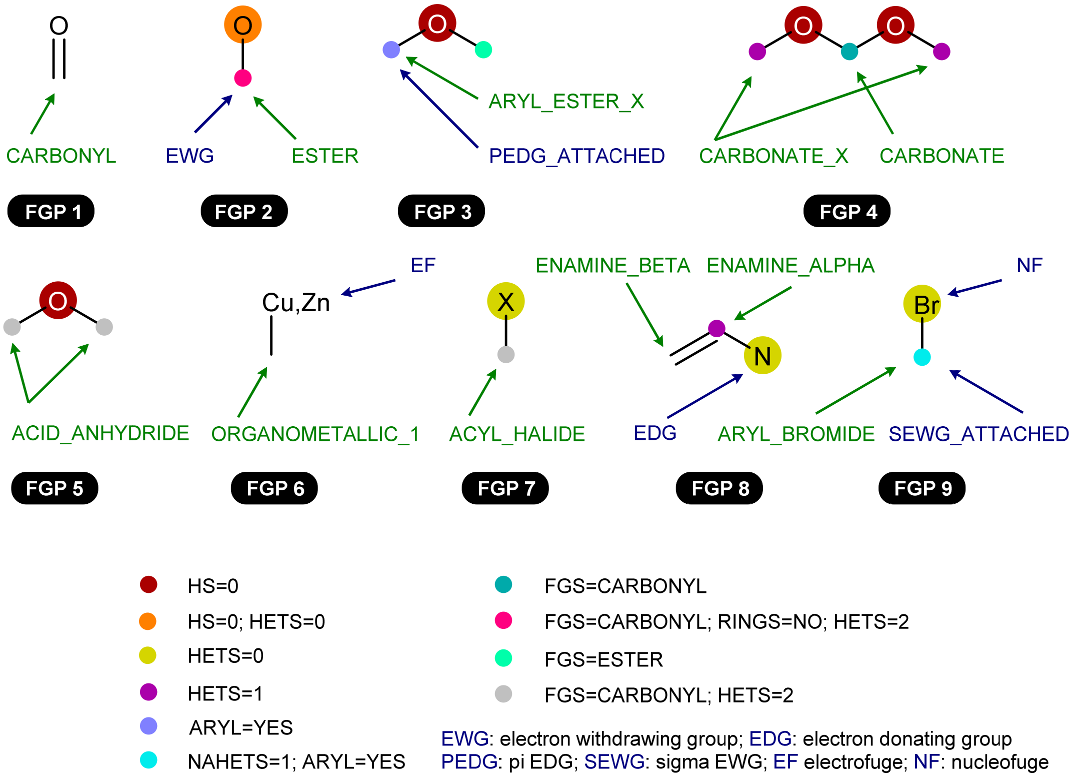


Figure 89 Sample functional group (FG) recognition patterns showing functional group definitions (in green) and property definitions (in blue). The FG names, properties and PATRAN constraints are created using the atom/bond annotation tool in the MarvinSketch drawing application.

constraints expressions were added using the "*Attach data to objects*" tool (*vide supra*) after the required atom or bond had been selected. The same tool was also used to add functional group name definitions and atom and bond property names. Constraints are represented by PATRAN feature statements (*e.g.* HETS=2; HS=0) while functional group definitions and properties are represented by plain text tokens (*e.g.* ESTER, EWG *etc.*) attached to an atom or a bond.

The goal of functional group perception is to identify and name all carbon atoms directly attached to heteroatoms or groups of heteroatoms. An important set of non-heteroatom bond environments are also named (*e.g.* unfunctionalised alkenes, alkynes and cyclopropanes). The exception to the hetero atom rule is that carbon atoms attached to heteroatoms in aromatic rings are not named.

The naming convention for ‘multivalent’ functional groups (such as esters or amides) is that carbons attached as *substituents* to the principal functional group by oxygen or sulphur have a ‘_X’ suffix added to the functional group name, while substituents attached to nitrogen atoms have an added ‘_Z’ suffix (Figure 90).

A carbon atom may receive a set of names ranging from a generic functional group class (*e.g.* AMIDE or AMIDE_Z) to a specific qualified environment name (*e.g.* TERTIARY_AMIDE or TERTIARY_ALKYL_AMIDE_Z). The assigned functional group names are targets for the FGS and FGNOT constraint expressions which are used in both functional group recognition and retrosynthetic transform rule patterns.

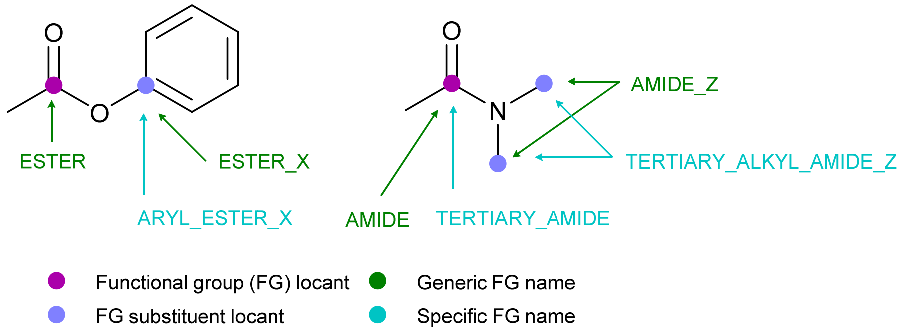
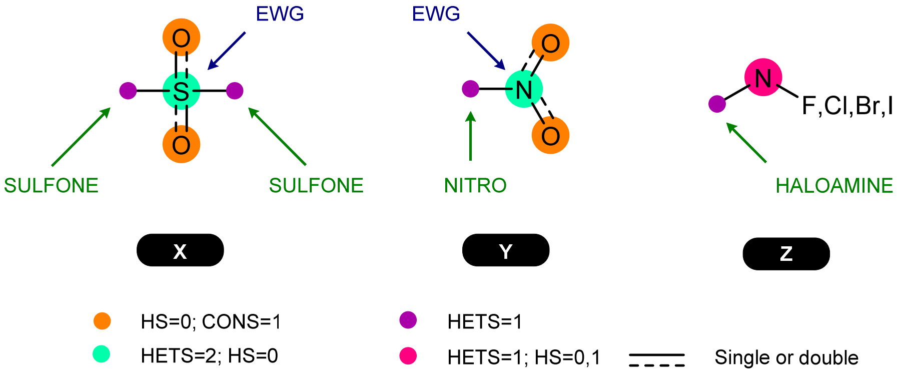


Figure 90 **Example named functional group locants. Atoms or bonds may have multiple assigned functional group names ranging from a generic name to a name indicating a specific environment. Functional group substituents have ‘_X’ or ‘_Z’ suffixes indicating the type of attached heteroatom.**

A range of functional and electronic properties are perceived during functional group perception by assigning named properties to specific atoms in the functional group patterns (*c.f.* FGP 2, FGP 3, FGP 6, FGP 8 and FGP 9 in Figure 89). These properties include: the attachment locants on aromatic rings of σ and π electron-donating and withdrawing groups (SEWG_ATTACHED, PEDG_ATTACHED, PEWG_ATTACHED[oo]); the atom locants of electron-donating and withdrawing centres in functional groups (EDG, EWG); the locants of nucleofuge and electrofuge leaving groups (NF, EF). These properties are referenced by name in transformation rules using the PROP and PROPNOT statements in a PATRAN constraints expression.

[oo] SEWG is *sigma electron-withdrawing group*; PEDG is *pi electron-donating group*; PEWG *is pi electron-withdrawing group*.

The patterns FGP 1, FGP 2 and FGP 3 (Figure 89) illustrate a series of interdependent patterns that must be applied in strict order when executing the functional group perception module. FGP 1 is a simple pattern that defines the CARBONYL moiety. FGP 2 defines the ESTER moiety *via* a reference to the CARBONYL moiety in a constraint expression and thus can only be applied after FGP 1. FGP 3 in turn depends on the ESTER functional group defined by FGP 2 and must be applied after FGP 2.



Screen Tables			
AND table		OR Table	
Fragment Key	Pattern Set	Fragment Key	Pattern Set
C	{ X, Y, Z }	S-	{ X }
C-	{ X, Y, Z }	S=	{ X }
S	{ X }	N-	{ Y }
S-	{ X }	N=	{ Y }
N	{ Y, Z }	O-	{ X, Y }
N-	{ Y, Z }	O=	{ X, Y }
O	{ X, Y }	F	{ Z }
		F-	{ Z }
		Cl	{ Z }
		Cl-	{ Z }
		Br	{ Z }
		Br-	{ Z }
		I	{ Z }
		I-	{ Z }

Table 19 A screen table constructed for functional group recognition patterns X, Y and Z.

FGP 6 and FGP 7 illustrate the use of atom symbol lists and special atom symbols such as X (for halogen) that are provided as standard query features in the Marvin chemical drawing program. These features enable generalised function groups such as HALOAMINE (Table 19: pattern Z) and ACYL_HALIDE (Figure 89: pattern FGP 7) to be efficiently defined.

Functional groups that may be drawn with alternative bonding representations can be accommodated. For example sulphoxide or sulphone groups can be drawn using multiple bond

(S=O) or ylid/charge separated bond styles ($S^+–O^-$). Using a “double or single” bond type with appropriate PATRAN constraints applied to limit hydrogen counts and the number of bond connections overcomes the need to use multiple patterns to account for each representation (*c.f.* Table 19: patterns X and Y).

An off-line program (*build-fg-database*) is used to read a set of functional group definition files encoded in standard MOLFILE format.[pp] Each pattern is scanned and a note made of the functional group names definitions and the names of other functional groups referenced in FGS and FGNOT constraint expressions. This information enables a hierarchical tree of patterns to be constructed. Using the hierarchy the program sorts the patterns into a required processing order, sequentially numbers them to maintain that order and stores them into a database for later retrieval by the main retrosynthesis program. The *build-fg-database* program also records the details of the hierarchical relationships as parent-child pairs in a separate table in the database. The hierarchy is later loaded by the functional group perception module and used in a screening step during target molecule perception.

At main program start up all the functional group recognition patterns contained in the prepared database file are loaded into memory and various low computational cost perception tasks are performed. The generated pattern information is cached for the duration of the program execution. The exception is the longer running tasks such as ring, aromatic, stereochemical and symmetry perception which are performed on demand the first time each pattern is used (*lazy evaluation*). The deferred perception results are also permanently cached when they become available.

The fragment screening data structure consists of two maps labelled the ‘AND’ and ‘OR’ screen tables (see Table 19). The ‘AND’ table is used to identify screen fragments that *must* be present and the ‘OR’ table for those that *may* be present to establish if a pattern will automatically fail to match a target molecule due to the absence of necessary structural features in the target molecule. This technique is used to eliminate as many patterns as possible from the pattern matching phase of functional group perception as a backtracking search via a CSP solver is a potentially expensive operation.

The ‘AND’/’OR’ tables contain simple screen keys representing single atom fragments and atom/attached bond fragments. The atom keys are represented by an atomic number. The

[pp] The specification documents are accessed at http://accelrys.com/products/informatics/cheminformatics/ctfile-formats/no-fee.php

atom/attached bond keys are coded as a pair of concatenated values consisting of an atomic number and a bond order value. Each of these screen keys indexes a list of functional group patterns that contain the screen fragment.

The keys are constructed in the following manner. For all patterns, each atom is visited in turn. If the atom symbol is not a special atom type (*i.e.* it is found in the periodic table of elements) an

Inputs: **G** is a molecule graph; **AF** is the set of 'AND' fragments found in all functional group patterns; **OF** is the set of 'OR' fragments found in all functional group patterns; **H** is the functional group pattern hierarchy.

Operation: *MoleculeToFragments*(G) is a function that generates a set of fragments (MF) found in molecule G; *AndFragmentToPatterns*(F) is a function that returns the set of patterns containing the 'AND' fragment F; *OrFragmentToPatterns*(F) is a function that returns the set of patterns containing 'OR' fragment F; *DependentPatternsOf*(P, H) is a function that recursively finds all the dependent patterns of P using the pattern hierarchy H

PAF is the set of present 'AND' fragments; AAF is the set of absent 'AND' fragments;
POF is the set of present 'AND' fragments; AOF is the set of absent 'OR' fragments;
PAP is the set of present 'AND' patterns; AAP is the set of absent 'AND' patterns;
POP is the set of present 'OR' patterns; AOP is the set of absent 'OR' patterns;
EP is the set of eliminated patterns.
All sets are empty on first use.

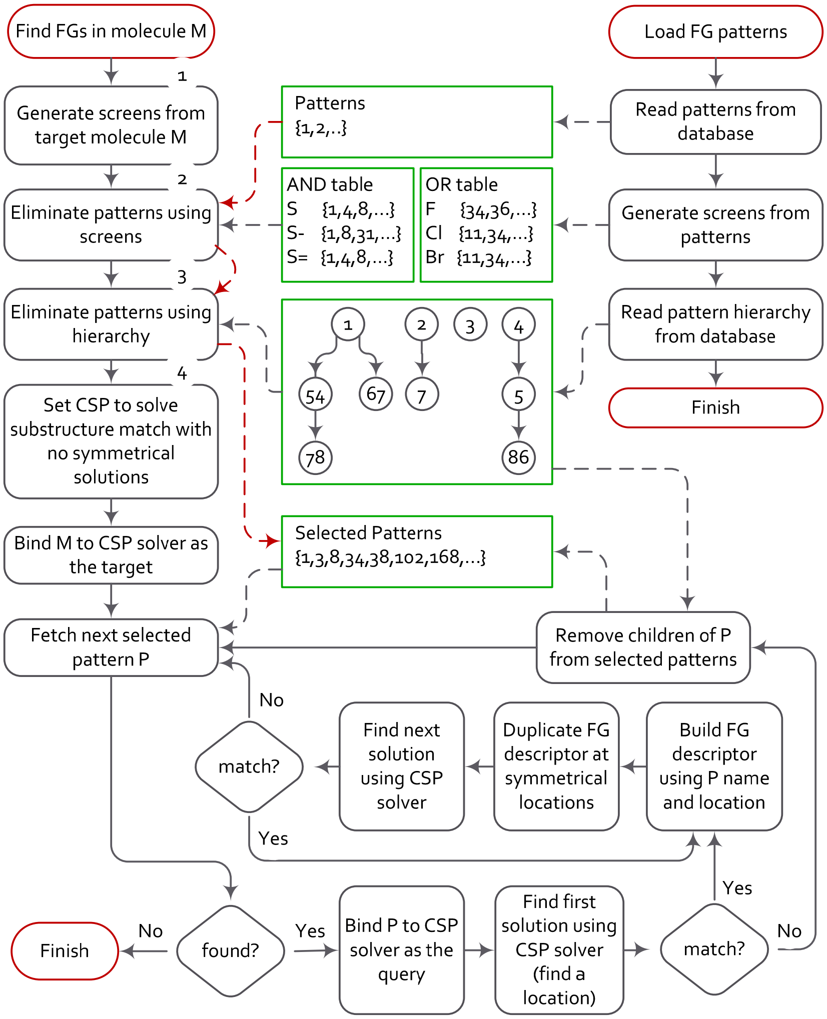
Outputs: SP is the set of screened patterns that will be matched to the molecule graph G

```
( find molecule fragments and classify them against the pattern screen fragments )
MF ← MoleculeToFragments (G)
PAF ← AF ∩ MF; AAF ← AF – PAF
POF ← OF ∩ MF; AOF ← OF – POF
( convert fragments to patterns using the 'AND'/'OR' screen tables )
FOR ALL: F ∈ PAF
    PAP ← PAP ∪ AndFragmentToPatterns (F)
FOR ALL: F ∈ AAF
    AAP ← AAP ∪ AndFragmentToPatterns (F)
FOR ALL: F ∈ POF
    POP ← POP ∪ OrFragmentToPatterns (F)
FOR ALL: F∈ AOF
    AOP ← AOP ∪ OrFragmentToPatterns (F)
( compute the initial set of selected patterns from the present and absent pattern sets )
SP ← PAP – AAP – (AOP – POP)
( find the eliminated patterns and remove their dependents from the selected patterns )
EP ← (PAP ∪ AAP ∪ POP ∪ AOP)  –  SP
FOR ALL: P ∈ EP
    SP ← SP  –  DependentPatternsOf  (P, H)
```

Listing 13 Eliminating non matching patterns using the fragment screens and the pattern hierarchy (detailing steps 1 – 3 in Flow chart 3).

atom key is created and added to the ‘AND’ screen table. The index of the pattern containing the atom is added to the pattern list associated with the atom key. If on the other hand the atom symbol is in a list of atom types (*e.g.* O, S) or is one of a special set of symbols representing a predefined list of atom types (*e.g.* X for halogen), then a set of corresponding atom key codes are created for each alternative atom value in the list and these are added to the ‘OR’ screen table. The pattern lists associated with these ‘OR’ atom keys are then updated by inserting the pattern index into each atom key list.

The bonds connected to each atom are now visited in turn in the final phase of screen table

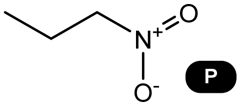

Find FGs in molecule M
1
Generate screens from target molecule M
2
Eliminate patterns using screens
3
Eliminate patterns using hierarchy
4
Set CSP to solve substructure match with no symmetrical solutions
Bind M to CSP solver as the target
Fetch next selected pattern P
Load FG patterns
Read patterns from database
Generate screens from patterns
Read pattern hierarchy from database
Finish
Patterns {1,2,..}
AND table
S {1,4,8,...}
S- {1,8,31,...}
S= {1,4,8,...}
OR table
F {34,36,...}
Cl {11,34,...}
Br {11,34,...}
1
2
3
4
54
67
7
5
78
86
Selected Patterns {1,3,8,34,38,102,168,...}
Remove children of P from selected patterns
No
match?
Yes
Find next solution using CSP solver
Duplicate FG descriptor at symmetrical locations
Build FG descriptor using P name and location
Finish
No
found?
Yes
Bind P to CSP solver as the query
Find first solution using CSP solver (find a location)
match?
Yes
No

Flow chart 3 The functional group perception routine.

preparation. A single concatenated atom/attached bond key code is constructed if both the parent atom and the bond have fixed values and this key is inserted into the ‘AND’ screen table.

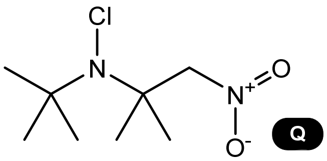
Otherwise a set of concatenated key codes is made by permuting all combinations of the atom and bond values from the list(s) and these keys are added to the 'OR' screen table. In all these cases the associated pattern list for each screen key is updated with the index of the pattern containing the atom/attached bond pair. Table 19 illustrates the outcome of this procedure when applied to a limited set of patterns (in this case just X, Y and Z).

The procedure to perform functional groups perception is outlined in Flow chart 3. This

 P	Screen Fragments	Patterns containing the selected fragments (see Table 19)

All fragments present in P ⟶
AND fragments present in P ⟶
AND fragments absent in P ⟶
OR fragments present in P ⟶
OR fragments absent in P ⟶
	{ C, C-, N, N-, N=, O, O-, O= }
	{ C, C-, O, N, N- }
	{ S, S- }
	{ N-, N=, O-, O= }
	{ S-, S=, F, F-, Cl, Cl-, Br, Br-, I, I- }
		a = { X, Y, Z }
		b = { X }
		c = { X, Y }
		d = { X, Z }

Selected patterns = a – b – (d – c) = { X, Y, Z } – { X } – ({ X, Z } – { X, Y }) = { **Y** }

 Q	Screen Fragments	Patterns containing the selected fragments

All fragments present in Q ⟶
AND Fragments present in Q ⟶
AND fragments absent in Q ⟶
OR fragments present in Q ⟶
OR fragments absent in Q ⟶
	{ C, C-, N, N-, N=, O, O-, O=, Cl, Cl- }
	{ C, C-, O, N, N- }
	{ S, S- }
	{ N-, N=, O-, O=, Cl, Cl- }
	{ S-, S=, F, F-, Br, Br-, I, I- }
		a = { X, Y, Z }
		b = { X }
		c = { X, Y, Z }
		d = { X, Z }

Selected patterns = a – b – (d – c) = { X, Y, Z } – { X } – ({ X, Z } – { X, Y, Z }) = { **Y, Z** }

Key: a = set of patterns with present AND fragments;
b = set of patterns with absent AND fragments
c = set of patterns with present OR fragments;
d = set of patterns with absent OR fragments

Table 20 Screen analysis of fragments found in structures P and Q against fragments extracted from patterns X, Y, and Z.

procedure is also used to label atoms and bonds with properties but this detail is omitted for clarity. The first step (Flow chart 3: box 1; Listing 13: line 1) generates a set of atom and atom/bond fragments from the target molecule using the same procedure described for finding the fragments in the recognition patterns. These fragments are compared to those in both the 'AND' and 'OR' screen tables to identify the subsets of present and absent fragments (Flow chart 3: box 2; Listing 13: lines 2-3). These are used to identify the corresponding sets of patterns containing these fragments by applying a fragment to pattern translation using the 'AND' and 'OR' screen tables (Listing 13: lines 4-11). The patterns that cannot possibly match the target molecule are identified as those containing AND (*must have*) fragments that are absent in the target molecule including those patterns where *all* alternatives values of a required OR (*may have*) fragment are also absent (Listing 13: line 12).

A worked example showing the outcomes of the screening logic is shown in Table 20 with an illustrative (i.e. limited) set of functional group recognition patterns applied to two sample molecules.

The pattern selection phase ends by removing all dependent patterns of the excluded patterns (Flow chart 3: box 2 and Listing 13: lines 13-15). These dependent patterns are identified by consulting the prebuilt pattern hierarchy that was loaded during program start up.

With the selected patterns at hand, the detailed pattern matching process now begins. A CSP solver is set to operate in substructure mode on the target molecule and symmetrical solutions are set to be rejected. Each selected pattern is bound to the CSP solver in turn and a solution requested. If no substructure match is possible the selected patterns are further pruned by removing any dependants of the failed pattern. Otherwise the solution map is inspected and the functional group definition atoms and group names are used to build descriptors identifying the locants of each functional group in the target molecule. An identical process is used to construct property descriptors from property definitions found in the functional group pattern.

The elimination of symmetry duplicate matches initially results in missed functional group locants. To overcome this limitation the following steps are carried out: first each functional group definition atom in the pattern is selected and all its symmetry equivalents within the pattern are excluded from further consideration; second the representative definition atom is mapped from the pattern to its location in the target molecule using the current CSP solver map. A functional group descriptor is then constructed using the definition name (e.g. ESTER) and copies of this descriptor are assigned to the mapped target atom and each of its symmetry equivalents.

The algorithm now proceeds to repeat matching the pattern finding all non-symmetrical occurrences in the target molecule, until all matches are exhausted. This process is repeated for each selected pattern until all patterns have been attempted. This concludes the functional group and property perception. At the end of the perception task the molecule is marked up with the locations of all recognised functional groups, and properties such as aromatic directing effects, electron-withdrawing or donating centres and potential nucleofuge or electrofuge leaving groups.

Results and Discussion

The performance and behaviour of functional group perception was investigated with respect to the use of fragment screening tactics and symmetry pruning. The various screening approaches were tested individually and in combination and compared to non-screened runs.

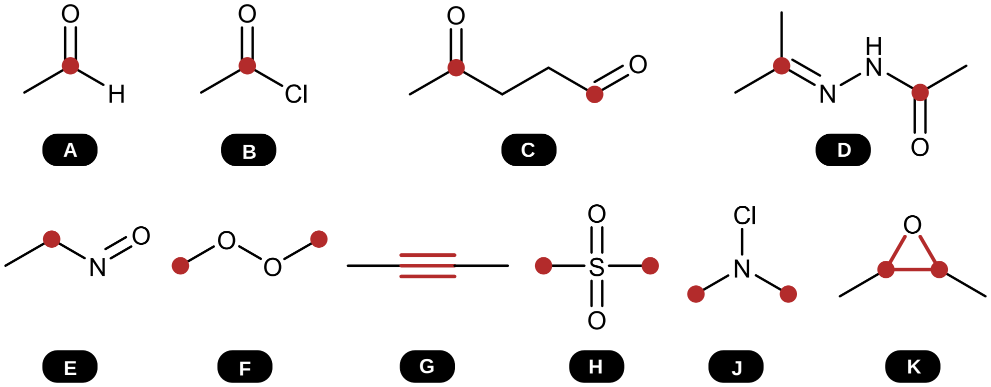


Figure 91 Structures used to measure screen efficiency of functional group perception.

Structure (Figure 91)	No screens (time ms)	Hierarchy only (time ms)	Fragments only (time ms)	Fragments + hierarchy (time ms)	Performance gain
A	10.2	8.0	3.3	1.4	7.3
B	5.0	4.1	2.8	1.6	3.1
C	5.6	4.4	2.8	1.1	5.1
D	7.1	5.9	6.0	4.1	1.7
E	5.1	3.4	3.8	1.8	2.8
F	5.0	3.2	2.9	1.3	3.8
G	5.0	3.1	2.4	0.9	5.6
H	5.3	3.7	3.3	2.0	2.7
J	5.0	3.3	3.0	1.4	3.6
K	5.5	3.7	4.0	1.9	2.9
Total time/ mean gain	58.8	42.8	34.3	17.5	3.4

Table 21 Performance times of screening in functional group perception.

The experiment used structures A to K to investigate the improvements made by the screening algorithms. Completely disabling screening forces all patterns to be tried on every molecule and gives the poorest performance as expected. The mean performance gain (the ratio of execution times) measured after eliminating patterns using the pattern hierarchy alone is 1.5, while the use of only the screen fragments gives a mean gain of 1.7. Combining the two approaches is more beneficial giving an overall mean performance gain of 3.4 (Table 21).

Most of this performance gain is due to the overall number of patterns eliminated during the pre-search screening phase as shown in Table 22. This table shows the progressive reduction of pattern numbers as each screening tactic is applied both before and during the search. The combined tactic of applying fragment screening followed by the hierarchy guided screen reduction significantly reduces the number of tested patterns in most cases. A much smaller gain is made on average during the search phase when dependants of failed patterns are removed.

The smaller gain observed during the search phase is largely due to the sparse nature of the pattern hierarchy which is dominated by a large carbonyl hierarchy with a number of smaller independent hierarchies formed by aryl and alkyl variants of non-carbonyl-based functional groups.

Structure (Figure 91)	Number of Patterns Passing Screening			Functional groups recognised (locants indicated in Figure 91)
	AND/OR fragments	Pre search w. hierarchy	During search w. hierarchy	
A	27	6	6	CARBONYL, ALDEHYDE
B	34	13	13	CARBONYL, ACID_CHLORIDE, ACYL_HALIDE
C	27	6	6	CARBONYL (2), KETONE, ALDEHYDE
D	62	48	37	CARBONYL, HYDRAZONE, ACYL_HYDRAZONE, HYDRAZONE_Z
E	49	19	15	NITROSO
F	40	13	9	PEROXIDE (2)
G	24	2	2	ACETYLENE
H	34	16	13	SULFONE (2), ALKYL_SULFONE (2)
J	54	19	17	HALOAMINE (2)
K	40	13	10	EPOXIDE (2), EPOXIDE_C_C, EPOXIDE_C_O (2)

Table 22 Screening efficiency in functional group perception as counts of selected patterns.

There is good evidence that when pre-search screening is less effective (meaning many patterns that will fail to match are not detected) the subsequent in-search hierarchy screening plays a useful role in eliminating additional non-matching patterns (*c.f.* structure D and structures E, F, J and K to a lesser extent).

The notably slower performance for the simple aldehyde when fragment screens are disabled needs a comment (see structure A, columns 2 and 3 in Table 21). The perception tasks performed on the functional group recognition patterns are triggered on first use of the pattern by the CSP substructure solver and then cached for subsequent uses of the pattern. Structure A is the first molecule tested in the experiment and consequently triggers the perception tasks on all the 258 recognition patterns tried. The processing of subsequent molecules does not invoke any further recognition pattern perception so the processing times for these molecules are much lower. When full screening is enabled these one-off perception costs are now distributed across the processing of many molecules as determined by an arbitrary sequence of first pattern uses. Consequently part of the performance gain measured is due to the absence of perception costs for the unused patterns.

Conclusions

A graphical approach to transform rule representation has been described along with the role of the PATRAN language to set atom and bond constraints. The application of the rules to build detailed transform rule hierarchies with evidence-based ratings is described and evaluated in the next chapter.

Functional group perception has been improved. The patterns are expressed in a drawing notation containing fragments of PATRAN to set the pattern match constraints. The need to carefully order functional group recognition patterns by hand has been removed. A simple but detailed screening system was introduced that copes with both AND and OR features present in the functional group patterns. A pattern hierarchy was automatically established during the construction of the functional group database and used to provide additional screening. The combination of screening and pattern hierarchies proved effective in speeding up the perception process by an average of 3 fold. The use of isomorph free pattern matching solves the problem of excess matches when mapping symmetric functional group patterns onto a target molecule.

Chapter 5

Designing Stereochemical Transform Rules

Introduction

The previous chapters have set out the principles and algorithms used to underpin the construction of stereoselective or stereospecific transform rules. This chapter describes two of the methodologies employed to create families of transform rules used to solve three problems: representing the full scope of a reaction; calculating a rating for a proposed retrosynthetic transformation best matched to the circumstances of the target molecule; and transforming a target molecule into precursor molecules while obeying any stereochemical constraints set in the rule. The latter problem is the subject of the final chapter.

An evidence-based approach using known reaction examples is used to gather support for a proposed stereoselective transformation. The reputation of a transform rule is measured by counting the number of matching examples.[40] The quality rating of a transform rule is measured using the set of reported yields and enantiomeric or diastereoselective excess values to estimate likely outcomes when applied to a target molecule. The yield and stereoselective estimates are also supported by reliability measures derived from the datasets. Additional support for a proposed transformation is provided by noting the number of examples that tolerate the functional groups found within the target molecule. Complexity reduction parameters are taken for the retrosynthetic reduction of stereocentres, cleavage of key bond types and the stereotopicity of the reaction centre. These reputation, tolerance, quality and complexity ratings will be used to rank the proposed transformation to allow comparisons to be made between alternative retrosynthetic routes (see Chapter 1).

The evidence-based approach is coupled to the use of rule networks that factor the reaction examples into linked clusters containing progressively more detailed retron environments. The choice of constraints used to define the environments is directed both by mechanistic knowledge of the reaction and by scanning the extracted examples and looking for recurring patterns in the reaction neighbourhood. This approach is designed to find rational evidence for supported and unsupported environments around the reaction centre and to improve the relevance of the supporting examples to the target molecule for route scoring purposes.

The practical design and construction of the rule networks used to extract the supporting information from reaction examples is illustrated using the enantioselective epoxidation of alkenes, the enantioselective aldol reaction and the enantioselective Stetter reaction. Rules for extracting supporting evidence for diastereoselective control are explored *via* chiral auxiliary chemistry employed in the aldol reaction.

Reaction Example Sources

A limited number of reaction databases were available for the purposes of this study. ARChem is configured to use the Elsevier (Reaxys)[qq], ChemInform (CIRX)[239] and Methods of Organic Synthesis (MOS)[rr] databases with end-users permitted to use only those databases that they have independently licensed from the database vendor. Permission was granted by Wiley to use the legacy CIRX 2010 file for this study. The large Reaxys database was not suitable for this study as the necessary stereoselectivity data was absent from the import file provided to load it into the ARChem database. Additional reaction data was available from the now discontinued commercial database REFLIB, which was loaned to this project by Accelrys Inc (now Biovia)[159] and accessed locally on a University of Leeds server. Except for REFLIB access, all database experiments were conducted on the sponsor's servers via remote login.

Database	Examples	Period	Yield (%)	CS (%)	DE (%)	EE (%)
Reaxys	10,898,402	~1850 – 2010	✓			
CIRX 2010	1,427,076	1992 – 2010	✓	✓	✓	✓
REFLIB	209,777	1968 – 1991	✓		✓	✓
MOS	42,963	1998 – 2010	✓			
Total	12,578,218					

Table 23 Available reaction databases with supporting data. Available data key: CS - chemoselectivity; DE - diastereomeric excess; EE - enantiomeric excess.

Table 23 lists the number of reactions available in each database, the publication periods covered and the availability of essential stereoselectivity information required for this study. Only CIRX 2010 and REFLIB were suitable for the stereoselective rule experiments due to the

[qq] Reaxys is accessed at http://www.elsevier.com/online-tools/reaxys

[rr] MOS is available from the Royal Society of Chemistry accessed at http://www.rsc.org/Publishing/CurrentAwareness/MOS/

availability of enantiomeric excess and diastereomeric excess values quoted in some reactions. These two databases cover a large part of the primary literature in a continuous time sequence from 1968 to 2010. This period usefully encapsulates the growth of enantioselective synthetic methodology begun within this period and continuing to the present day. [103, 240, 241]

The reaction data for both databases is available in RDFILE file format and this was readily converted into an equivalent JSON format [ss]using a custom program written specifically for this purpose (*molrxn-to-json*). The hierarchical organisation of the reaction data within the RDFILE format was strictly maintained in the conversion to JSON to retain the structure of the reaction variation data. Each reaction involving the same educts and products may refer to multiple experiments (variations) describing different reaction conditions and often employing different reagents, solvents and catalysts.

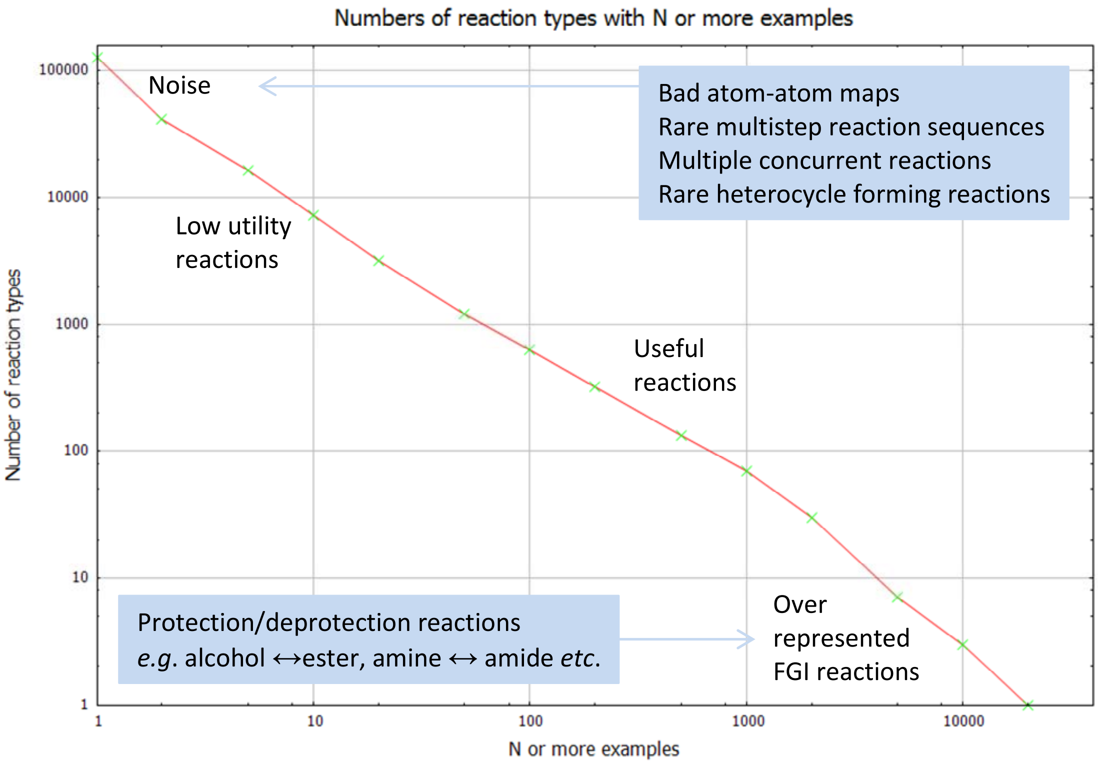


Figure 92 The distribution of distinct reaction types *versus* example counts in the reaction database.

The JSON data was loaded into a MongoDB document database using the *mongoimport* program. [tt] The loaded data was then indexed using the custom program *db-rxn-examples-*

[ss] The JSON format is formally described at http://www.json.org/

[tt] MongoDB and MongoImport are downloaded from http://www.mongodb.org/

indexer. This program first calculates CRC and SCRC Morgan codes derived from each reaction example, then stores and indexes them. The CRC codes the pattern of bond changes and the SCRC codes stereocentre modifications (*vide supra)* and are used for pre-screening the reaction examples to speed up processing times when matching examples to rules. The aggregate running time for coding and indexing 1.6 million reactions was in the order of 5 hours.

Database scripts were used to filter and eliminate unsuitable and erroneous reactions (*e.g.* those with no atom-atom maps, and those with missing essential information such as yields). The reactions examples were clustered by CRC code to count the number of distinct reaction types using database map/reduce scripts. The reaction type clusters were sorted by the number of examples in each cluster and the distribution plotted on a log/log graph (Figure 92).

Database content	# examples	Notes
Number of examples	1,636,853	Aggregate of CIRX + REFLIB
Number of unmapped examples	233,752	Reaction type unknown
Number of examples belonging to reactions with 5000 or more examples (7 reactions)	66,275	Ubiquitous protection / deprotection reactions
Number of examples belonging to reactions with 20 or less examples (123,312 reactions)	266,338	Bad atom maps (database errors) Multistep reaction sequences Rare heterocycles
Useable examples	**1,070,448**	**65 %**

Table 25 The estimated number of generally useable examples in the aggregated reaction database.

Selectivity measure	Example count	% of aggregated database
Yield	1,476,894	90
Chemoselectivity (CS)	1,125,608	69
Diastereoselectivity (DE)	49,069	3
Enantioselectivity (EE)	80,580	5

Table 24 Proportions of the aggregated reaction database with yield and chemo/stereoselectivity data.

The reactions with very low numbers of examples were sampled and it was noted that four general types of reaction were represented: incorrect atom-atom maps; conflated sequences of reactions from one-pot tandem or cascade reactions; substrates undergoing concurrent reactions at multiple reaction sites; and the construction of rare heterocycles.[uu] At the other extreme are a small number of reactions represented by a very large number of examples, the top seven being ubiquitous FGI reactions used in functional group protection/deprotection.

Reaction category	# of distinct reaction types (≥50 examples , ee ≥95%)
Enantioselective construction or transposition of stereocentres	***45***
Chiral resolutions	10
Non-racemising FGIs on enantio-enriched substrates	7
Stereospecific substitutions on enantio-enriched substrates	4
Erroneous reactions (bad atom-atom maps)	5

Table 27 Breakdown of the top 71 enantioselective reactions into categories.

Stereoselectivity (M)		**Number of examples (N)**						
% ee	**er**	**≥200**	**≥100**	**≥50**	**≥20**	**≥10**	**≥5**	**≥1**
≥ 99	≥995:5	7	12	21	62	149	355	2196
≥ 98	≥99:1	8	19	36	106	247	535	3153
≥ 97	≥98:2	11	24	44	134	296	643	3574
≥ 95	≥97:3	19	32	***71***	191	418	892	4654
≥ 90	≥95:5	25	53	109	303	599	1297	5814
≥ 80	≥9:1	35	73	153	371	757	1563	6667
≥ 60	≥4:1	46	83	181	440	900	1815	7500
> 0	>1:1	52	88	197	478	973	1931	7982

Table 26 Counts of distinct reaction types with ≥ N examples having ≥ M% quoted enantiomeric excess.

[uu] ARChem has procedures to identify and promote the importance of otherwise poorly supported heterocycle forming reactions as these are considered valuable reactions.

Table 25 shows the outcome of slicing off the reaction clusters with the highest and lowest numbers of examples to give a rough estimate of 1 million reaction examples that may usefully support most retrosynthesis transforms.

Additional data reduction was performed to isolate enantioselective reactions. Only 5% of the reaction examples quoted enantiomeric excess values (Table 24) reducing the number of useful examples for supporting enantioselective transforms to around 80,580.

The enantioselective reaction examples were clustered to count the number of reaction

Reaction type	Bond alterations	# Examples ≥95% *ee*	Notes
Addition of C nucleophiles to C=C	CH + C=C → CCCH	1603	Mostly conjugate additions
Reduction of C=O	C=O → HCOH	1553	Many types of carbonyl
Addition of C nucleophiles to C=O	CH + C=O → CCOH	1265	Includes mostly aldols, Henry reaction + alkynylations
Reduction of C=C	C=C → HCCH	1120	Wide variety of C=C environments
Addition of C nucleophiles to C=N	CH + C=N → CCNH	639	Diverse nucleophile types and methods
Epoxidation of C=C	C=C → C1CO1	415	Sharpless, Jacobsen, Shi etc.
Addition of R_3B to C=C	C-B + C=C → CCCH	329	Mostly conjugate addition to enones
Addition of R_2Zn to C=O	C-Zn + C=O → CCOH	306	Mostly addition to aldehydes
Dihydroxylation of C=C	C=C → HOCCOH	266	Sharpless dihydroxylation
Reduction of C=N	C=N → HCNH	256	Many types of C=N reduction
Diels-Alder	C=C + C=C-C=C → C1CCC=CC1	222	Carbocyclic Diels-Alder
Cyclopropanation of C=C	C=N + C=C → C1CC1	222	Via diazo precursor (carbene)
Mukaiyama Aldol	SiOC=C + C=O → O=CCCOH	210	Silyl enol ethers of ketones, esters etc.
Enolate C substitution of CBr	CH + C-Br → CC	199	Diverse nucleophile types and methods
[2+3] azomethine ylid cycloaddition	C=NCH + C=C → N1CCCC1	198	Cycloaddition of *in-situ* azomethine ylids
Addition of R_2Zn to C=C	C-Zn + C=C → CCCH	162	Mostly conjugate addition to enones
Addition of R_3B to C=O	C-B + C=O → CCOH	141	Aryl/alkyl ketones and aryl aldehydes
Oxidation of sulphides	S → S=O	137	Chiral sulphoxide products
N S_N2' substitution of allylic O	NH + C=CCO → NCC=C	131	Diverse nitrogen nucleophile types and methods
Addition of R_4Si to C=O	C-Si + C=O → CCOH	112	Vinylic, allylic and silylcyanation reactions
Mannich or Pictet-Spengler reaction	CH + NH_2 + C=O → CCN	107	Diverse reaction types and methods

Table 28 The top 21 enantioselective reaction types.

examples having a *% ee* value above a given threshold *per distinct reaction type*. These counters allow the best represented enantioselective reaction types to be selected, ranked and sampled (Table 26).

Sampling the best 71 reactions (those with ≥50 examples having ≥95% *ee*) and manually filtering out unsuitable reactions reduced the number of qualifying reactions to 45 (Table 27).

Chiral resolution reactions were discounted at this stage due to a problem in interpreting how these are represented in CIRX[vv]. It was noted that reactions with quoted *% ee* values are not necessarily enantioselective as they can be resolved substrates undergoing non-racemising or stereospecific reactions. The cluster bins set out in Table 26 indicate that the upper limit in useful enantioselective reactions represented in the aggregate database is possibly in the order of 600 reaction types (those with ≥10 examples having ≥90% *ee* values), but adjusting for the sampled usability rate this is likely to be reduced to around 400.

Table 28 details the top 21 enantioselective reactions with best reputation. This list was used as the primary guide to plan transform selection and construction.

Transform Rule Hierarchies

The structure of a stereoselective transform described in this chapter differs from earlier approaches described in the literature. The novel approach to transform representation is based on rule hierarchies and reaction example evidence. The key differences between the new approach and the older transform language methodology employed by LHASA and SECS are set out below.

The excerpt shown in Listing 14 summarises the essential features of a LHASA retrosynthetic transform (*vide supra*). LHASA and SECS transforms had a number of limitations: they were time consuming to set up, requiring a rigorous literature study of the reaction and its substrate scope by a chemist who had to then design, write, compile and test the transform statements; they only captured the reaction scope known at the time they were constructed; the ratings assigned to the transform represented a subjective consensus chosen by the chemist for the then known reaction substrate types; there was no non-subjective method to normalise ratings between

[vv] A resolution reaction is represented as a single educt molecule without drawn stereochemistry, and two stereodefined product molecules representing the resolved and differentiated products. Only *one* of these is atom mapped to the educt. This representation is incomplete and cannot be directly processed.

different transforms; the retron pattern and transform mechanism used to key the transformation and then generate the precursor had to be explicitly described in CHMTRN language statements using a complex naming scheme for identifying target atoms and bonds.

```
TRANSFORM 101
NAME Mannich Reaction

TYPICAL*YIELD                       GOOD                                        A
RELIABILITY                         FAIR                                        A
REPUTATION                          EXCELLENT                                   A
…
PATH 2 BONDS                                                                    B
GROUP*1 MUST BE AMINE* OR AMINE*3                                               B
GROUP*2 MUST BE KETONE                                                          B
…
KILL IF THERE IS A MULTIPLY BONDED ATOM ON ATOM*1                               C
…
SUBTRACT 25 IF THERE IS A LEAVING GROUP &                                       D
          ON ALPHA TO ATOM*3 OFFPATH
ADD 5 IF BOND*1 IS IN A RING                                                    D
…
          BREAK BOND1*1                                                         E
          ATTACH A CARBONYL TO ATOM*1                                           E
          BREAK BOND*1                                                          E
…
```

Listing 14 An excerpt from a LHASA transform rule for the Mannich reaction.[91] Key: A - substrate-independent nominal ratings; B – keying retron statements; C – kill qualifiers; D – rating adjustment qualifiers; E – mechanism commands.

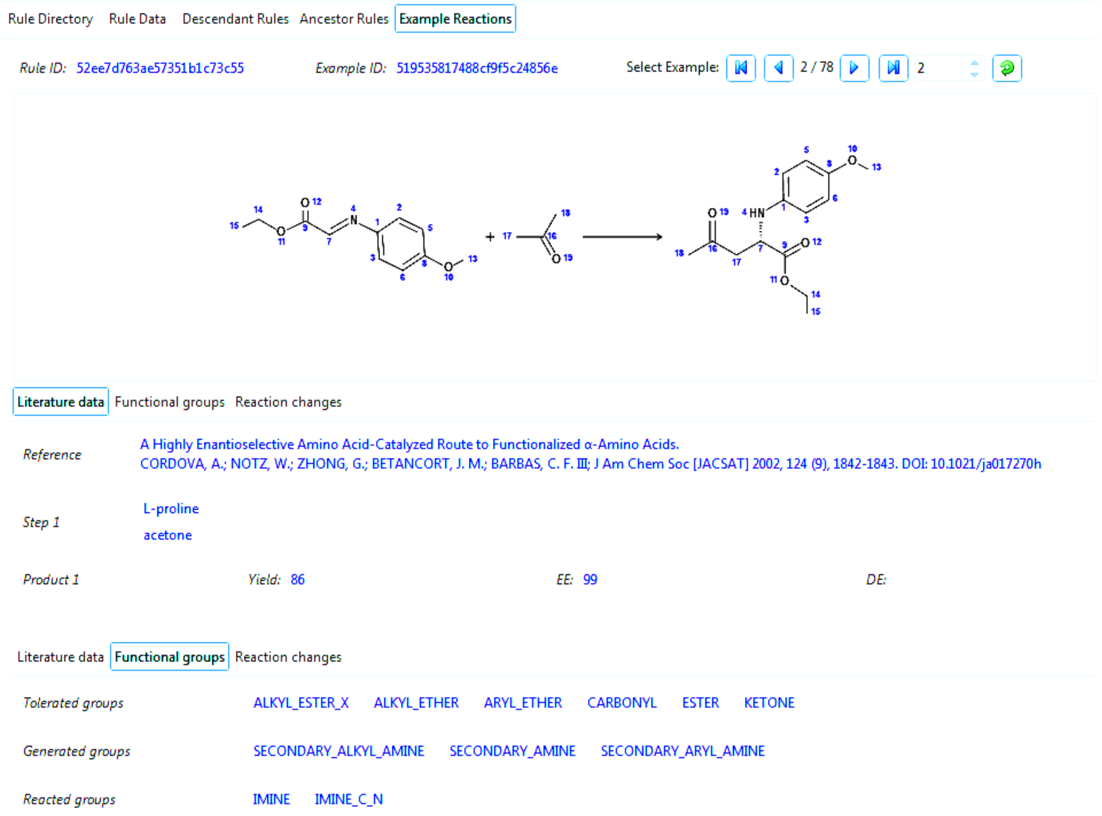
The new approach replaces or improves the sections of the classic transform in the following ways:

- A transform is now represented by a hierarchy (or a directed network) of drawn graphical reaction rules (*vide supra*). This hierarchy is used to establish the scope of the reaction using data from a database of known reactions.
- Overall substrate independent ratings and rating adjustments are abandoned and replaced by rule specific ratings automatically derived from matched reaction example data. This approach is non-subjective and normalises the calculated ratings as each transform rule is scored using the same algorithm.
- Kill qualifiers are represented by graphical rules that have no matching reaction examples.
- The keying retron statements and mechanism commands are automatically derived from the graphical reaction rule notation.

- The scope and ratings of the transform can be continuously updated by reapplying the rule hierarchy to the most current reaction databases.

Searching for Reaction Examples

A program (*rswb-find-rule-examples*)[ww] was created to find the set of reaction examples that match a reaction rule. It performs the following tasks: it accepts a single rule for processing in JSON format; converts the JSON representation into an in-memory reaction hyperstructure graph; analyses the graph to find the reaction core; generates the CRC and SCRC Morgan codes derived from the reaction core; uses the logical '*and*' combination of CRC and SCRC codes as a query to retrieve an initial subset of the examples containing the required pattern of core bond and stereochemical changes; and finally tests each retrieved reaction example candidate against



Rule ID: 52ee7d763ae57351b1c73c55 *Example ID:* 519535817488cf9f5c24856e

Reference A Highly Enantioselective Amino Acid-Catalyzed Route to Functionalized α-Amino Acids.
CORDOVA, A.; NOTZ, W.; ZHONG, G.; BETANCORT, J. M.; BARBAS, C. F. III; J Am Chem Soc [JACSAT] 2002, 124 (9), 1842-1843. DOI: 10.1021/ja017270h

Step 1 L-proline
acetone

Product 1 *Yield:* 86 *EE:* 99 *DE:*

Tolerated groups ALKYL_ESTER_X ALKYL_ETHER ARYL_ETHER CARBONYL ESTER KETONE

Generated groups SECONDARY_ALKYL_AMINE SECONDARY_AMINE SECONDARY_ARYL_AMINE

Reacted groups IMINE IMINE_C_N

Figure 93 **A screen shot of the Retrosynthesis Workbench (RSWB) showing the reaction example review page. The 'function group' and 'reaction changes' tabs show data calculated by the application.**

[ww] The 'rswb' prefix refers to the Retrosynthetic Workbench web application built for this study. It supports reaction mining, rule generation, rule hierarchy construction, viewing reaction example statistics and provides a single step retrosynthesis tool.

the reaction rule using a CSP solver configured to check if the stereochemistry constraints and PATRAN constraints are satisfied. The identifiers of the matched example reactions are cross-referenced to the rule and this cross-reference table is stored in the database. A batch mode was also implemented so that the entire set of stored transforms can be processed in one operation.

Statistics for the rule are generated that identify and count the frequency of each type of tolerated functional group found in the example set, along with the number and types of: made, broken and modified carbon-carbon, hetero-carbon and hetero-hetero bonds; created, removed, and modified stereocentres; and any observed ring modifications. These statistics contribute to the scoring parameters used to rank alternative routes in a retrosynthetic analysis (*vide infra*).

A separate program (*rswb-get-rxn-example*) retrieves a specific example reaction for presentation in the Retrosynthesis Workbench (RSWB) along with the computed statistics (Figure 93). The RSWB web application was routinely used to inspect matched reaction examples to manually validate that the rules were producing the intended results, and to observe recurring features and trends in the example sets as aids in the design of daughter rules used to divide the examples into more detailed subsets.

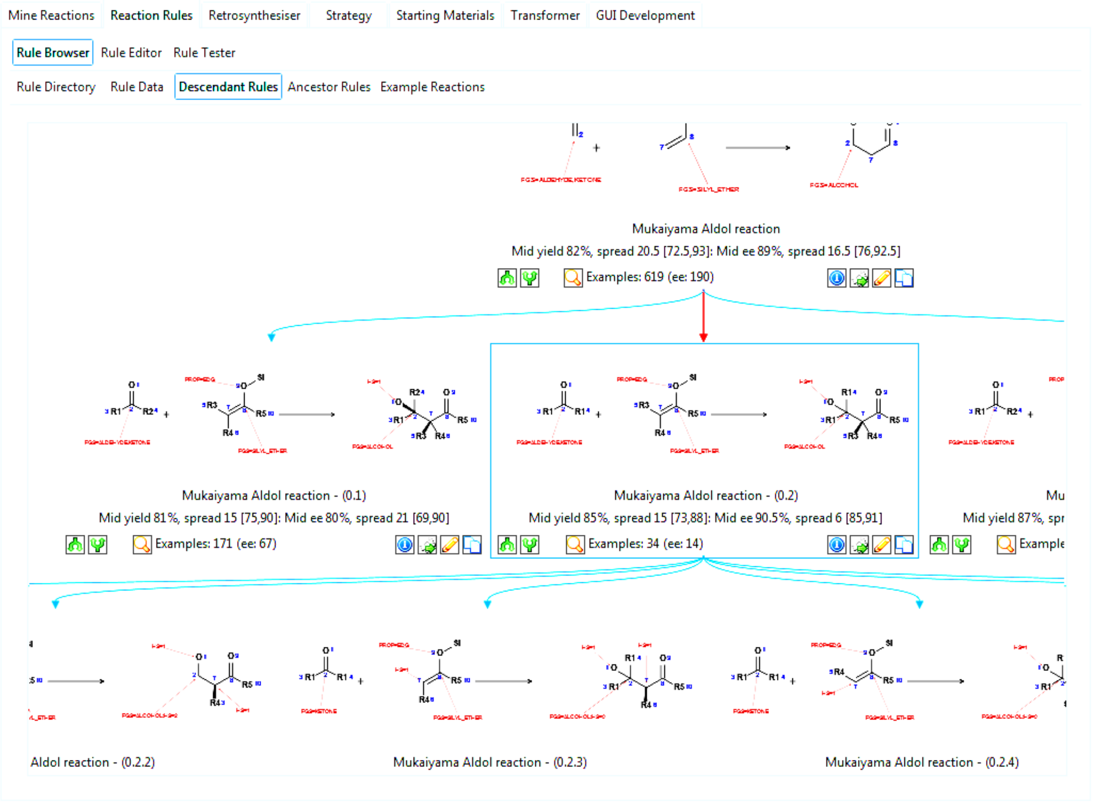
Transform Design

The transform design approach described in this section is used to assemble networks of rules that probe the outcomes of a wide variety of enantioselective, diastereoselective and stereospecific reaction types selected by data mining techniques discussed in the next chapter. The goal is to develop a retrosynthetic transform from a hierarchy of rules that find the most relevant extended retron match for a particular target molecule.

The design methodology focuses on first selecting a specific reaction type and then, using knowledge of the known synthetic methods and reaction mechanisms, to map out a tree of *required* and/or *controlling* environments around the minimum reaction substructure. These environments are coded in PATRAN annotated structural diagrams to represent individual reaction rules and these are used to find matching examples in the manner described in the previous chapter.

The design process begins with a reaction rule that defines the minimum necessary structural modifications made to the bonds (the reaction core) and then progressively introduces a network of daughter rules containing, but not limited to, the following:

- The addition of non-reacting functional groups to separate out distinct reaction types that share the same core bond modifications but rely on different types of adjacent activating or controlling functionality.
- Stereocentre constraints to establish the scope and limitations of stereoselective control or to define the stereospecific rules of the transformation.
- Substitution pattern constraints to provide additional evidence of the effects of substituents on enantioselective or diastereoselective control and selectivity.
- If appropriate, the addition of atom or bond constraints to probe:
    - The differences between intermolecular and intramolecular examples.
    - Any limitations on formed ring sizes.
    - The effect of substituent π conjugation.
    - The effect of the steric bulk of substituents (via R_L, R_M, R_S atom constraints).
    - The effect of adding specific functional groups to take the place of the generic types EWG, EDG, EF and NF.
    - Many other considerations that depend on the specific reaction type.



Mukaiyama Aldol reaction

Mid yield 82%, spread 20.5 [72.5,93]: Mid ee 89%, spread 16.5 [76,92.5]

Examples: 619 (ee: 190)

Mukaiyama Aldol reaction - (0.1)

Mid yield 81%, spread 15 [75,90]: Mid ee 80%, spread 21 [69,90]

Examples: 171 (ee: 67)

Mukaiyama Aldol reaction - (0.2)

Mid yield 85%, spread 15 [73,88]: Mid ee 90.5%, spread 6 [85,91]

Examples: 34 (ee: 14)

Mu

Mid yield 87%, spr

Example

Aldol reaction - (0.2.2)

Mukaiyama Aldol reaction - (0.2.3)

Mukaiyama Aldol reaction - (0.2.4)

Figure 94 The RSWB rule hierarchy navigation tool. This tool allows the user to inspect parent or child rules, rule parameters and the matching reaction examples.

The enumeration of stereocentre and substitution patterns should be considered a standard procedure for all stereoselective rules. After this the transform design process is conducted interactively and incrementally. Typically the reaction examples associated with a parent rule are inspected manually with the interactive RSWB application to discover possible important discriminating features that can be used to split the examples into subsets. For example, it might

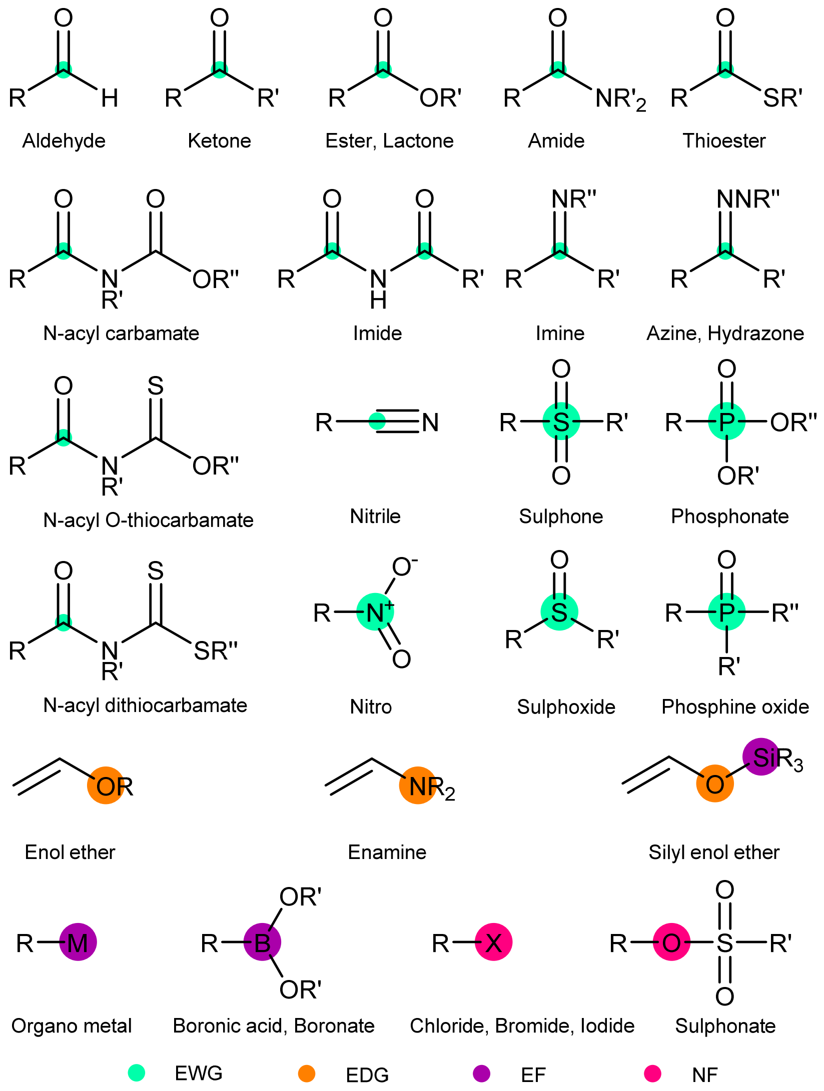


Figure 95 **EWG, EDG, NF and EF properties assigned to atoms in a selection of common functional groups.**

be apparent when sampling an example set that π conjugation between a reacting double bond and a particular substituent is highly represented. A trial split can be made on this feature to produce a pair of 'yes'/'no' daughter rules to test if the rule split should be retained.

The best outcomes occur when one branch is strongly supported by example counts while the other is poorly supported, as a likely necessary mechanistic requirement may have been identified. Alternatively the split should be retained if it separates distinct yield and/or stereoselectivity data clusters within the parent set. The retention of poorly supported branches in the rule network is important as it provides positive evidence that the associated retron is not retrosynthetically favourable based on known scope. Branches with poor support are not subject to further division. Figure 94 shows one of the tools in the RSWB application that is used to inspect the rule hierarchy in a selected transform.

Experimental Techniques

A single common definition for the special atoms types EWG, EDG, EF and NF was applied to all the transform rules. Figure 95 (*vide supra*) lists the specific functional groups that have been

Number of examples	Assigned reputation	Rationale
0 – 1	None	No evidence to support the rule. Singletons are included as many were found to be the result of incorrect atom-atom maps.
2 – 5	Very poor	Likely very low substrate diversity. High likelihood all examples are from one paper. Inadequate knowledge of the scope of tolerated functional groups is very likely.
6 – 10	Poor	A fair likelihood the examples are from a single paper suggesting low adoption (reputation) in general synthesis.
11 – 20	Fair	Likely moderate support in terms of diversity, adoption and tolerated functional group scope.
21 – 50	Good	Likely reasonable substrate diversity and evidence of adoption in general synthesis. Evidence for the most important tolerated functional groups is likely.
51 – 100	Very good	Likely to have very good substrate diversity with good evidence of adoption and utility in general synthesis. Good knowledge of a wide range of tolerated functional groups is likely.
101 and higher	Excellent	Likely excellent substrate diversity with routine adoption and utility in general synthesis. Very good knowledge of most tolerated functional groups is likely.

Table 29   The empirical support grades for rating the reputation of reaction rules are based on the numbers of qualifying reaction examples in a 7 point Likert scale.

labelled as electron-withdrawing groups (EWG), electron-donating groups (EDG) and leaving groups (NF, EF) for the transform design experiments.

The experimental results are presented using Likert-type scales[242] to indicate the likely level of support provided by the examples over a number of measured parameters (Table 29). The use of ordinal anchor phrases such as “poor”, “fair”, “good”, “very good”, “excellent” *etc*. to represent a level of support or an estimated expectation is designed to simplify rating a reaction. A colour coded scheme is used to present the results in this thesis as it provides a concise alternative to using anchor phrases (the latter is probably preferable when presenting results to the users of the retrosynthesis program). The measured parameters are: the total number of examples (reputation); the number of examples with enantiomeric excess values (enantioselective reputation); the estimated yield; the reliability of the yield estimate; the estimated enantioselectivity or diastereoselectivity; and the reliability of the stereoselectivity estimate.

The choice of bin positions and bin ranges were roughly estimated for the reputation parameter by counting and grading the number of supporting examples. The reputation grades were chosen after sampling a number of example sets in an effort to derive a simple rationale for estimating the likely substrate and tolerated functional group diversity as a function of set size. The chosen bin ranges increase using a non-linear 1-2-5 series scale[xx] with the narrowest bin holding the examples sets with the smallest numbers of reactions (Table 29). The non-linear

Yield (%)	Support rating
95 – 100	Superb
90 – 95	Excellent
80 – 90	Very good
70 – 80	Good
50 – 70	Fair
30 – 50	Poor
0 – 30	Very poor

A

ee /de (%)	er / dr	Support rating
98 – 100	99:1 – ∞:1	Superb
96 – 98	49:1 – 99:1	Excellent
90 – 96	19:1 – 49:1	Very good
80 – 90	9:1 – 19:1	Good
66 – 80	5:1 – 9:1	Fair
33 – 66	2:1 – 5:1	Poor
0 – 33	1:1 – 2:1	Very poor

B

Yield / *ee* / *de* spread (%)	Support rating
0 – 5	Superb
5 – 10	Excellent
10 – 15	Very good
15 – 20	Good
20 – 25	Fair
25 – 30	Poor
30 – 100	Very poor

C

Table 30 Empirical support grades with intervals selected to rate reaction rules by estimated yield (A) and enantiomeric or diastereomeric excess values (B). Each rating is augmented by an estimate of its reliability measured by the interquartile range (IQR) of the set of yield, % *ee* or % *de* values (C).

[xx] The 1-2-5 sequence expands as follows: 1, 2, 5, 10, 20, 50, 100 *etc.* It used as a convenient approximation for a logarithmic sequence.

scale was selected to approximately follow the initial rapid gain in information when growing the example set size and the subsequent slowdown in information gain at moderate set sizes and beyond. The chosen nominal scale accounts for the likelihood that small numbers of examples provide very little evidence of the substrate scope and functional group tolerance and that the examples often originate from a single published study. As the numbers of examples increase the likelihood grows that the method has been applied to a broader set of substrates with more types of tolerated functional groups and has been used by a larger set of authors to solve practical synthetic problems. The scale is curtailed at 100 examples as sampling showed that in most cases little additional information was gained beyond this set size.

The likely yield and enantioselectivity measure for a specific rule is estimated by taking the median values from the reported yield and % *ee*/*de* values of the matched examples. The median was chosen as a robust statistic resistant to outliers typically found in mildly contaminated datasets.[243] The nominal rating scheme for reaction yield has been adopted from Vogel's practical chemistry text book[244] while the enantiomeric excess ratings are empirically scaled using a 1-2-5 sequence matched to the stereoselective *ratios* (Table 30). Masamune has suggested that a minimum useful stereoselectivity ratio in a practical synthesis is 20:1 [245] (equivalent to a stereoselective excess value of close to 90%) so this value is anchored to the "good" – "very good" boundary. Median value statistics are not calculated for the "none" and "very poor" bins as the sample set is too small to give reliable estimates.

A simple reliability measure for the yield and *% ee* value estimates was made using the interquartile range (IQR) of the respective data sets. The IQR bounds the middle 50% of the data points and is a robust trimmed estimator with a breakdown point of 25% which is well suited to skewed distributions. [243]

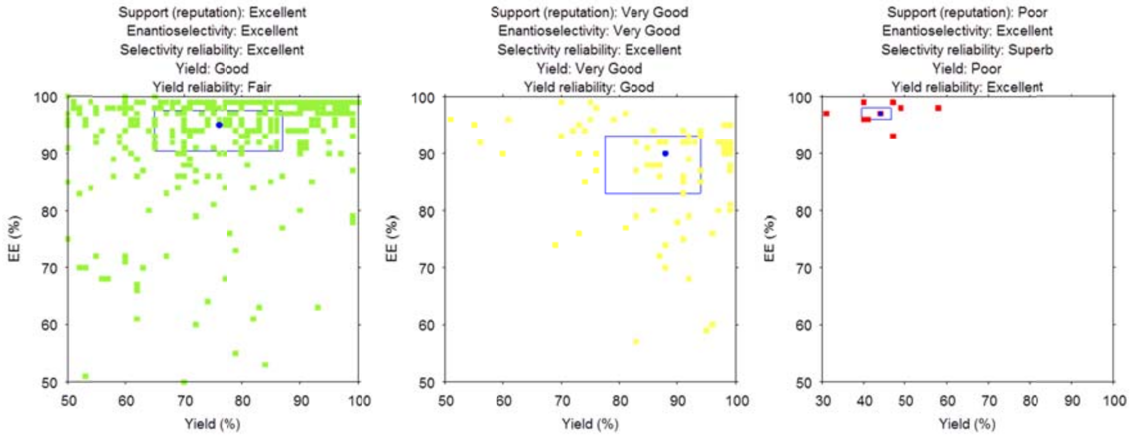


Figure 96 Stereoselectivity and yield plots of some reaction example sets.

The distribution of yields and ee values of the examples sets are conveniently visualised using 2D scatter plots to create a combined yield and enantioselectivity distribution map (Figure 96). The overlaid boxplots mark the centroid representing the median yields and stereoselectivity values while the width and height give the IQR spread for the yield and *% ee*/de values respectively.[yy] The plots enable the transform designer to observe if daughter rules are uncovering emerging data clusters when compared to the parent and sibling rule plots.

Selected Stereoselective Transforms

This section discusses methods used to design hierarchical rule sets for selecting supporting reaction examples. This includes examples of enantioselective and diastereoselective reaction types involving both stereospecific and stereoselective mechanisms. Rule development is illustrated with the enantioselective alkene epoxidation reaction, the stereoselective aldol reaction and the enantioselective Stetter reaction. The epoxidation and aldol reactions are well explored broad scoped synthetic methods with well-known substrate and substitution pattern limitations. The enantioselective Stetter reaction is a much narrower scoped reaction that is currently under development and requires careful rule design to expose its current limitations. The aldol and Stetter are carbon-carbon bond forming reactions allowing the probing of inter and intramolecular variants to be considered as part of the rule design, a consideration absent in alkene epoxidation reactions.

Enantioselective Alkene Epoxidation

Overview

The rule design process begins with a survey of the types of alkene that can be epoxidised taking into account the activating, deactivating and directing effects of neighbouring functional groups and the effects of substitution patterns.

The epoxidation of alkenes is an important reaction that introduces one or two new stereogenic centres often under stereospecific control depending on the method employed. The susceptibility of the strained 3 membered oxirane ring to facile stereospecific S_N2 ring opening reactions using a wide range of carbon and heteroatom nucleophiles makes these compounds valuable synthesis intermediates that lead to many types of 2-substituted alcohols. A wide

[yy] Note the boxplot area has no significance as yield and stereoselective excess values are rarely correlated.

variety of unfunctionalised and functionalised alkenes are readily converted to enantiopure epoxides using a diverse set of enantioselective methods. [246, 247, 248, 249, 250]

Allylic alcohols can be converted to enantiopure 2,3-epoxy alcohols using the Katsuki-Sharpless method. [247] This uses Ti(Oi-Pr)$_4$ as a catalyst with either a natural (*S,S*) or an unnatural (*R,R*) tartrate ester along with stoichiometric amounts of an oxidant, usually t-BuOOH. The stereochemistry is substrate controlled and the selected enantiomer is predicted via the simple model shown in Figure 98. The presence of the substrate alcohol group is essential as it coordinates to an intermediate chiral titanium tartrate complex which directs the delivery of an oxygen atom of a coordinated hydroperoxide ligand to a specific stereotopic face of the alkene.

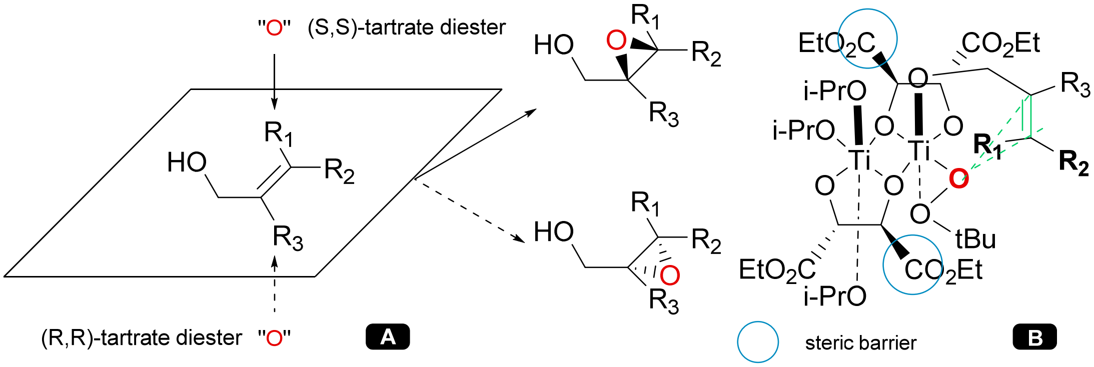


Figure 98 A) A predictive model for the outcome for the Katsuki-Sharpless epoxidation of allylic alcohols. B) A proposed transition state model used to explain enantioselective outcomes.

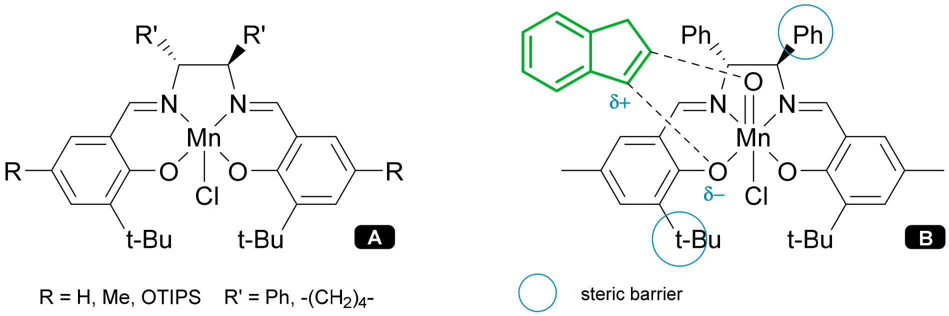


Figure 97 (A) Jacobsen Mn(III)salen enantioselective epoxidation catalysts. (B) A proposed pre-transition state assembly showing the steric and electronic factors favouring the enantioselective epoxidation of π conjugated *cis*-alkenes.

A number of intermediate titanium species that could explain the stereochemical outcome have been postulated, one of which is shown in Figure 98, B. [251, 252, 253] In this model the orientation of the tethered allylic alcohol sterically favours highest enantioselectivity with *trans* alkenes or tri-substituted alkenes without the *cis* substituent.

Jacobsen epoxidation works well with unfunctionalised alkenes and most substitution patterns are reported to give good enantioselectivity, especially unsymmetrical *cis*-disubstituted alkenes. [254] The exceptions are *trans*-disubstituted alkenes which often exhibit poor enantioselectivities. The catalyst is a Mn(III) salen complex of the type shown in Figure 97 with ligands formed from an (*S,S*) or (*R,R*) chiral diamine condensed with a salicylaldehyde derivative containing additional bulky substituents. These designed ligands provide the dissymmetric environment required to promote good enantioselectivity. A pre-transition state complex in which the substrate is tightly orientated both sterically and by electrostatic stabilisation of a partial positive charge forming on a benzylic carbon with a salen oxygen atom has been postulated to explain why mono π conjugated *cis*-alkenes are the most favoured enantioselective substrates (Figure 97). [255] The stoichiometric oxidant is commonly NaOCl (bleach) or iodosylbenzene, optionally with catalytic amounts of an amine-*N*-oxide as a rate enhancing additive.

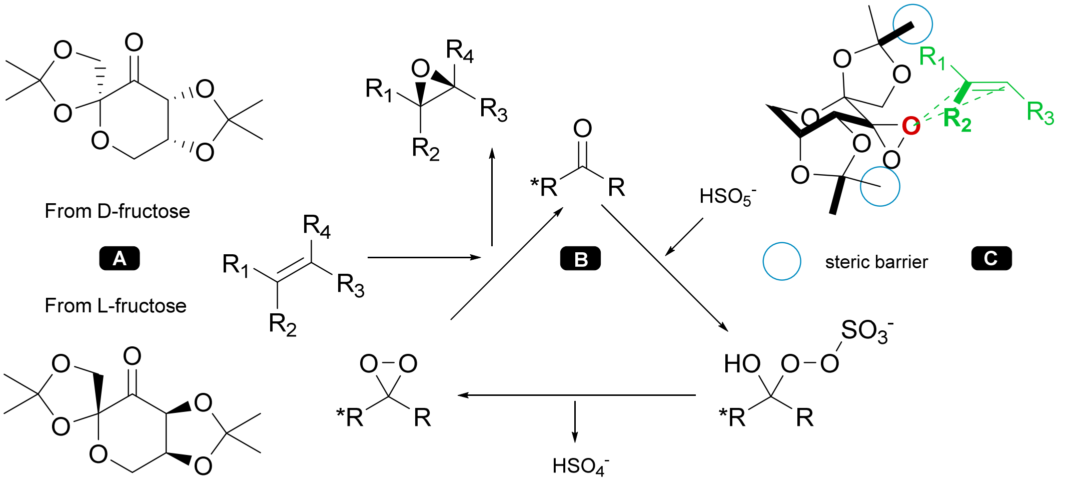


Figure 99 **(A) Fructose derived catalyst precursor ketones for the Shi epoxidation method. (B) The catalytic ketone/dioxirane cycle. (C) A proposed chiral pre-transition state assembly favouring high enantioselectivity for *trans*-disubstituted and tri-substituted alkenes.**

Shi epoxidation of unfunctionalised alkenes uses catalytic quantities of a chiral dioxirane formed *in-situ* from ketones derived from either natural D-fructose or unnatural L-fructose (prepared *via* natural L-sorbose) (Figure 99). The dioxirane intermediate is the active epoxidation agent reacting with the alkene. The stoichiometric oxidant used to form the dioxirane intermediate in the catalytic cycle is $KHSO_5$, usually in the form of the mixed salt Oxone ($2KHSO_5.KHSO_4$). The scope of the reaction complements the Jacobsen method in that unfunctionalised *trans*-disubstituted alkenes usually give the best results. [249] The proposed transition state involves spiro geometry between the sterically accessible dioxirane oxygen atom and the approaching alkene. In this model the orientation of the alkene sterically favours highest enantioselectivity with *trans* and tri-substituted alkenes. [256, 257]

A wide variety of methods exist for the enantioselective nucleophilic epoxidation of *electron-deficient* alkenes, principally α,β unsaturated carbonyl compounds. The mechanisms are either based on chiral Lewis acid activation of α,β unsaturated carbonyls[258] or involve chiral ion-pair intermediaries operating under phase transfer conditions.[259] The diverse methods include: chiral pyrroldinomethanols (Figure 100: A) that form reactive face selective chiral iminium intermediates; [260] quaternary cinchona alkaloid derivatives (Figure 100: B) operating *via* phase transfer; [261] lithium or magnesium t-butyl peroxides incorporating chiral DET ligands (Figure 100: C); [262] and chiral complexes formed from the alkoxides of yttrium or the lanthanide metals La, Gd, Sm and Yb with BINOL and triphenylphosphine oxide (Figure 100: D).[263] The stoichiometric oxidants are usually NaOCl, H_2O_2 or t-BuOOH.

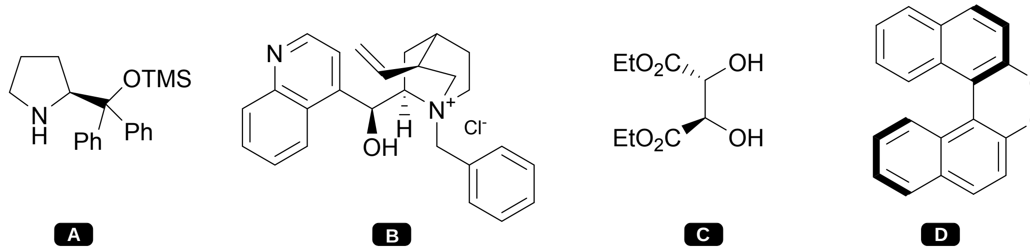


Figure 100 **Example chiral ligands and catalysts employed in a variety of methods for the enantioselective epoxidation of electron-deficient alkenes.**

Rule Development

The preceding overview provides the rule designer with a variety of functional group constraints that can be applied to divide the rule set into three main branches. These branches are designed to separate allylic alcohols, general unfunctionalised alkenes and Michael acceptors as each category has specific epoxidation methods and each of these methods in turn favours particular substitution patterns.

The details of the specific rule patterns devised for the enantioselective epoxidation of alkenes are shown in Table 31 to Table 37. The parent-daughter organisation of these rules is summarised in Figure 101.

Rule E (Table 31) is the base rule that establishes the minimum structure containing the alkene epoxidation retron and the corresponding alkene precursor. All rules are atom-atom mapped but for brevity the daughter rules are shown in the tables without the mapping numbers. The base rule defines no stereochemical or substitution constraints as it is exists to retrieve all alkene epoxidation reactions and to act as the single entry point for the epoxidation retron search during retrosynthetic analysis.

Rule E.1 (Table 31) establishes the allylic alcohol branch with the FGS=ALCOHOL constraint on one substituent atom. The EWGS=0 constraints on the alkene carbons ensures that no other substituents are electron-withdrawing. The special case of electron-deficient allylic alcohols is not covered by this rule and is handled by a separate branch (not shown). The expectation is that the E.1 branch will capture all the Sharpless epoxidation examples.

Rule E.2 (Table 31) uses the FGNOT=ALCOHOL and EWGS=0 constraint in combination to isolate predominately unfunctionalised alkenes that are not allylic alcohols. The expectation is most of the Jacobsen and Shi examples will be captured in this rule branch.

Rule E.3 (Table 31) establishes the electron-deficient alkene branch and uses a single EWG substituent atom to ensure that the alkene is a Michael acceptor while the remaining substituents are constrained via FGNOT=ALCOHOL to eliminate the special case of electron-deficient allylic alcohols.

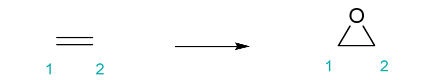
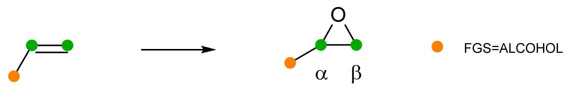
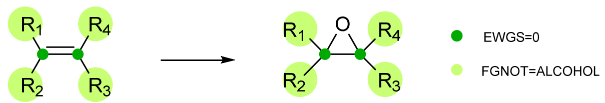
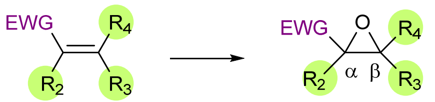
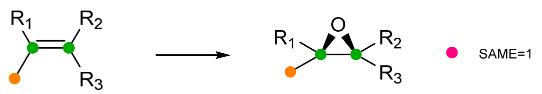
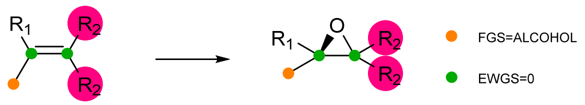
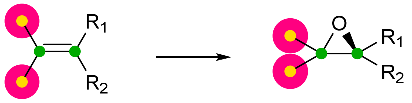
The second level of daughter rules derived from E.1, E.2 and E.3 split the examples according to the number of chiral centres in the epoxide and the relationship of these chiral centres to the activating or directing functional groups (Table 31, Table 34 and Table 36). These stereochemical constraints are directly represented in the rule diagram using wedged or hashed bond styles. The guiding principle is all expected stereocentre patterns should be represented.

The third level of daughter rules of the E.1.x. E.2.x and E.3.x branches split the examples according to all possible hydrogen substitution patterns to probe which types of substituted alkene are well supported and which are poorly supported for each type of epoxide.

The fourth level of daughter rules adds atom hybridisation constraints to the substituents to split the examples according to patterns of non-conjugated and π conjugated substituents. The unfunctionalised *cis*- and *trans*-disubstituted alkene branches (E.2.1.1 and E.2.1.2) are probed to find the level of example support provided by the Jacobsen method which is reported to favour at least one π conjugated substituent. [255]

The *rswb-find-rule-examples* program was created and used to match rules to reaction examples and count supporting examples and to calculate likely yields and *% ee* values by the methods described.

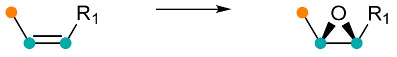
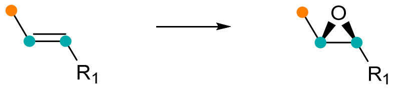
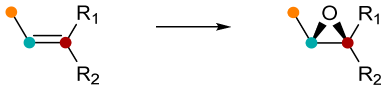
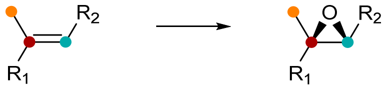
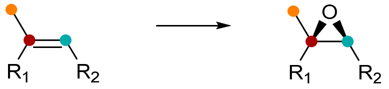
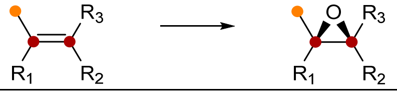
The current set of 43 alkene epoxidation rules where applied to both the CIRX and REFLIB databases to select supporting examples. The results of each rule search are available for review in the Retrosynthesis Workbench (RSWB) web application to allow the rule designer to validate the outcomes of each rule search and observe trends in the examples as an aid to develop further daughter rules. The results with measured reputations and estimated yields and enantioselectivities are tabulated and discussed in the next section.

Rule identifier	Rule Pattern	A	B	C	D	E	F	G	Notes
E		5385	1210	81	21	92	14	-	Epoxidation of all types of alkene Daughter rules are E1, E2, E3
E.1		945	249	80	22	94	8	21	Epoxidation of allylic alcohols Covers the Sharpless method Daughter rules are E.1.x
E.2	 R1 R4 R2 R3 R1 O R4 R2 R3 EWGS=0 FGNOT=ALCOHOL	3111	526	79	22	91	15	44	Epoxidation of electron rich or neutral alkenes Covers Jacobsen/Shi methods + others Daughter rules are E.2.x
E.3	 EWG R4 R2 R3 EWG O R4 R2 α β R3	1233	430	84	20	92	14	35	Epoxidation of electron-deficient alkenes Covers Julia-Colonna method + many others Daughter rules are E.3.x
E.1.1		659	187	80	21	94	6	75	Allylic alcohols Product has α,β stereocentres Daughter rules are E.1.1.x
E.1.2	 R1 R2 R2 O R1 R2 R2 FGS=ALCOHOL EWGS=0	154	40	66	28	91	7	16	Allylic alcohols Product only has an α stereocentre
E.1.3		0	0					0	Allylic alcohols Product only has a β stereocentre. No face selectivity *via* substrate control is possible using Sharpless method.

Rating Key: ● Superb ● Excellent ● Very good ● Good ● Fair ● Poor ● Very poor ● No support

Table 31 **Enantioselective alkene epoxidation rules (part 1) showing: root rule (E); allylic alcohol branch (E.1); unfunctionalised alkene branch (E.2); electron- deficient alkene branch (E.3); patterns of chiral and non-chiral centres in the allylic alcohol branch (E.1.1 – E.1.3).**

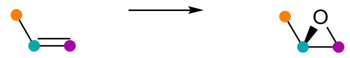
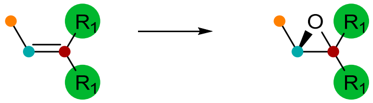
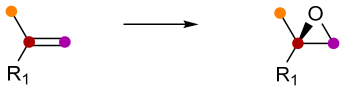
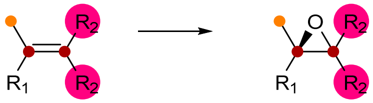
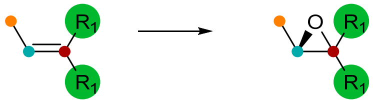
Columns:
A – Number of examples
B – Number of examples with *% ee* values
C – Estimated yield (median %)
D – Reliability of the estimated yield (IQR spread in %)
E – Estimated *% ee* value (median %)
F – Reliability of the estimated *% ee* value (IQR spread in %)
G – Percentage of parent rule with respect to B

Rule identifier	Rule Pattern	A	B	C	D	E	F	G	Notes
E.1.1.1		177	31	72	18	88	14	16	*cis* allylic alcohols 60% Sharpless, 20% other 16% Shi/Jacobsen
E.1.1.2		230	78	80	21	95	6	42	*trans* allylic alcohols 98% Sharpless 2% other
E.1.1.3		82	31	80	20	93	10	16	70% Sharpless 30% other
E.1.1.4	 ● FGS=ALCOHOL ● HS=1;EWGS=0 ● HS=0;EWGS=0	25	5	93	9			3	80% Sharpless 20% other
E.1.1.5		130	40	84	22	95	4	21	90% Sharpless 7% Shi 3% other
E.1.1.6		17	2	80	17			1	100% Sharpless

Rating Key: ● Superb ● Excellent ● Very good ● Good ● Fair ● Poor ● Very poor ● No support

Table 32 **Enantioselective alkene epoxidation rules (part 2) showing: patterns of substitutions in the allylic alcohol branch (E.1.1.1 – E.1.1.6).**

Columns:
A – Number of examples
B – Number of examples with *% ee* values
C – Estimated yield (median %)
D – Reliability of the estimated yield (IQR spread in %)
E – Estimated *% ee* value (median %)
F – Reliability of the estimated *% ee* value (IQR spread in %)
G – Percentage of parent rule with respect to B

Rule identifier	Rule Pattern		A	B	C	D	E	F	G	Notes
E.1.2.1		● FGS=ALCOHOL ● HS=2;EWGS=0 ● HS=1;EWGS=0 ● HS=0;EWGS=0	62	20	56	12	92	7	50	Terminal allylic alcohols 100% Sharpless
E.1.2.2		● SAME=1; ARYL = ANY	24	8	80	7	49	42	20	100% other Daughter rules E.1.2.2.1 and E.1.2.2.2
E.1.2.3		● FGS=ALCOHOL ● HS=2;EWGS=0 ● HS=0;EWGS=0	60	10	62	25	95	5	25	50% Sharpless 40% other 10% Shi
E.1.2.4		● SAME=1	4	1					3	
E.1.2.2.1		● SAME=1; ARYL = NO	5	4					50	β,β’ di-aryl allylic alcohols 100% other
E.1.2.2.2		● SAME=1; ARYL = YES	19	4	84	29			50	β,β’ di-alkyl allylic alcohols 100% other

Rating Key: ● Superb ● Excellent ● Very good ● Good ● Fair ● Poor ● Very poor ● No support

Table 33 **Enantioselective alkene epoxidation rules (part 3) showing: patterns of substitutions in the allylic alcohol branch (E.1.2.1 – E.1.2.4).**

Columns:

A – Number of examples

B – Number of examples with *% ee* values

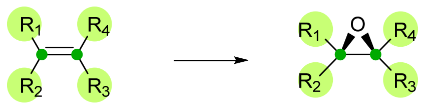
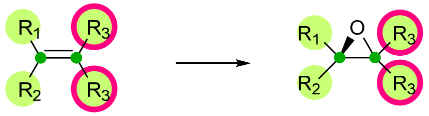
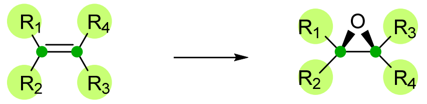
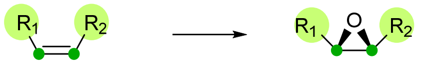
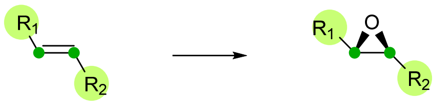
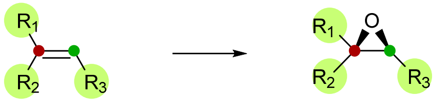
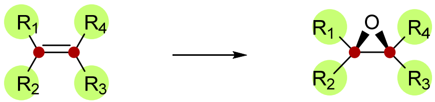
C – Estimated yield (median %)

D – Reliability of the estimated yield (IQR spread in %)

E – Estimated *% ee* value (median %)

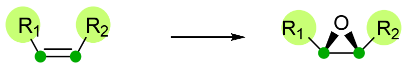
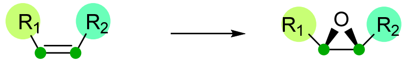
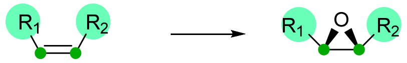
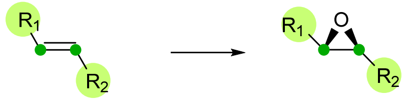
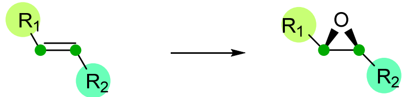
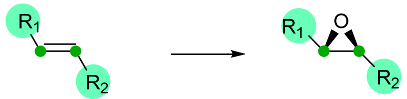
F – Reliability of the estimated *% ee* value (IQR spread in %)

G – Percentage of parent rule with respect to B

Rule identifier	Rule Pattern	A	B	C	D	E	F	G	Notes
E.2.1	 ● EWGS=0 ● FGNOT=ALCOHOL	1766	311	72	18	88	14	59	Unfunctionalised alkenes 2 stereocentres Daughter rules are E.2.1.x
E.2.2	 ● SAME=1	348	119	80	21	95	6	23	Unfunctionalised alkenes 1 stereocentre Daughter rules are E.2.3.x
E.2.3		48	18	80	21	93	11	3	2 stereocentres via 2 step mechanism with C-C bond rotation favouring R_s *cis* to R_L (occurs in the Jacobsen reaction)
E.2.1.1	 ● FGNOT=ALCOHOL ● EWGS=0; HS=2 ● EWGS=0; HS=1 ● EWGS=0; HS=0	889	149	78	22	94	9	48	Daughter rules are E.2.1.1.x
E.2.1.2		312	77	85	20	90	17	25	Daughter rules are E.2.1.2.x
E.2.1.3		470	82	81	20	92	9	26	61% Shi 22% other 17% Jacobsen
E.2.1.4		94	3	58	27	⬛	⬛	1	

Rating Key: ● Superb ● Excellent ● Very good ● Good ● Fair ● Poor ● Very poor ● No support

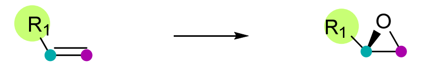
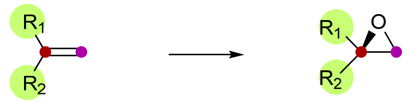
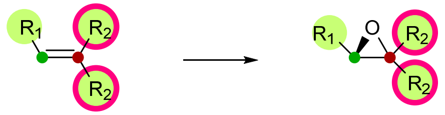
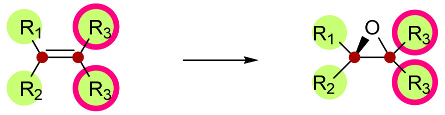
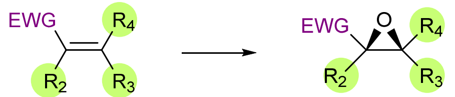
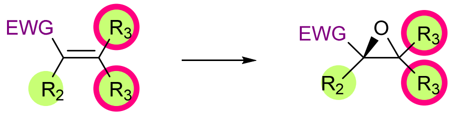
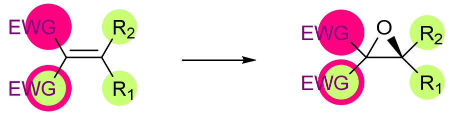
Table 34 **Enantioselective alkene epoxidation rules (part 4) showing: patterns of chiral centres in unfunctionalised alkenes (E.2.1 – E.2.2); a two-step mechanism with C-C bond rotation (E.2.3); substitution patterns in the E.2.1.x branch. Columns key: see Table 33.**

Rule identifier	Rule Pattern	A	B	C	D	E	F	G	Notes
E.2.1.1.1		17	0					0	No enantioselective support
E.2.1.1.2	 ● FGNOT=ALCOHOL;SPS=1,2 ● FGNOT=ALCOHOL;SPS=3 ● EWGS=0; HS=1	251	120	78	22	94	8	80	Excellent steric differentiation of substituents 66% Jacobsen 28% other 6% Shi
E.2.1.1.3		582	29	79	20	91	21	20	Non-enantioselective epoxidation of dialkyl cis-alkenes is very well supported in contrast to the enantioselective variants.
E.2.1.2.1		84	20	85	22	88	25	26	65% other 20% Jacobsen 15% Shi
E.2.1.2.2	 ● FGNOT=ALCOHOL;SPS=1,2 ● FGNOT=ALCOHOL;SPS=3 ● EWGS=0; HS=1	87	28	73	24	91	13	36	Excellent steric differentiation of substituents 65% Shi 25% other 10% Jacobsen
E.2.1.2.3		120	20	87	7	89	23	26	65% Shi 35% other 0% Jacobsen

Rating Key: ● Superb ● Excellent ● Very good ● Good ● Fair ● Poor ● Very poor ● No support

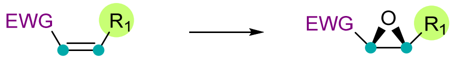
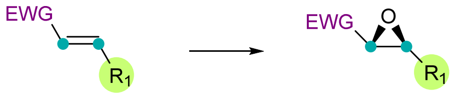
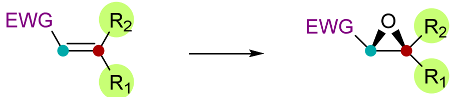
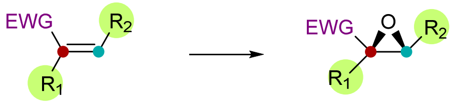
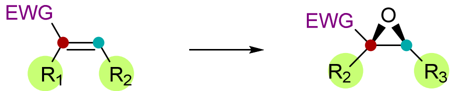
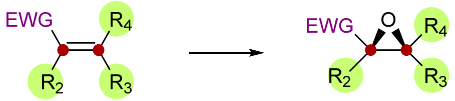
Table 35 **Enantioselective alkene epoxidation rules (part 5) showing: patterns of sp/sp^2 and sp^3 substituents in the E.2.1.1.x (*cis*-alkenes) and in the E.2.1.2.x (*trans*-alkenes) branches.**

Columns: **A – Number of examples** **B – Number of examples with *% ee* values**
C – Estimated yield (median %) **D – Reliability of the estimated yield (IQR spread in %)**
E – Estimated *% ee* value (median %) **F – Reliability of the estimated *% ee* value (IQR spread in %)**
G – Percentage of parent rule with respect to B

Rule identifier	Rule Pattern	A	B	C	D	E	F	G	Notes
E.2.2.1		141	69	78	18	83	21	58	84% other (predominately enzymatic) 6% Jacobsen 6% Shi
E.2.2.2	 ● FGNOT=ALCOHOL ● EWGS=0; HS=2 ● EWGS=0; HS=1 ● EWGS=0; HS=0	116	22	77	26	89	14	19	75% other (predominately enzymatic) 25% Shi 0% Jacobsen
E.2.2.3	 ● SAME=1	63	21	92	15	90	25	18	53% Shi 33% other 13% Jacobsen
E.2.2.4		20	7	72	32	82	17	5	100% other
E.3.1		885	392	84	19	92	13	91	Electron-deficient alkenes α,β stereocentres Daughter rules are E.3.1.x
E.3.2	 ● FGNOT=ALCOHOL ● SAME=1	22	12	78	13	92	12	3	Electron-deficient alkenes α stereocentre
E.3.3		9	5	92	29			1	Electron-deficient alkenes β stereocentre

Rating Key: ● Superb ● Excellent ● Very good ● Good ● Fair ● Poor ● Very poor ● No support

Table 36 **Enantioselective alkene epoxidation rules (part 6) showing: substitution patterns in the unfunctionalised alkene E.2.2.x branch; chiral centre patterns in electron-deficient alkenes (E.3.1 – E.3.2). Columns key: see Table 35.**

Rule identifier	Rule Pattern	A	B	C	D	E	F	G	Notes
E.3.1.1		129	24	68	21	93	17	6	30% lanthanide/BINOL/Ph_3(P,As)O catalysts. 17% other 17% Jacobsen 13% chiral quaternary ammonium catalysts. 13% chiral peroxides
E.3.1.2		469	302	85	17	93	11	77	36% lanthanide/BINOL/Ph_3(P,As)O cat. 22% Julia-Colonna (poly-L-leucine cat.) 15% pyrrolidine catalysts *etc.* 15% other 12% chiral quaternary ammonium catalysts.
E.3.1.3	 ● FGNOT=ALCOHOL ● HS=2 ● HS=1 ● HS=0	82	22	80	25	82	26	6	40% chiral peroxides 25% other 20% Shi 15% chiral quaternary ammonium catalysts.
E.3.1.4		93	29	90	22	79	18	7	48% chiral peroxides 36% pyrrolidine catalysts *etc.* 12% other 8% chiral quaternary ammonium catalysts.
E.3.1.5		147	42	90	21	84	13	11	27% pyrrolidine catalysts *etc.*. 20% Julia-Colonna (poly-L-leucine cat.) 17% chiral peroxides 17% lanthanide/BINOL/Ph_3(P,As)O cat. 7% chiral quaternary ammonium catalysts. 7% other
E.3.1.6		25	1					0	

Rating Key: ● Superb ● Excellent ● Very good ● Good ● Fair ● Poor ● Very poor ● No support

Table 37 **Enantioselective alkene epoxidation rules (part 7) showing: substitution patterns in the electron-deficient alkene E.3.1.x branch. Columns key: see Table 35.**

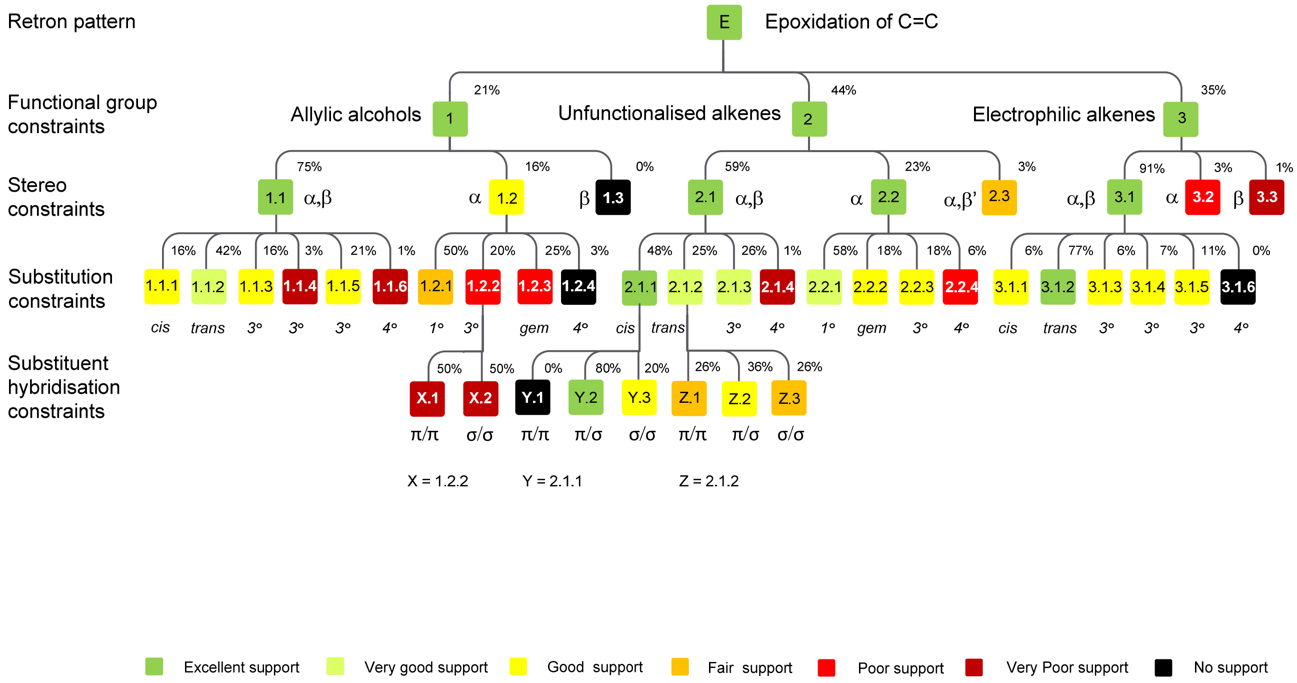


Figure 101 The enantioselective alkene epoxidation rule hierarchy with evidence-based support ratings.

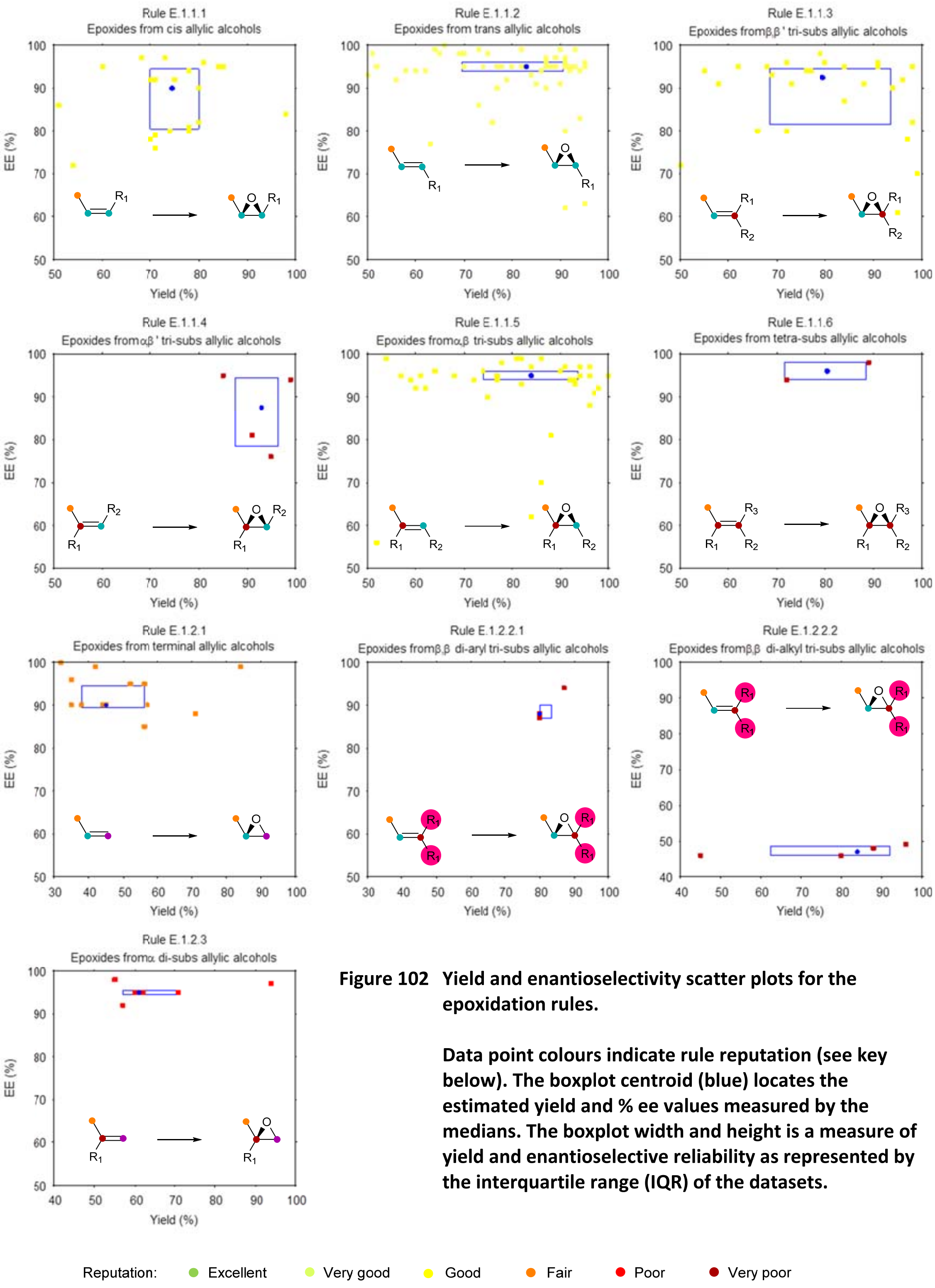


Figure 102 **Yield and enantioselectivity scatter plots for the epoxidation rules.**

Data point colours indicate rule reputation (see key below). The boxplot centroid (blue) locates the estimated yield and % ee values measured by the medians. The boxplot width and height is a measure of yield and enantioselective reliability as represented by the interquartile range (IQR) of the datasets.

Results and Discussion

The counts of total reaction examples and examples with good *% ee* values are shown for each rule in Table 31 through to Table 37. The example support counts (reputation), likely reaction yields and likely enantiomeric excess values are colour coded with a 'traffic-light' scheme to indicate the empirically derived support levels. Figure 101 summarises the parent-daughter rule hierarchy and is colour coded to indicate the measured reputation of each rule.

Alkene epoxidation is well represented in the source databases with an aggregate total of 5385 examples of which 1210 have useful *% ee* values. The rules with the best overall reputation, stereoselectivity and reliability are the epoxidation of:

- *trans*-allylic alcohols (rule E.1.1.2 – *% ee* median: 95%, IQR: 91 – 97%).[zz]
- *cis* aryl-alkyl unfunctionalised alkenes (rule E.2.1.1.2 – *% ee* median: 94%, IQR: 90 – 98%).
- *trans* Michael acceptors (rule E.3.1.2 – *% ee* median: 93%, IQR: 85 – 97%).

Each set of reaction examples was manually reviewed and tallies made for the principle synthetic methods employed. These surveys are reported in the tabulated results in the "notes" columns. An automated method for clustering reaction examples by synthetic method has not yet been developed due in part to the difficultly of recognising relationships between catalyst and reagent variants, and reaction conditions when represented in textual form.

The epoxidation of allylic alcohols is represented by 945 examples of which 249 have useful *% ee* values. A survey of the reaction examples for the rules that introduce two stereogenic centres (E.1.1.1 through E.1.1.6) shows that the Sharpless method dominates. The predominance of a single method allows a tentative comparative analysis to be made of the observed enantioselectivities against the measured substitution patterns under the assumption that a common transition state model dictates the reported *% ee* values:

- Rules E.1.1.2 and E.1.1.5: the <u>presence</u> of a *trans*-substituent (with respect to the allylic alcohol group) in conjunction with the <u>absence</u> of the *cis* substituent ensures a very good enantioselectivity (ee median: 95%) with excellent reliability (ee IQR: 91 – 97%).
- Rules E.1.1.1 and E.1.1.3: the <u>presence</u> of a *cis*-substituent in conjunction with the <u>absence</u> of the *gem* substituent reduces the reliability of the reaction (ee IQR: 80 – 95%)

[zz] IQR is the interquartile range.

as well as reducing the likelihood of obtaining as high a selectivity value (ee median: 88 and 93% respectively).

- Rules E.1.1.4 and E.1.1.6: the concurrent presence of the *cis*- and *gem*-substituent is very poorly supported in terms of examples suggesting that this combination is not normally successful with respect to generating the desired product.

Caution must be applied with the latter interpretation as other likely explanations for low numbers of examples include a lack of effective routes to the required precursor molecule or a lack of synthetic value in generating products with a particular substitution pattern.

The effects of *trans*- and *cis*-substituents in the mined data observations are presented in the example reaction yield/enantioselectivity scatter plots (Figure 102). These plots indicate that the *cis* substituent lowers the selectivity of the two competing enantioselective Sharpless transition states by reducing the Gibbs free energy difference ($\Delta\Delta G^{\ddagger}$) between them from approximately 1.8 – 2.6 kcal mol^{-1} (at 25 ^{o}C)[aaa] for *trans*-substituted allylic alcohols to 1.2 – 2.0 kcal mol^{-1} for *cis*-substituted allylic alcohols. These substituent effects encoded *via* the rule ratings are thus directly accessible to the ARChem retrosynthesis executive for comparing and rating alternative routes.

The epoxidation of unfunctionalised alkenes is represented by 3111 examples of which 526 have useful *% ee* values. The best rules in order of decreasing reputation are the epoxidation of:

- *Cis*-aryl/alkyl alkenes, predominantly by the Jacobsen method (rule E.2.1.1.2 – *% ee* median: 94%, IQR: 90 – 98%).
- *Trans*-aryl/alkyl alkenes, predominantly by the Shi method (rule E.2.1.2.2 – *% ee* median: 91%, IQR: 82 – 95%).
- Trisubstituted alkenes, predominantly by the Shi method (rule E.2.1.3 – *% ee* median: 92%, IQR: 86 – 95%).
- Monosubstituted alkenes, predominately using enzymatic methods (rule E.2.2.1 – *% ee* median: 83%, IQR: 73 – 94%).

Rules E.2.1.1.1 to E.2.1.1.3 enumerate the combinations of alkyl and aryl substituents in *cis*-alkenes. The mined example data provides useful insights for the retrosynthesis executive about available support for the epoxide disconnection to the *cis*-alkene:

[aaa] The energy ranges are derived from the IQR of the observed % *ee* values.

- There is no support for the retrosynthetic generations of di-aryl *cis*-alkenes. This gives the retrosynthesis executive the necessary evidence to reject this disconnection if the E.2.1.1.1 retron is matched. Note that the absence of rule E.2.1.1.1 in the rule hierarchy would result in a successful match to the parent *cis*-alkene rule E.2.1.1 which would normally suggest a potentially viable disconnection[bbb] to a generalised *cis*-alkene. This observation emphasises the importance of including rules with negative or poor outcomes to prevent inappropriate scoring.
- Even though the non-enantioselective disconnection to di-alkyl *cis*-alkenes has an excellent reputation (582 examples) the enantioselective counterpart has much weaker support (only 29 examples) and comparatively poor enantioselective reliability (ee IQR: 74 – 95%). The weaker support and reliability parameters for rule E.2.1.1.3 will demote this retrosynthetic disconnection below higher ranking matched retrons and this transformation would only come to prominence if better rules are not matched.

The data mining capabilities suggest uses beyond retrosynthetic analysis. Detection of rules with a high general reputation and low enantioselective reputation offers a method to alert chemists to those heavily exploited reactions performed on certain substrate classes that have no adequate enantioselective counterpart. A ranked list of these problems could identify key enantioselective reactions requiring further study and development.

Similar observations were made for the *trans*-alkene counterparts. Some support for di-aryl and di-alkyl alkenes is present but the enantioselective reliability of these reactions (ee IQR: 71 – 96 and 70 – 93% respectively) is lower than for aryl-alkyl alkenes (ee IQR: 82 – 95%) in line with the expected lower steric differentiation between similar substituent types. The observed results are in line with the trends reported in the review literature. [254]

Rule E.2.3 tests support for a two-step pathway where the intermediate species undergoes C-C bond rotation before the epoxide product is formed. This outcome has been reported for the Jacobsen reaction where quaternary ammonium halide additives favour the conversion of certain *cis*-alkenes to *trans*-epoxides. [264] The data mining evidence indicates that less than 3% of unfunctionalised alkene epoxidation examples are produced *via* this apparent two-step

[bbb] Non-terminal rules are penalised if matched to a target molecule as this implies the presence of an unmatched feature not covered by any daughter rule. The penalty is derived from the number of residual examples remaining after subtracting all daughter rule examples.

pathway. However comparison of the database examples to the original literature revealed that most were the result of erroneously drawn stereo bonds.

The epoxidation of electron-deficient alkenes is represented by 1233 examples of which 430 have useful *% ee* values. A wide variety of methods were apparent when reviewing the mined examples including: the use of chiral ligand-metal peroxides; alkoxy-lanthanide / BINOL / triphenylphosphine oxide catalysts with TBHP (Shibaski); phase transfer catalysis using chiral ammonium salts derived from Cinchona alkaloids; Julia-Colonna type methods with polyamino acid catalysts; as well as a small number of examples of the Shi and Jacobsen methods.

The rule with the best reputation is the epoxidation of *trans*-electron-deficient alkenes to form *trans*-epoxides (rule E.3.1.2 – *% ee* median: 93%, IQR: 86 – 97%). Only modest support exists for the formation of *cis*-epoxides from *cis*-electron-deficient alkenes (rule E.3.1.1 – *% ee* median: 93%, IQR: 79 – 96%) which exhibit lower and less reliable yields and lower enantioselective reliability. The generation of *trans*-epoxides from *cis*-electron-deficient alkenes was tested using rule E.3.1.7 (Figure 103). Evidence for this outcome was negative with no enantioselective examples found.

E.3.1.7

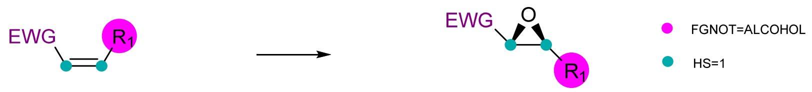


Figure 103 Rule E.3.17 selects examples of the formation of *trans*-epoxides from *cis*-electron-deficient alkenes via a two-step pathway.

The Stereoselective Aldol Reaction

Overview

The aldol reaction is one of the most widely used reactions for the construction of carbon-carbon bonds with the concurrent controlled construction of up to two stereogenic centres.[265, 266] The Zimmerman-Traxler model[267] is used to explain the diastereoselective control observed when reacting a preformed Z or E metal or boron enolate with an aldehyde or, less reliably, a ketone. The model invokes a highly organised six-membered transition state involving the enolate and aldehyde components assembled in a chair conformation, though boat conformations are known to be involved if the metal has octahedral coordination. [268] The key controlling elements in the model are the chelation of the aldehyde oxygen atom with the enolate metal/boron atom to bring the reacting partners together into the necessary pre-transition state conformation and a preferred orientation of the aldehyde to minimise repulsive 1,3 diaxial strain in the transition state (Figure 104). The Zimmerman-Traxler model is rarely applicable to aldol reactions mediated by non-metal catalysts (*vide infra*).

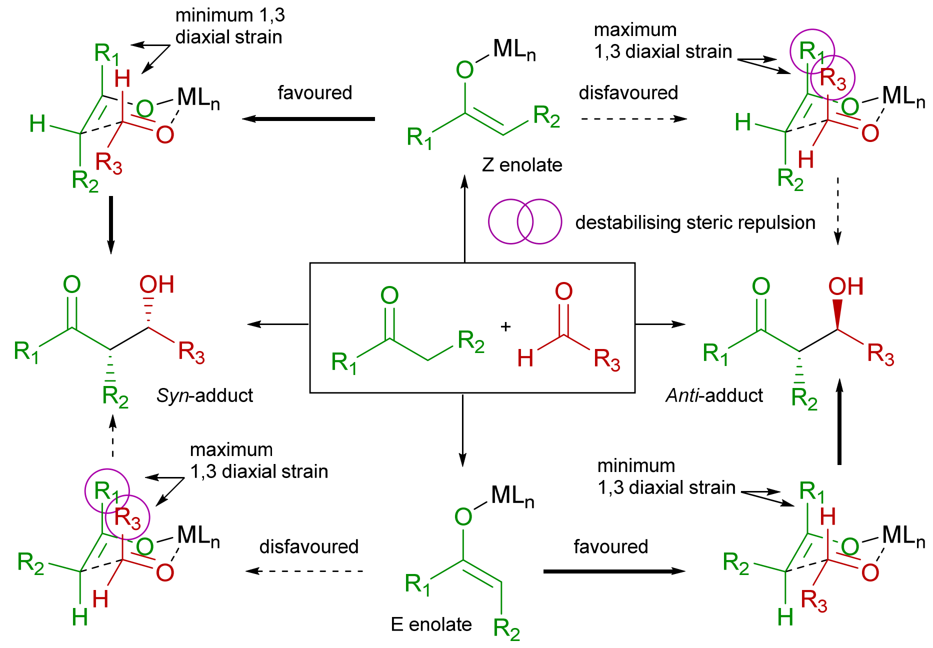


Figure 104 The origin of diastereoselective control in the aldol reaction *via* the Zimmerman-Traxler cyclic chair transition state. *Z*-enolates favour *syn*-adducts; *E*-enolates favour *anti*-adducts.

There has been extensive development over the last 30 years or so of effective enantioselective methods that include the use of removable chiral controlling groups (chiral auxiliaries); [269, 270] chiral ligands bound to the enolate metal/boron atom; [271, 272] and the use of metal free organocatalysts. [266, 273, 274]

The development of strategies using chiral auxiliaries were amongst the earliest regularly employed methods for the enantioselective aldol reaction. The auxiliary method was extensively explored by Evans,[275] Heathcock [276] and Crimmins [277, 278] and a number of tactics

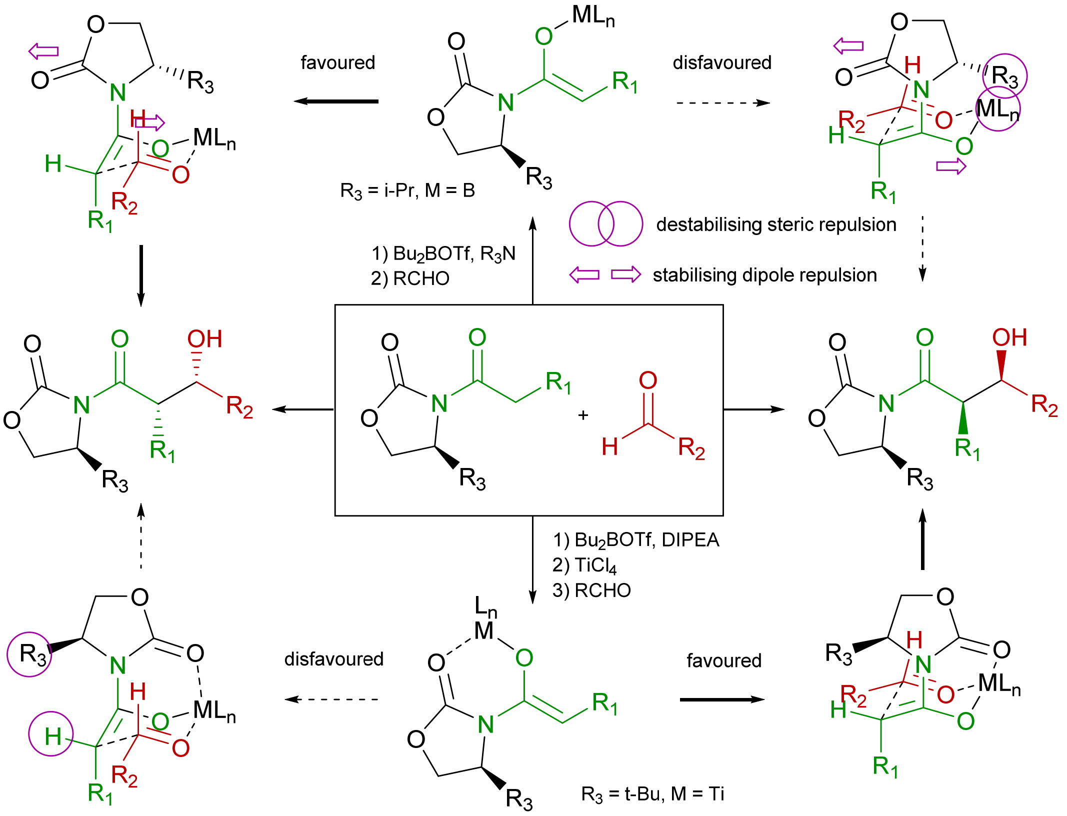


Figure 105 The origin of enantioselective control in the aldol reaction utilising oxazolidinone chiral auxiliaries. Bicoordinate *Z*-enolate complexes favour the Evans *syn*-adduct. Tricordinate *Z* -enolate complexes favour the 'non-Evans' *syn*-adduct.

were developed to control the formation of all four possible *syn* and *anti* enantiomeric aldol products (isolated after the auxiliary was removed). Figure 105 details the mechanism of formation of so-called "Evans *syn*" and "non-Evans *syn*" adducts via intermediate boron and titanium Z-enolates, operating with and without chelation control.[270] A method devised by Evans for the formation of the *anti*-adducts employs octahedral coordinated magnesium enolates that favour a boat transition state. [268] Chelation and non-chelation control using magnesium enolates was switched on or off by substituting an oxazolidinethione (or thiazolidinethione) for the oxazolidinone auxiliary to take advantage of the affinity of magnesium to chelate to oxygen but not so readily to sulphur.

Over the last two decades the organo-catalysed direct aldol reaction has come to prominence and has largely overtaken chiral auxiliary methodologies due to its greater efficiency in avoiding the additional steps required to install and remove the chiral controller. Examples include the use of catalytic amounts of chiral secondary amines to form intermediate iminium species that can be readily deprotonated to operate as a reactive enamine substitute for the aldol enolate partner (Figure 106).[279] The array of usable chiral organocatalysts and techniques for the aldol reaction is extensive and includes chiral primary amines, secondary amines, ionic liquids, [280, 281] and immobilised catalysts [282] amongst many others.

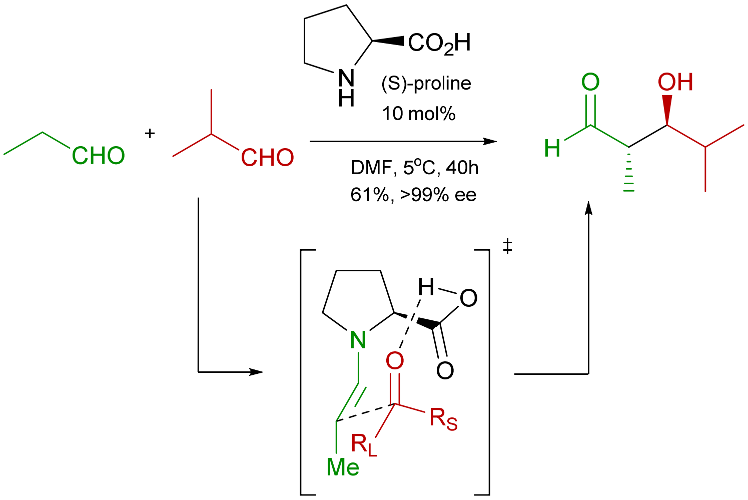


Figure 106 The stereoselective outcome of (S)-proline catalysed direct aldol reaction is predicted by the Houk-List transition state model. [52]

The Mukaiyama aldol is a modified variant well suited to directed cross-aldol reactions in which the enolate partner is replaced by a less reactive but relatively stable and isolable silyl enol ether group. The preformed enolate equivalent avoids the side reactions that plague regular aldols such as self-aldol, inverse cross-aldol and polymerisations. Activation of the aldehyde or ketone partner in the enantioselective Mukaiyama modification is effected using a variety of chiral metal Lewis acids.[283]

Rule Development

The construction of the rule decision tree follows the same principles established for the alkene epoxidation. The root rule (Table 38: rule A) establishes the minimum requirements of the retron consisting of a ketone or aldehyde component reacting with an activated carbon atom capable of being deprotonated. Note that the Mukaiyama aldol variant is handled by a separate rule hierarchy as it involves a transform retron that is distinct from the regular aldol retron.

The second level in the hierarchy separates the tree into distinct reaction types differentiated by the type of electron-withdrawing functional group that activates the nucleophilic carbon atom. The tree is divided into the Henry reaction (rule A.1: EWG is NO_2), the aldol reaction (rule A.2: EWG is C=O) and aza-aldol reaction (rule A.3: EWG is C=N[ccc]) branches of which only the aldol branch is described here.

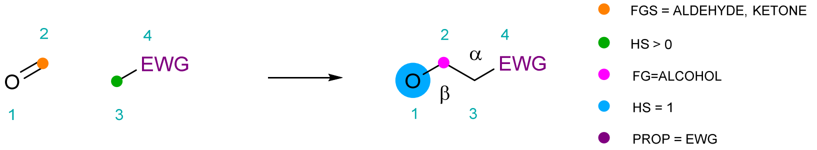
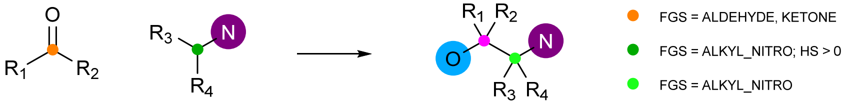
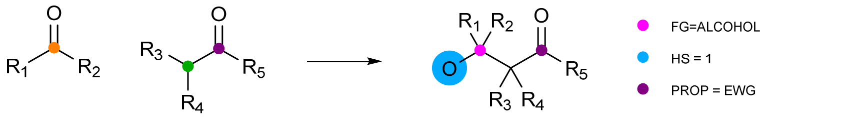
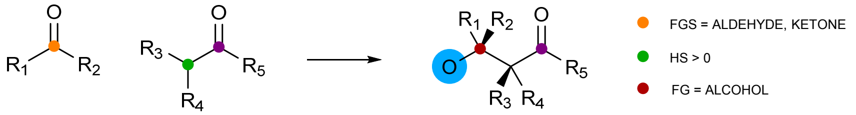
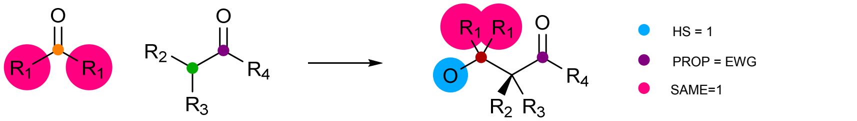
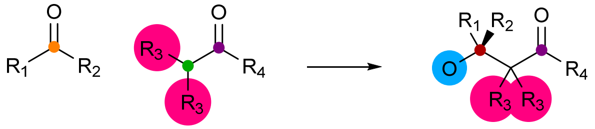
The third level in the rule hierarchy divides the examples into those that create two stereogenic centres (rule A.2.1), one stereogenic centre α to the product carbonyl group (rule A.2.2) or one stereogenic centre β to the product carbonyl group (rule A.2.3).

The fourth level enumerates all possible substitution patterns. Rule A.2.1 heads the branch containing *syn*- and *anti*-adducts formed from aldehydes and ethyl (or higher) carbonyl compounds as well as adducts from asymmetric ketones. Rule A.2.2 heads a minor branch that selects reactions with formaldehyde or symmetric ketones. Rule A.2.3 heads the branch containing adducts formed from methyl carbonyl compounds or symmetrical α branched carbonyl compounds.

The aldol is a two component reaction so the fifth level uses the bond constraint RINGS=NO and RINGS=YES to divide the substitution rules into inter and intramolecular branches. Inter and intramolecular reactions variations often place different demands on the formation of the transition state. In particular the number of bonds between the reacting functional groups in intramolecular reactions determines if the active transition state is attainable.

The sixth level enumerates the specific electron-withdrawing functional groups in the nucleophilic partner to determine which types are best supported. The daughter rules of the intermolecular aldehyde *syn*- and *anti*-aldol reactions A.2.1.1.1.X and A.2.1.2.1.X (Table 43: *vide infra*) include rules that cover the Evans type auxiliary methodologies using oxazolidinones, oxazolidinethiones and thiazolidinethiones. Table 42 lists the A.2.1.1.1.X and A.2.1.2.1.X daughter rules that seek supporting evidence for ketone, aldehyde, amide and ester derived enolate nucleophiles.

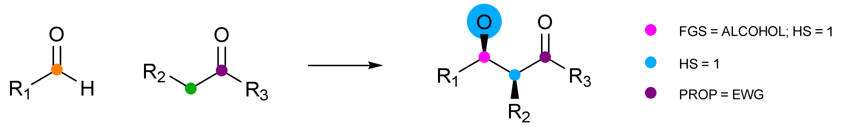
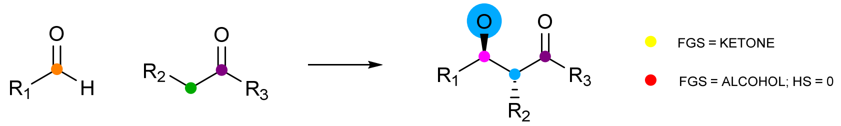
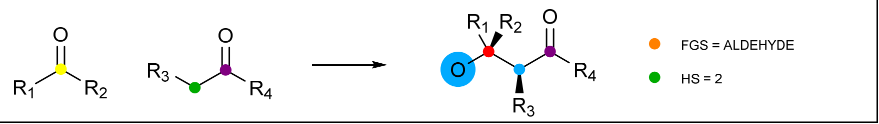
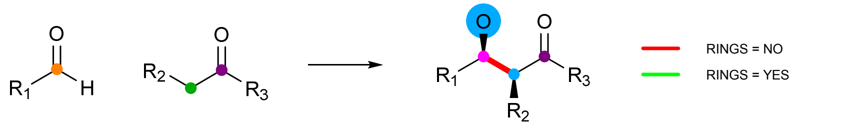
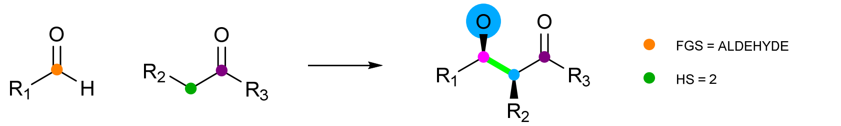
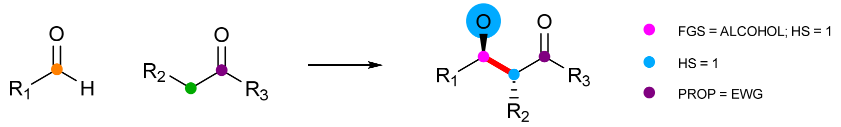
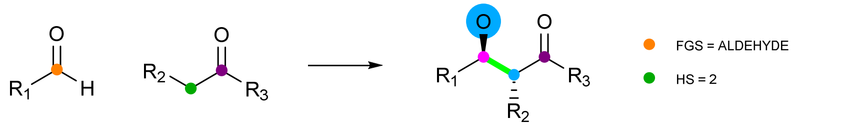
[ccc] For example Ender's RAMP/SAMP chiral auxiliary methodology.[284] This rule branch is not shown.

Rule identifier	Rule Pattern	A	B	C	D	E	F	G	Notes
A	 2 4 O EWG 1 3 2 4 α O EWG β 1 3 FGS = ALDEHYDE, KETONE HS > 0 FG=ALCOHOL HS = 1 PROP = EWG	7251	2056	79	25	93	13	-	Stabilised carbanion addition to aldehydes or ketones
A.1	 O R_1 R_2 R_3 N R_4 R_1 R_2 O N R_3 R_4 FGS = ALDEHYDE, KETONE FGS = ALKYL_NITRO; HS > 0 FGS = ALKYL_NITRO	812	422	83	22	92	11	21	Henry reaction (nitro aldol) (*rule hierarchy not elaborated*)
A.2	 O R_1 R_2 R_3 O R_5 R_4 R_1 R_2 O O R_5 R_3 R_4 FG=ALCOHOL HS = 1 PROP = EWG	4649	1362	77	27	94	12	66	Aldol and closely related reactions
A.2.1	 O R_1 R_2 R_3 O R_5 R_4 R_1 R_2 O O R_5 R_3 R_4 FGS = ALDEHYDE, KETONE HS > 0 FG = ALCOHOL	2546	853	78	26	95	8	63	Aldol – forming α, β stereocentres
A.2.2	 O R_1 R_1 R_2 O R_4 R_3 R_1 R_1 O O R_4 R_2 R_3 HS = 1 PROP = EWG SAME=1	156	40	86	14	89	8	3	Aldol – forming α stereocentre via symmetrical ketones or formaldehyde
A.2.3	 O R_1 R_2 R_3 O R_4 R_3 R_1 R_2 O O R_4 R_3 R_3	828	358	80	26	91	14	26	Aldol – forming β stereocentre

Rating Key: ● Superb ● Excellent ● Very good ● Good ● Fair ● Poor ● Very poor ● No support

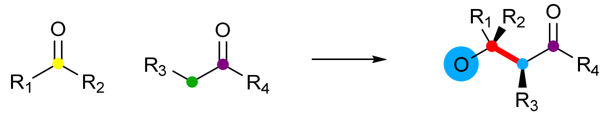
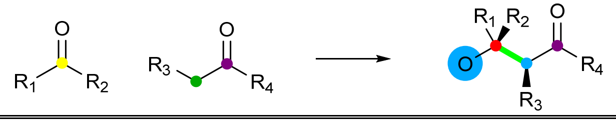
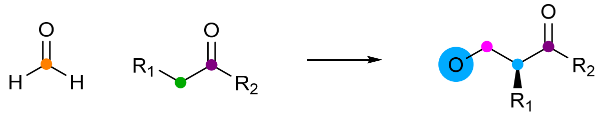
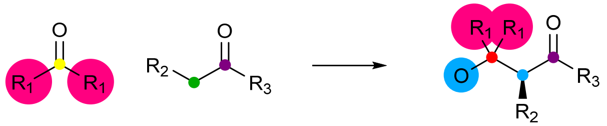
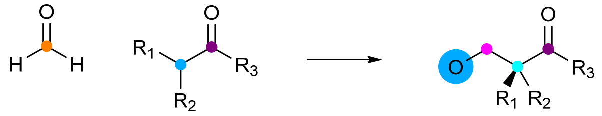
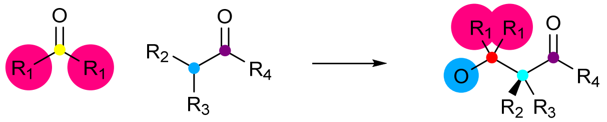
Table 38 **Enantioselective aldol rules (part 1) showing: root rule (A); nitro aldol branch (A.1); aldol branch (A.2); patterns of stereocentre formation in the aldol sub-branches (A.2.1 - α, β, A.2.2 - α, A.2.3 - β).**

Columns:
A – Number of examples
B – Number of examples with *% ee* values
C – Estimated yield (median %)
D – Reliability of the estimated yield (IQR spread in %)
E – Estimated *% ee* value (median %)
F – Reliability of the estimated *% ee* value (IQR spread in %)
G – Percentage of parent rule with respect to B

Rule identifier	Rule Pattern	A	B	C	D	E	F	G	Notes
A.2.1.1		1204	266	77	22	94	7	31	Ethylcarbonyl (and higher) addition to an aldehyde forming the *syn* product
A.2.1.2		1050	525	76	25	96	7	62	Ethylcarbonyl (and higher) addition to an aldehyde forming the *anti*-product
A.2.1.3		164	32	77	35	99	5	4	Ethylcarbonyl (and higher) addition to a ketone
A.2.1.1.1		1188	261	77	22	94	8	98	Intermolecular ethylcarbonyl (and higher) addition to an aldehyde forming the *syn*-product
A.2.1.1.2		16	5	80	17	⬛	⬛	2	Intramolecular ethylcarbonyl (and higher) addition to an aldehyde forming the *syn*-product
A.2.1.2.1		1034	520	76	25	96	7	99	Intermolecular ethyl carbonyl (and higher) addition to an aldehyde forming the *anti*-product
A.2.1.2.2		16	5	75	11	⬛	⬛	1	Intramolecular ethylcarbonyl (and higher) addition to an aldehyde forming the *anti*-product

Rating Key: ● Superb ● Excellent ● Very good ● Good ● Fair ● Poor ● Very poor ● No support

Table 39 **Enantioselective aldol rules (part 2) showing: substitution patterns derived from rule A.2.1. Column key: see Table 38.**

Rule identifier	Rule Pattern		A	B	C	D	E	F	G	Notes
A.2.1.3.1		— RINGS = NO — RINGS = YES ● FGS = KETONE ● FGS = ALCOHOL; HS = 0	93	19	77	20	99	4	59	Intermolecular ethylcarbonyl (and higher) addition to a ketone
A.2.1.3.2		● HS = 1 ● PROP = EWG	71	13	65	29	96	10	41	Intramolecular ethylcarbonyl (and higher) addition to a ketone
A.2.2.1		● FGS = ALDEHYDE ● HS = 2 ● FGS = ALCOHOL; HS = 2	24	10	44	7	97	2	25	Ethylcarbonyl (and higher) addition to formaldehyde
A.2.2.2		● HS = 1 ● PROP = EWG	32	4	89	18	●	●	10	Ethylcarbonyl (and higher) addition to symmetrical ketones
A.2.2.3		● SAME=1	33	25	92	10	87	8	63	α substituted ethylcarbonyl (and higher) addition to formaldehyde
A.2.2.4		● FGS = KETONE ● FGS = ALCOHOL; HS = 0 ● HS = 0	4	0	●	●	●	●	0	α substituted ethylcarbonyl (and higher) addition to symmetrical ketones

Rating Key: ● Superb ● Excellent ● Very good ● Good ● Fair ● Poor ● Very poor ● No support

Table 40 Enantioselective aldol rules (part 3) showing: substitution pattern sub-branch rules descended from the aldol formation of an α stereocentre rule A.2.2.

Columns: A – Number of examples

B – Number of examples with *% ee* values

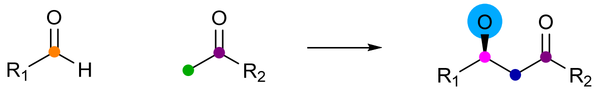
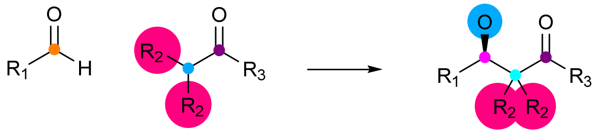
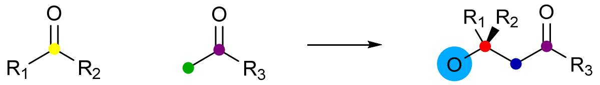
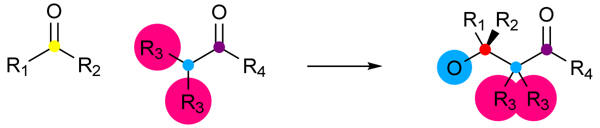
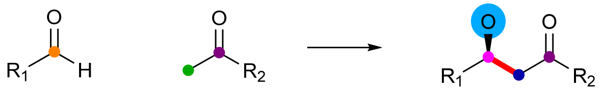
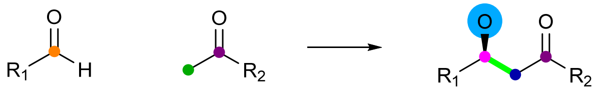
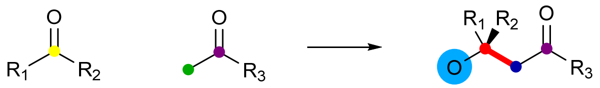
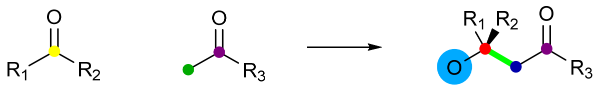
C – Estimated yield (median %)

D – Reliability of the estimated yield (IQR spread in %)

E – Estimated *% ee* value (median %)

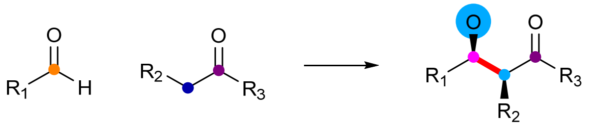
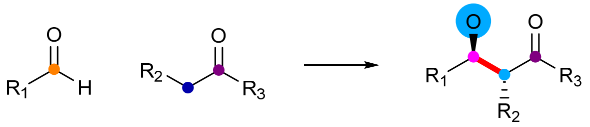
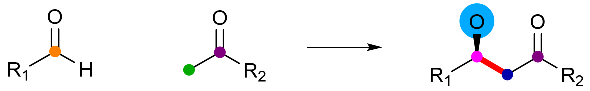
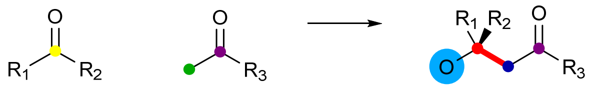
F – Reliability of the estimated *% ee* value (IQR spread in %)

G – Percentage of parent rule with respect to B

Rule identifier	Rule Pattern		A	B	C	D	E	F	G	Notes
A.2.3.1		● FGS = ALDEHYDE ● HS = 3 ● FGS = ALCOHOL; HS = 1	565	241	74	26	91	16	67	Methylcarbonyl addition to an aldehyde
A.2.3.2	 O R$_1$ H R$_2$ O R$_3$ R$_2$ O O R$_1$ R$_3$ R$_2$ R$_2$	● HS = 2 ● HS = 1 ● HS = 0 ● PROP = EWG	18	10	91	14	94	2	3	Symmetrical substituted carbonyl addition to an aldehyde
A.2.3.3	 O R$_1$ R$_2$ O R$_3$ R$_1$ R$_2$ O O R$_3$	● FGS = KETONE ● FGS = ALCOHOL; HS = 0	164	90	87	19	90	12	25	Methylcarbonyl addition to an asymmetric ketone
A.2.3.4	 O R$_1$ R$_2$ R$_3$ O R$_4$ R$_3$ R$_1$ R$_2$ O O R$_4$ R$_3$ R$_3$	● SAME=1	10	4	87	25	■	■	1	Symmetrical substituted carbonyl addition to an asymmetric ketone
A.2.3.1.1		— RINGS = NO — RINGS = YES	564	241	74	26	91	16	100	Intermolecular methylcarbonyl addition to an aldehyde
A.2.3.1.2		● FGS = ALDEHYDE ● HS = 3 ● FGS = ALCOHOL; HS = 1	1	0	■	■	■	■	0	Intramolecular methylcarbonyl addition to an aldehyde
A.2.3.3.1	 O R$_1$ R$_2$ O R$_3$ R$_1$ R$_2$ O O R$_3$	● HS = 2 ● HS = 1 ● HS = 0	131	88	87	19	90	12	98	Intermolecular methylcarbonyl addition to an asymmetric ketone
A.2.3.3.2	 O R$_1$ R$_2$ O R$_3$ R$_1$ R$_2$ O O R$_3$	● PROP = EWG ● FGS = KETONE ● FGS = ALCOHOL; HS = 0	33	2	72	1	■	■	2	Intramolecular methylcarbonyl addition to an asymmetric ketone

Rating Key: ● Superb ● Excellent ● Very good ● Good ● Fair ● Poor ● Very poor ● No support

Table 41 **Enantioselective aldol rules (part 4) showing: aldol substitution pattern sub-branch rules descended from the formation of a β stereocentre rule A.2.3 Column key: see Table 40.**

Rule identifier	Rule Pattern		A	B	C	D	E	F	G	Notes
A.2.1.1.1.9		● PROP = EWG ; FGS = KETONE	606	211	78	26	93	10	81	Ketones
A.2.1.1.1.10		● PROP = EWG ; FGS = ESTER	107	8	73	10	96	6	3	Esters
A.2.1.1.1.11		● PROP = EWG ; FGS = AMIDE	377	19	82	18	98	4	7	Amides
A.2.1.1.1.12		● PROP = EWG ; FGS = ALDEHYDE	24	18	73	10	98	3	7	Aldehydes
A.2.1.1.1.13		● PROP = EWG ; FGS = CARBOXYLIC_ACID	24	5	76	4	■	■	2	Acids
A.2.1.2.1.7		● PROP = EWG ; FGS = KETONE	685	458	76	27	96	7	88	Ketones
A.2.1.2.1.8		● PROP = EWG ; FGS = ESTER	162	16	77	22	91	31	3	Esters
A.2.1.2.1.9		● PROP = EWG ; FGS = AMIDE	75	0	■	■	■	■	0	Amides
A.2.1.2.1.10		● PROP = EWG ; FGS = ALDEHYDE	34	31	80	15	97	4	6	Aldehydes
A.2.1.2.1.11		● PROP = EWG ; FGS = CARBOXYLIC_ACID	9	3	70	14	■	■	1	Acids
A.2.3.1.1.1		● PROP = EWG ; FGS = KETONE	105	87	87	18	90	13	99	Ketones
A.2.3.3.1.1		● PROP = EWG ; FGS = KETONE	405	228	74	27	91	16	95	Ketones
A.2.3.3.1.2		● PROP = EWG ; FGS = ESTER	41	6	77	12	92	4	2	Esters

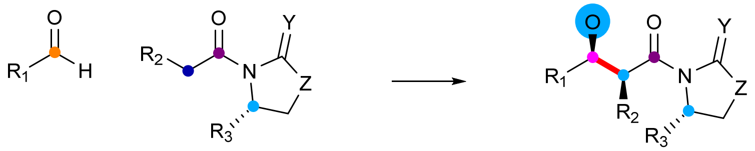
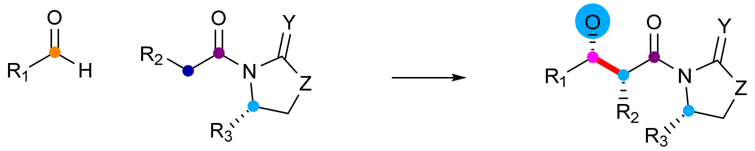
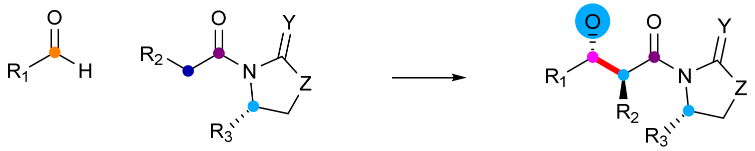
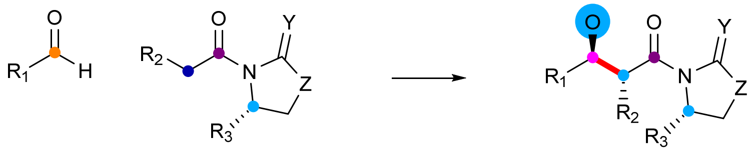
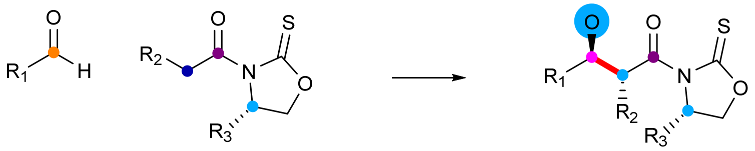
— RINGS = NO ● HS = 3 ● HS = 2 ● HS = 1 ● HS = 0 ● FGS = ALDEHYDE ● FGS = KETONE ● FGS = ALCOHOL; HS = 0 ● FGS = ALCOHOL; HS = 1

Rating Key: ● Superb ● Excellent ● Very good ● Good ● Fair ● Poor ● Very poor ● No support

Table 42 **Enantioselective aldol rules (part 5) showing: daughter rules with specific electron-withdrawing functional groups for the intermolecular *syn*-aldol (A.2.1.1.1.X), intermolecular *anti*-aldol (A.2.1.2.1.X), aldehyde plus methylcarbonyl (A.2.3.1.1.X) and ketone plus methylcarbonyl (A.2.3.3.1.X).**

Columns:

A – Number of examples	**B – Number of examples with *% ee* values**
C – Estimated yield (median %)	**D – Reliability of the estimated yield (IQR spread in %)**
E – Estimated *% ee* value (median %)	**F – Reliability of the estimated *% ee* value (IQR spread in %)**
G – Percentage of parent rule with respect to B	

Rule identifier	Rule Pattern		A	B	Reagents and counts
A.2.1.1.1.1		Y = O ; Z = O	125	50	R_2BOTf / R_3N (90)
A.2.1.1.1.3		Y = S ; Z = O	21	15	$TiCl_4$ / R_3N / (NMP) (13)
A.2.1.1.1.5		Y = S ; Z = S	19	14	$TiCl_4$ / (-)-sparteine (12)
A.2.1.1.1.2		Y = O ; Z = O	12	0	TMSOTf / $TiCl_4$ / R_3N (4) R_2 = CF_3
A.2.1.1.1.4		Y = S ; Z = O	7	5	$TiCl_4$ / (-)-sparteine (5)
A.2.1.1.1.6		Y = S ; Z = S	15	12	$TiCl_4$ / (-)-sparteine (8) or $Sn(OTf)_4$ / R_3N (5)
A.2.1.2.1.1		Y = O ; Z = O	2	1	-
A.2.1.2.1.3		Y = S ; Z = O	0	0	-
A.2.1.2.1.5		Y = S ; Z = S	5	0	TMSCl /R_3N /$MgBr_2.OEt_2$
A.2.1.2.1.2		Y = O ; Z = O	22	5	TMSCl / R_3N/ $MgCl_2$ (14) or R_2BOTf / R_3N inverse addition (6)
A.2.1.2.1.4		Y = S ; Z = O	10	1	See daughter rules A.2.1.2.1.4.x
A.2.1.2.1.6		Y = S ; Z = S	0	0	-
A.2.1.2.1.4.1		R_2 = OR	10	1	*N*-glycolyl examples 2eq $TiCl_4$ / (-)-sparteine
A.2.1.2.1.4.2		R_2 = C	0	0	Non *N*-glycolyl examples

● HS = 2 ● HS = 1 ● FGS = ALDEHYDE ● PROP = EWG ● FGS = ALCOHOL; HS = 1

Rating Key: ● Superb ● Excellent ● Very good ● Good ● Fair ● Poor ● Very poor ● No support

Table 43 Diastereoselective aldol rules showing: rules for selecting oxazolidinone, oxazolidinethione and thiazolidinethione chiral auxiliary examples.
Columns: A – number of examples
B – number of examples with *% de* values

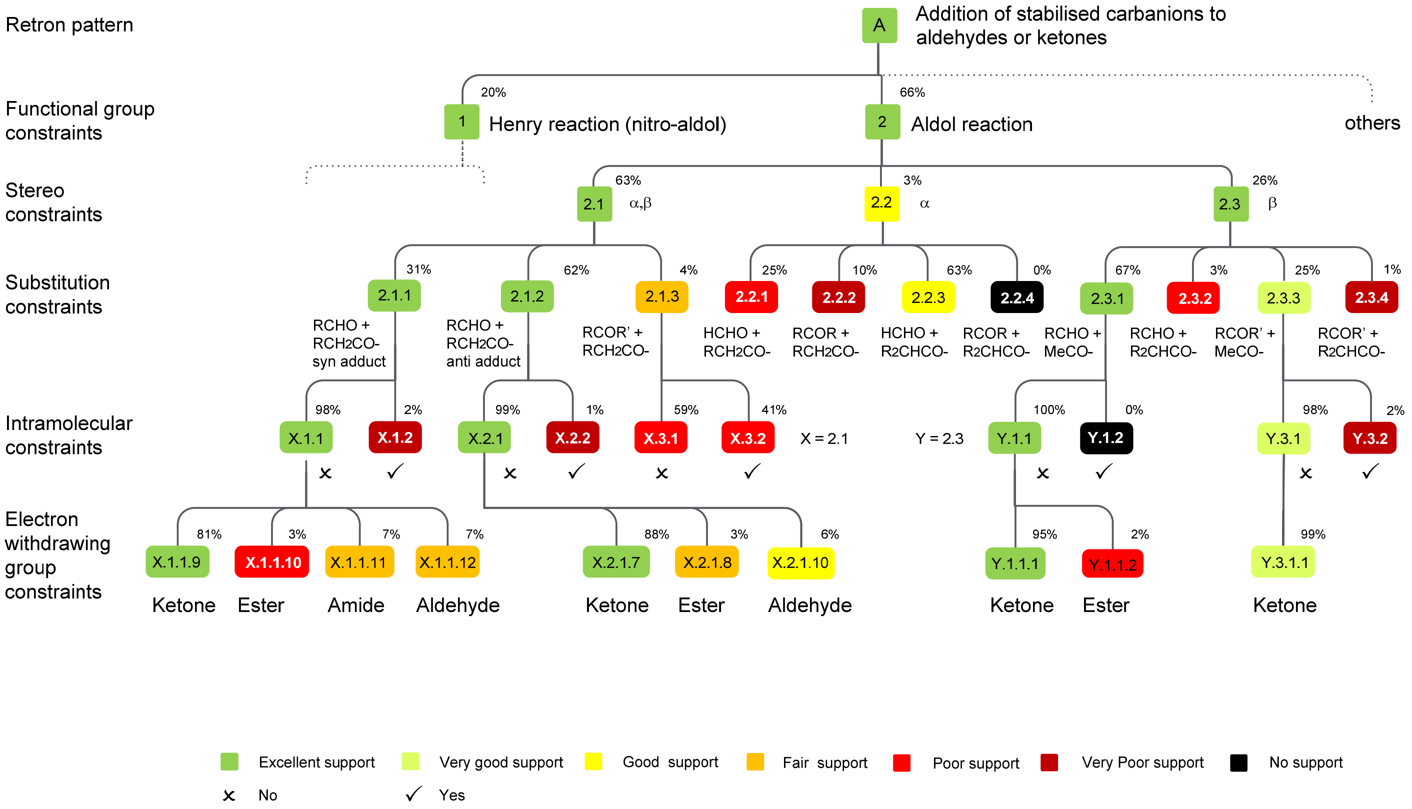
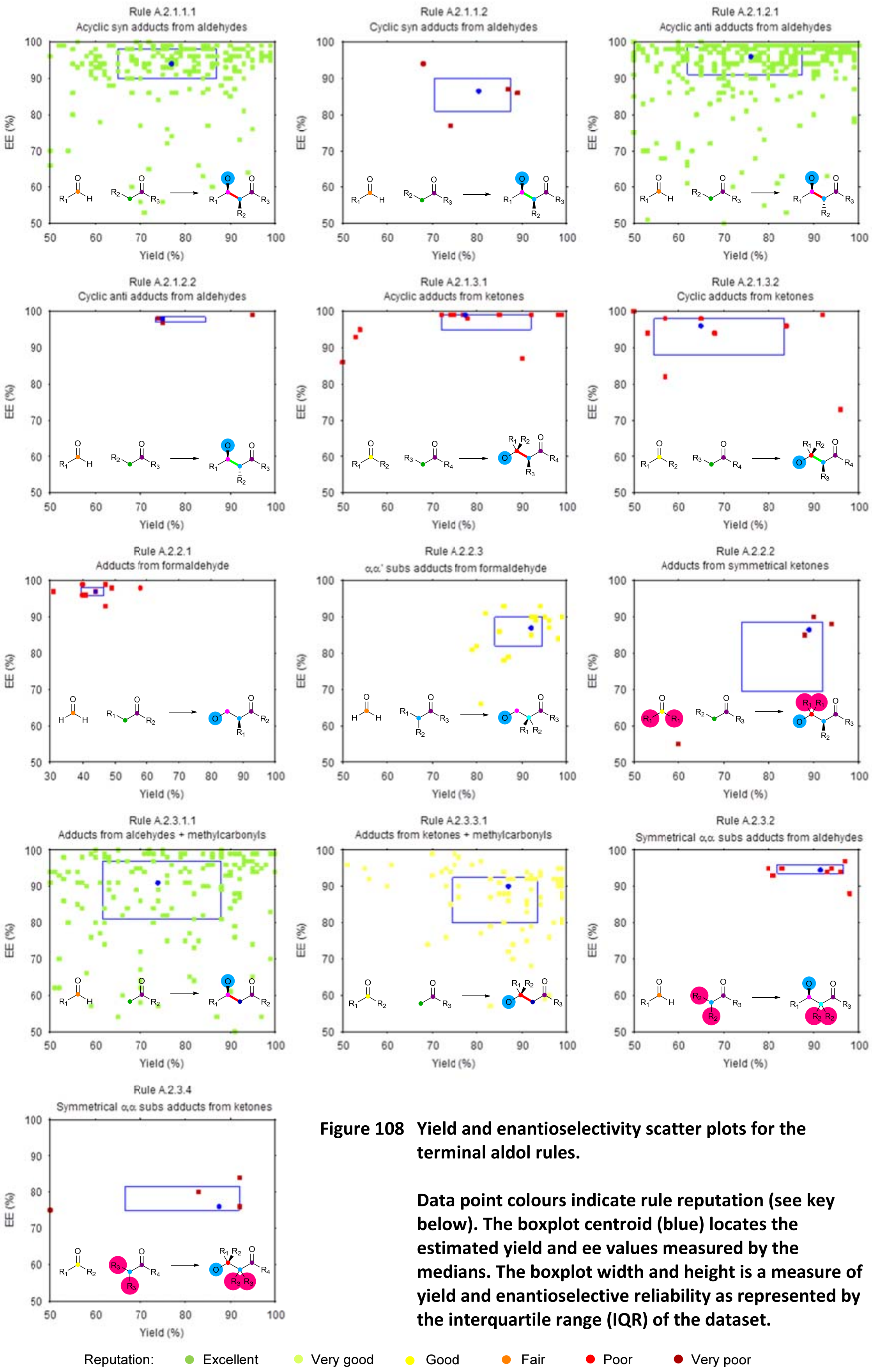


Figure 107 **The enantioselective aldol rule hierarchy.**



Rule A.2.1.1.1 Acyclic syn adducts from aldehydes

Rule A.2.1.1.2 Cyclic syn adducts from aldehydes

Rule A.2.1.2.1 Acyclic anti adducts from aldehydes

Rule A.2.1.2.2 Cyclic anti adducts from aldehydes

Rule A.2.1.3.1 Acyclic adducts from ketones

Rule A.2.1.3.2 Cyclic adducts from ketones

Rule A.2.2.1 Adducts from formaldehyde

Rule A.2.2.3 α,α' subs adducts from formaldehyde

Rule A.2.2.2 Adducts from symmetrical ketones

Rule A.2.3.1.1 Adducts from aldehydes + methylcarbonyls

Rule A.2.3.3.1 Adducts from ketones + methylcarbonyls

Rule A.2.3.2 Symmetrical α,α subs adducts from aldehydes

Rule A.2.3.4 Symmetrical α,α subs adducts from ketones

EE (%)
Yield (%)
50
60
70
80
90
100
30
40
R1
R2
R3
R4
H
O
Reputation: ● Excellent ● Very good ● Good ● Fair ● Poor ● Very poor

Figure 108 Yield and enantioselectivity scatter plots for the terminal aldol rules.

Data point colours indicate rule reputation (see key below). The boxplot centroid (blue) locates the estimated yield and ee values measured by the medians. The boxplot width and height is a measure of yield and enantioselective reliability as represented by the interquartile range (IQR) of the dataset.

Results and Discussion

The counts of total reaction examples and examples with good *% ee* values are shown for each rule in Table 38 through to Table 43 (*vide supra*). The reaction reputation, likely reaction yields and likely enantiomeric excess values with reliability estimates are colour coded in the previously described manner. Figure 107 summarises the near complete aldol rule hierarchy and it is colour coded to indicate the enantioselective reputation for each rule. The Evans auxiliary rules were omitted from Figure 107 for space reasons but are presented in full in Table 43.

The stabilised carbanion addition to aldehydes or ketones (rule A) is very well represented in the source databases with an aggregate total of 7251 examples of which 2056 have good *% ee* values. The aldol subset (rule A.2) consisting of carbonyl stabilised carbanions (reacting as enolates) is represented by 4649 examples of which 1362 have good *% ee* values.

Within the aldol branch the rules with the best reputation, stereoselectivity and reliability are the *intermolecular* addition of:

- Ethyl or higher ketones to aldehydes to form *anti*-adducts (rule A.2.1.2.1.7 – 458 examples, *% ee* - median: 96%, *% ee* IQR: 91 – 98%).
- Ethyl or higher ketones to aldehydes to form *syn*-adducts (rule A.2.1.1.1.9 – 211 examples, *% ee* - median 93%, *% ee* IQR: 86 – 96%).
- Methyl ketones to aldehydes (rule A.2.3.1.1.1 – 228 examples, *% ee* - median: 91%, *% ee* IQR: 81 – 97%).
- Methyl ketones to prochiral ketones (rule A.2.3.3.1.1 – 87 examples, *% ee* - median: 90%, *% ee* IQR: 80 – 93%).

The corresponding rules for enantioselective *intramolecular* aldol have very poor reputations in contrast to the excellent reputations for the *intermolecular* variant. The low intramolecular example count is likely due to a number of factors such as: poor stereoselective control due to a transition state distorted by tethered substituents; the inability to pre-form the enolate in the absence of the electrophile; or a tendency for the initial β-hydroxy carbonyl product to dehydrate to form a cyclic enone product. This lack of intramolecular reputation should provide good supporting evidence to strongly disfavour offering these intramolecular disconnections during a retrosynthetic analysis.

The effects of varying the enolate substitution patterns are presented in selected aldol example reaction yield/enantioselectivity scatter plots (Figure 108). The lack of terminal substituents on the enolates derived from methyl carbonyls (rules A.2.3.1.1 and A.2.3.3.1) significantly lowers the enantioselective reliability of the reaction compared to those using *Z*- or *E*-enolates (rules A.2.1.1.1 and A.2.1.2.1) indicating a less organised transition state is in control. The observed interquartile range of *% ee* values for *Z*- or *E*-enolates represents a corresponding Gibbs free energy difference ($\Delta\Delta G^{\ddagger}$) of 1.8 to 2.6 kcal mol^{-1} (at 25 ^{o}C) between the two competing transitions states. The interquartile interval of $\Delta\Delta G^{\ddagger}$ values for methyl carbonyl derived enolates is observed to have lowered and broadened to 1.4 – 2.5 kcal mol^{-1} when reacting with aldehyde electrophiles and 1.3 – 1.9 kcal mol^{-1} when reacting with ketone electrophiles.

The application of chiral controllers was investigated using rules A.2.1.1.1.1 – 6 and A.2.1.2.1.1 – 6 which were designed to extract examples for reactions using oxazolidinone, oxazolidinethione and thiazolidinethione auxiliaries. The selected examples generate diastereomers and the outcomes are reported in the database as *% de* values, though in general these are absent in over half of all cases. The original method developed by Evans for the generation of *syn*-adducts[275] is well represented with 125 examples (rule A.2.1.1.1.1) and the method is used by a highly diverse set of authors indicating good acceptance of the method. In contrast the newer methods for generating non-Evans *syn*-adducts (due to Heathcock [276] and Crimmins [277]) and the two *anti*-adducts (due to Evans [268, 285]) are poorly represented and dominated by examples from the original papers, indicating that these methods did not receive the same level of utility. Indeed, the number of reaction examples exploiting the oxazolidinone class of auxiliaries (161 examples) has been overtaken by examples of the more efficient and easier to handle direct aldol reaction (2056 examples) and the Mukiayama aldol reaction (190 examples).

The Enantioselective Stetter Reaction

Overview

The Stetter reaction is the formal addition of an aldehyde to a Michael acceptor to generate 1, 4 keto-functionalised adducts with the creation of a new carbon-carbon bond and the introduction of one or two new chiral centres that can be created under stereo control. The β-chiral centre (using the Michael acceptor nomenclature) is directly connected to the new bond and can be constructed under enantioselective control while the more remote α-chiral centre requires stereoselective control of the re-protonation step.

To activate the reaction the aldehyde has to be converted *in-situ* to an acyl anion equivalent to act as the nucleophile and as such is an example of reverse polarity *(umpolung*) chemistry. The catalytic conversion of the aldehyde to a chiral nucleophilic species is typically accomplished using a chiral *N*-heterocyclic carbene (NHC) intermediate generated from thiazolium, imidazolium or triazolium salts using an appropriate base.[286] The proposed catalytic cycle is shown in Figure 109. The aldehyde A reacts with the NHC catalyst B to form the Breslow intermediate [287] C which acts as the acyl anion equivalent. This species attacks the Michael acceptor D followed by expulsion of the NHC catalyst (B) after proton transfer generates the chiral 1,4 keto-functionalised adduct E. The expulsion of the carbene restarts the catalytic cycle.

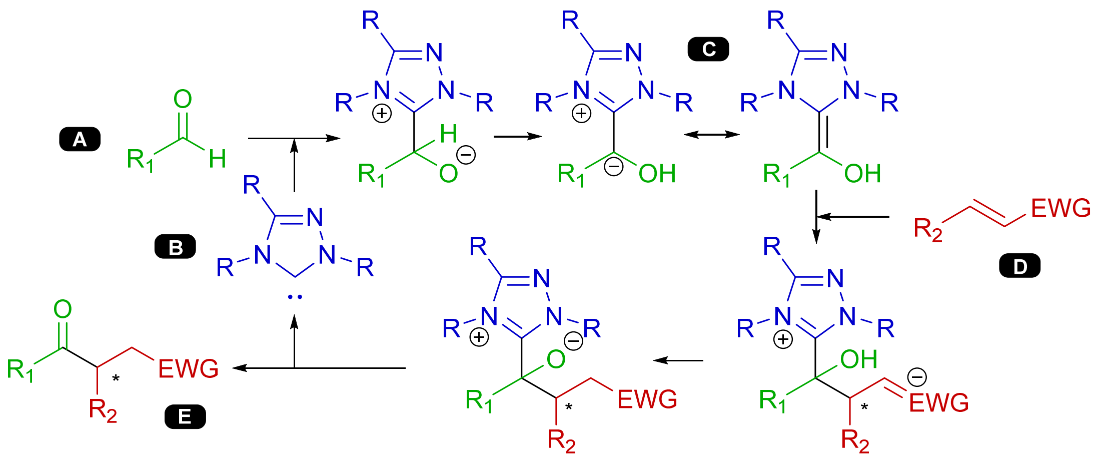


Figure 109 The mechanism of the *N*-heterocyclic carbene catalysed Stetter reaction.

A large range of chiral NHC catalysts have been explored for use the Stetter reaction. Successful catalysts have been found using chiral bicyclic triazolidene carbenes [288] and, to a lesser extent, chiral C_2-symmetric imidazolidene carbenes (Figure 110).[289] It is currently the case that the catalyst must be matched to the chosen pair of Michael acceptor and aldehyde to achieve an

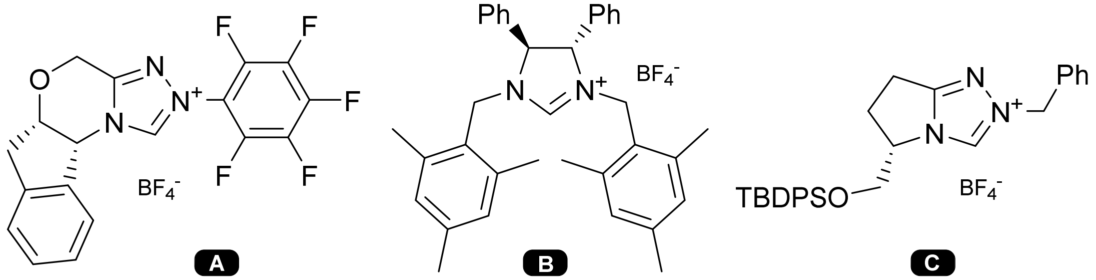


Figure 110 Chiral pre-catalyst triazolium salts and C_2-symmetric imidazolium salts used in the enantioselective Stetter reaction.

appropriate degree of reactivity and selectivity. For example the Matsumoto catalyst (Figure 110; B) is currently the optimum choice for the intramolecular Stetter reaction of α,β unsaturated esters or ketones reacting with aliphatic aldehydes. [290]

Greatest success has been achieved with the *intramolecular* Stetter reaction in terms of obtaining high enantioselectivity and Michael acceptor diversity. The high selectivity is due to an organised transition state with the tethered Michael acceptor adopting a chair-like configuration with respect to the chiral Breslow intermediate (Figure 111). [291]

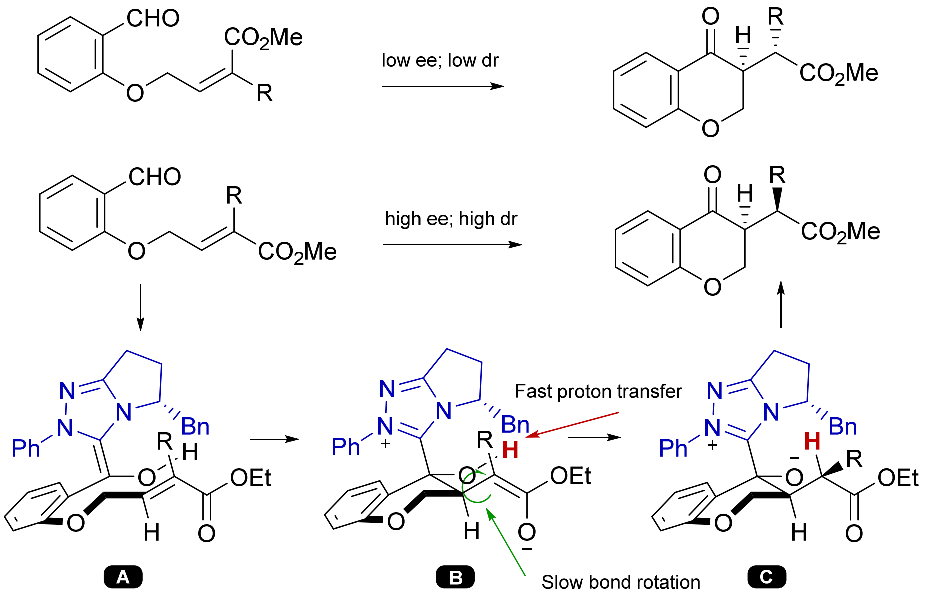


Figure 111 The mechanism of enantioselective and diastereoselective control in the intramolecular Stetter reaction of salicylaldehyde derivatives.

Rovis has shown that the diastereoselectivity of the α chiral centre formed from reacting α substituted Michael acceptors is consistently high in intramolecular Stetter reactions and is due to an internal proton transfer to the intermediate enolate (Figure 111: B) which is rapid enough to occur before the enolate bond can rotate to present the opposing face.[291] Intramolecular *cis*-Michael acceptors are less reactive and selective compared to the *trans*-analogues but due to the rapid proton transfer to form α stereogenic centres the diastereomeric outcome is predictable with respect to the substituent geometry of the tethered Michael acceptor. Although 5- and 6-membered intramolecular cyclisations are readily achieved, it was apparently not possible to create 7-membered rings.[292]

In contrast, during the period covered by the reaction database (up to 2010), the *intermolecular* asymmetric Stetter reaction had met with limited success largely due to a then general lack of enantioselective control in an open transition state.[289] Where success had been achieved it was with specific classes of substrate pairings allied with highly tuned catalysts able to achieve a more controlled (pre-) transition state. An example is the reaction of aryl aldehydes with nitroalkenes using a catalyst with a fluorinated backbone (Figure 112). Computational studies suggest the partial positive charge induced on the catalyst carbon atom by the fluorine substituent is responsible for increasing enantioselectivity of the reaction by locking the conformation of the nitroalkene in the pre-transition state assembly. [293]

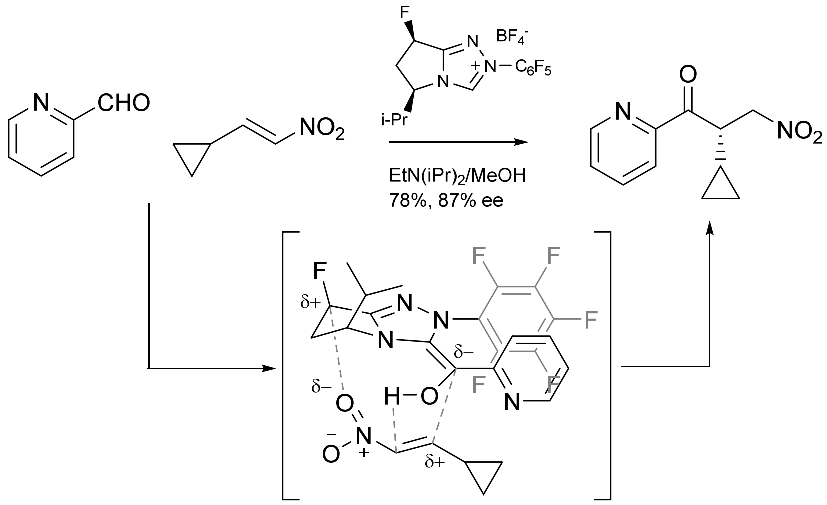


Figure 112 A proposed pre-transition state assembly used to explain increased enantioselectivities in the intermolecular Stetter reaction of aryl aldehydes with nitroalkenes when using a fluorinated NHC catalyst.

Rule Development

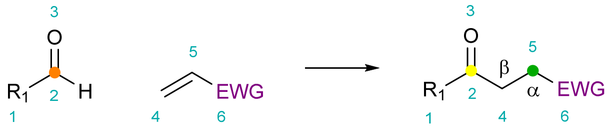
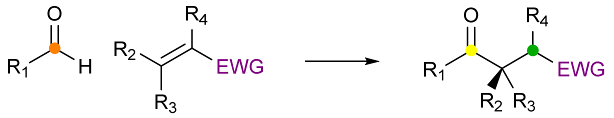
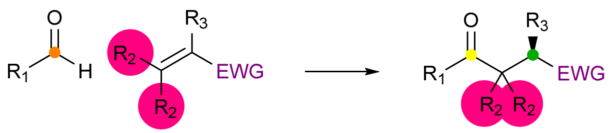
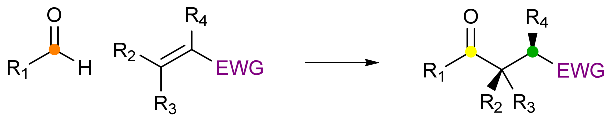
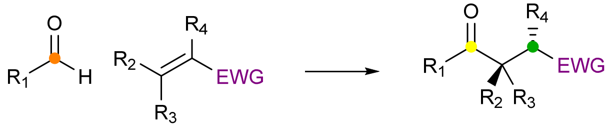
The Stetter rules are listed in Table 44 through Table 48 (*vide infra*) with each showing the rule identifier, rule pattern and the results of the example reaction searches in term of the number of examples found and the number of examples with quoted *% ee* values, and estimated yield and *% ee* values. Figure 113 (*vide infra*) shows the same rules assembled into a hierarchical tree organised by increasing retron detail.

The initial levels of the rule hierarchy follow the established protocol of describing the minimum retron requirements with the immediate daughter rules selecting all possible stereogenic patterns and then the granddaughter rules enumerating all substitution patterns. Subsequent levels of the rule hierarchy establish constraints tailored to the reported and observed trends seen in the parent rule examples.

The root rule (Table 44: rule S) sets out the minimum functional and structural requirements of the Stetter retron describing the necessary 1,4 relationship between a ketone and an electron-withdrawing functional group. The daughter rules S.1 to S.4 enumerate all possible stereocentre patterns in the retron. In particular rules S.3 and S.4 explore the diastereoselective outcomes of the re-protonation step for targets with adjacent stereogenic centres lying between the ketone and the electron-withdrawing group.

The fourth level of the rule hierarchy divides the examples into intermolecular and intramolecular variants by applying the RINGS=NO and RINGS=YES constraint to the formed C-C bond. The intramolecular branches are further divided into ring size divisions. The chosen sizes are <5, 5, 6 and >6 to mirror the reported observations that only the 5 or 6 membered cyclisations are favoured. The <5 and >6 rules are designed to confirm the absence of support for small or medium/large rings and thus enable the ARChem retrosynthesiser to judge that the Stetter disconnection may be inappropriate in these situations.

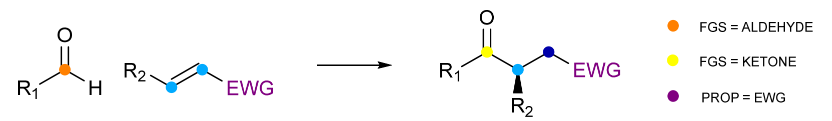
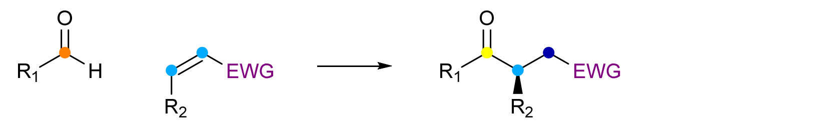
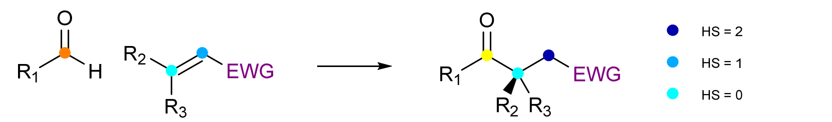
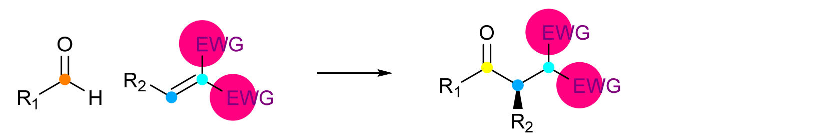
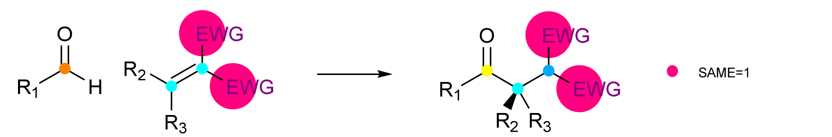
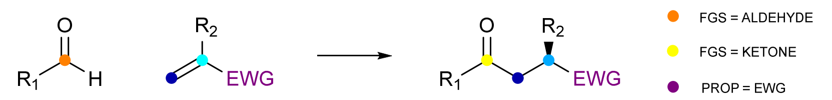
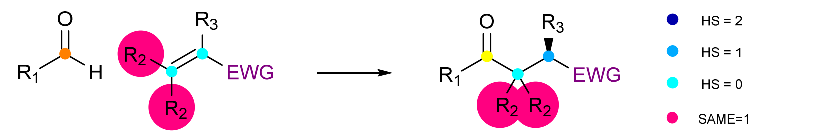
The final two levels of the rule hierarchy are designed to explore enantioselective trends for the limited types of aldehyde and Michael acceptors reported in the literature. Daughter rules are divided into aliphatic and aromatic aldehydes applying the ARYL=NO and ARYL=YES constraints to the aldehyde α carbon atom. Manual inspection of the example reactions for each rule at this depth in the tree revealed that very limited types of Michael acceptor had been reported. This limitation in scope is handled by adding rules that constrain the EWG atom to specific functional groups using the FGS constraint (*e.g.* FGS=ESTER).

Rule identifier	Rule Pattern		A	B	C	D	E	F	G	Notes
S	 3 O R_1 2 H 1 5 EWG 4 6 3 O β 5 R_1 2 α EWG 1 4 6	● FGS = ALDEHYDE ● FGS = KETONE ● HS > 0 ● PROP = EWG	412	164	84	18	92	12	-	Generalised Stetter reaction
S.1	 O R_1 H R_4 R_2 EWG R_3 O R_4 R_1 EWG R_2 R_3		122	110	85	21	90	13	67	Stetter – forming a single β stereocentre
S.2	 O R_1 H R_3 R_2 EWG R_2 O R_3 R_1 EWG R_2 R_2	● FGS = ALDEHYDE ● FGS = KETONE	28	20	81	13	97	3	12	Stetter – forming a single α stereocentre
S.3	 O R_1 H R_4 R_2 EWG R_3 O R_4 R_1 EWG R_2 R_3	● HS > 0 ● PROP = EWG	24	24	80	19	92	6	15	Stetter – forming syn α,β stereocentres. Protonation is on the same face as aldehyde addition
S.4	 O R_1 H R_4 R_2 EWG R_3 O R_4 R_1 EWG R_2 R_3	● SAME=1	1	0					0	Stetter – forming anti α,β stereocentres. Protonation is on the opposing face to aldehyde addition

Rating Key: ● Superb ● Excellent ● Very good ● Good ● Fair ● Poor ● Very poor ● No support

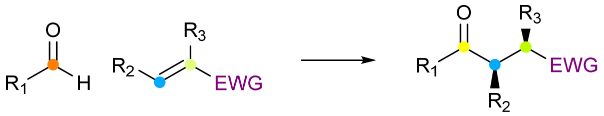
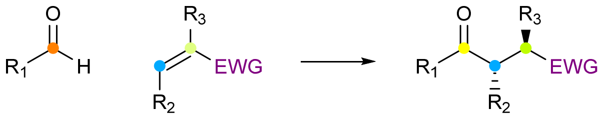
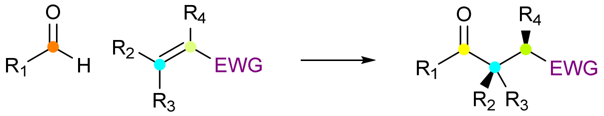
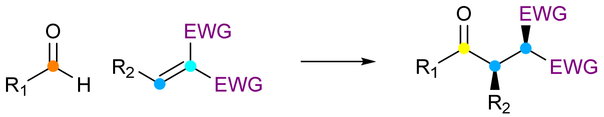
Table 44 **Enantioselective Stetter reaction rules (part 1) showing: root rule (S); patterns of stereocentre formation in the sub-branches S1.1 (β), S.2 (α), S.3 (*trans* α, β), S.4 (*cis* α, β). The atom-atom maps for the daughter rules are omitted for clarity.**

Columns:
A – Number of examples
B – Number of examples with *% ee* values
C – Estimated yield (median %)
D – Reliability of the estimated yield (spread in %)
E – Estimated *% ee* value (median %)
F – Reliability of the estimated *% ee* value (spread in %)
G – Percentage of parent rule with respect to B

Rule identifier	Rule Pattern	A	B	C	D	E	F	G	Notes
S.1.1	 O R1 H R2 EWG O R1 EWG R2 ● FGS = ALDEHYDE ● FGS = KETONE ● PROP = EWG	79	70	85	19	91	12	64	Addition of aldehyde to a trans Michael acceptor
S.1.2.		11	8	86	31	88	20	7	Addition of aldehyde to a cis Michael acceptor
S.1.3	 O R1 H R2 EWG R3 O R1 EWG R2 R3 ● HS = 2 ● HS = 1 ● HS = 0	18	18	85	29	96	8	16	Addition of aldehyde to a tri-substituted Michael acceptor
S.1.4	 O R1 H R2 EWG EWG O R1 EWG EWG R2	13	13	88	7	82	10	12	Addition of aldehyde to a tri-substituted Michael acceptor with identical EWGs
S.1.5	 O R1 H R2 EWG EWG R3 O R1 EWG EWG R2 R3 ● SAME=1	0	0					0	Addition of aldehyde to a tetra-substituted Michael acceptor with identical EWGs
S.2.1	 O R1 H R2 EWG O R2 R1 EWG ● FGS = ALDEHYDE ● FGS = KETONE ● PROP = EWG	28	20	81	13	97	3	100	Addition of aldehyde to a α subtituted Michael acceptor
S.2.2	 O R1 H R2 R3 EWG R2 O R3 R1 EWG R2 R2 ● HS = 2 ● HS = 1 ● HS = 0 ● SAME=1	0	0					0	Addition of aldehyde to a β,β tetra-substituted Michael acceptor

Rating Key: ● Superb ● Excellent ● Very good ● Good ● Fair ● Poor ● Very poor ● No support

Table 45 **Enantioselective Stetter reaction rules (part 2) showing: patterns of substitution in the β and α stereoselective Stetter sub-branches. Column key: see Table 44.**

Rule identifier	Rule Pattern	A	B	C	D	E	F	G	Notes
S.3.1	 O R1 H R3 R2 EWG O R3 R1 EWG R2	8	8	87	15	92	8	33	Addition of aldehyde to a Michael acceptor generating a *trans* adduct
S.3.2	 O R1 H R3 EWG R2 O R3 R1 EWG R2 ● HS = 2 ● HS = 1 ● HS = 0	4	4					17	Addition of aldehyde to a Michael acceptor generating a *cis* adduct
S.3.3	 O R1 H R4 R2 EWG R3 O R4 R1 EWG R2 R3 ● HS = 2 ; EWGS = 1 ● HS = 1 ; EWGS = 1 ● HS = 0 ; EWGS = 1	0	0					0	Addition of aldehyde to a tetra-substituted Michael acceptor
S.3.4	 O R1 H EWG R2 EWG O EWG R1 EWG R2	12	12	81	14	90	6	50	Addition of aldehyde to a Michael acceptor with two asymmetric EWGs

Rating Key: ● Superb ● Excellent ● Very good ● Good ● Fair ● Poor ● Very poor ● No support

Table 46 **Enantioselective Stetter reaction rules (part 3) showing: aldehyde addition to tri- and tetra-substituted Michael acceptors with protonation on same face as aldehyde addition.**

Columns:

A – Number of examples

B – Number of examples with *% ee* values

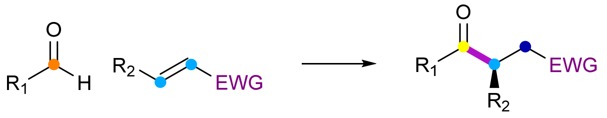
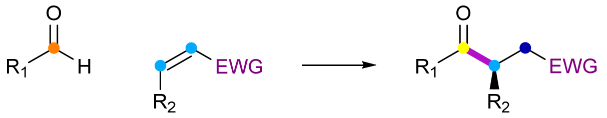
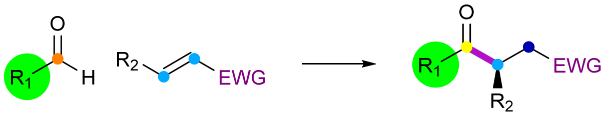
C – Estimated yield (median %)

D – Reliability of the estimated yield (IQR spread in %)

E – Estimated *% ee* value (median %)

F – Reliability of the estimated *% ee* value (IQR spread in %)

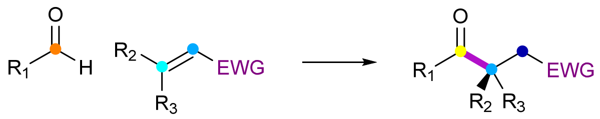
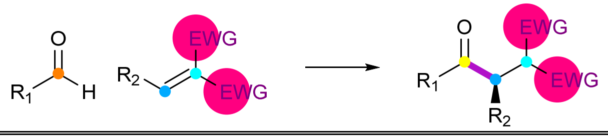
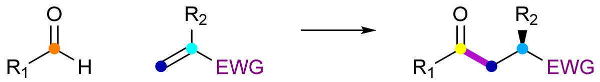
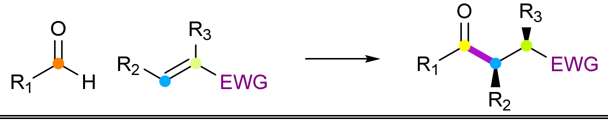
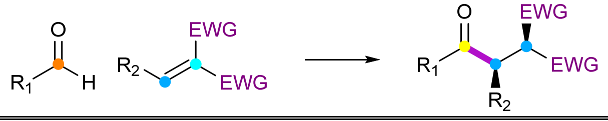
G – Percentage of parent rule with respect to B

Rule identifier	Rule Pattern		A	B	C	D	E	F	G	Notes
S.1.1.1		— RINGS = NO	16	16	83	20	87	12	23	Intermolecular trans Michael acceptor
S.1.1.2		— RINGS = YES	63	54	86	20	92	11	77	Intramolecular trans Michael acceptor
S.1.1.2.1		— RINGS < 5	0	0					0	Formation of 3,4 membered rings
S.1.1.2.2		— RINGS = 5	13	12	80	14	90	15	22	Formation of 5 membered rings
S.1.1.2.3		— RINGS = 6	49	42	87	21	92	12	78	Formation of 6 membered rings
S.1.1.2.4		— RINGS > 6	1	0					0	Formation of 6+ membered rings
S.1.2.1		— RINGS = NO	2	0					0	Intermolecular cis Michael acceptor
S.1.2.2		— RINGS = YES	9	8	86	31	88	20	100	Intramolecular cis Michael acceptor
S.1.2.2.1		— RINGS < 5	0	0					0	Formation of 3,4 membered rings
S.1.2.2.2		— RINGS = 5	8	8	86	31	88	20	100	Formation of 5 membered rings
S.1.2.2.3		— RINGS = 6	1	0					0	Formation of 6 membered rings
S.1.2.2.4		— RINGS > 6	0	0					0	Formation of 6+ membered rings
S.1.1.1.1		— RINGS = NO ● ARYL = YES	12	12	85	21	90	9	75	Aryl CHO forming acyclic adducts
S.1.1.1.2		— RINGS = NO ● ARYL = NO	4	4					25	Alkyl CHO forming acyclic adducts
S.1.1.2.2.1		— RINGS = 5 ● ARYL = YES	3	3					25	Aryl CHO forming 5 membered rings
S.1.1.2.2.2		— RINGS = 5 ● ARYL = NO	10	9	80	10	90	11	75	Alkyl CHO forming 5 membered rings
S.1.1.2.3.1		— RINGS = 6 ● ARYL = YES	45	42	87	21	92	12	100	Aryl CHO forming 6 membered rings
S.1.1.2.3.2		— RINGS = 6 ● ARYL = NO	4	0					0	Alkyl CHO forming 6 membered rings
S.1.1.2.3.1.1		— RINGS = 6 ● ARYL = YES EWG = ESTER	36	34	88	22	93	12	81	Aryl CHO forming 6 ring adducts with trans α,β unsaturated esters
S.1.1.2.3.1.2		— RINGS = 6 ● ARYL = YES EWG = PHOSPHINE_OXIDE	5	5					12	Aryl CHO forming 6 ring adducts with trans α,β unsaturated phosphine oxides
S.1.1.1.1.1		— RINGS = NO ● ARYL = YES EWG = NITRO	11	11	85	21	90	9	92	Aryl CHO forming acyclic adducts with trans nitroalkenes

● FGS = ALDEHYDE ● FGS = KETONE ● PROP = EWG ● HS = 2 ● HS = 1

Rating Key: ● Superb ● Excellent ● Very good ● Good ● Fair ● Poor ● Very poor ● No support

Table 47 **Enantioselective Stetter reaction rules (part 4) showing: addition of aldehydes to *trans*- and *cis*-substituted Michael acceptors forming acyclic and cyclic adducts; addition of aromatic and alkyl aldehydes to *trans*-substituted Michael acceptors forming 5 and 6 ring adducts. Column key: see Table 46.**

Rule identifier	Rule Pattern		A	B	C	D	E	F	G	Notes
S.1.3.1	 O R1 H R2 R3 EWG O R1 R2 R3 EWG	RINGS = NO	0	0					0	Intermolecular addition of aldehyde to a tri-substituted Michael acceptor
S.1.3.2		RINGS = YES	18	18	85	29	96	8	100	Intramolecular addition of aldehyde to a tri-substituted Michael acceptor
S.1.3.2.1		RINGS < 5	0	0					0	Formation of 3,4 membered rings
S.1.3.2.2		RINGS = 5	16	16	85	28	96	9	89	Formation of 5 membered rings
S.1.3.2.3		RINGS = 6	2	2					11	Formation of 6 membered rings
S.1.3.2.4		RINGS > 6	0	0					0	Formation of 7 + membered rings
S.1.4.1	 O R1 H R2 EWG EWG O R1 R2 EWG EWG	RINGS = NO	12	12	86	7	82	11	92	Intermolecular Stetter reaction
S.1.4.2		RINGS = YES	1	1					8	Intramolecular
S.2.1.1		RINGS = YES	5	5					25	Intramolecular Rh cat. hydroacylation
S.2.1.2		RINGS = NO	23	15	78	9	98	1	75	Intermolecular
S.2.1.2.1		EWG = ESTER	9	9	80	26	97	2	60	Intermolecular Stetter reaction
S.2.1.2.2		EWG = AMIDE	7	6	76	7	98	0	40	Intermolecular Rh cat. hydroacylation
S.3.1.1	 O R1 H R3 R2 EWG O R3 R1 EWG R2	RINGS = YES	8	8	87	15	92	8	100	Intramolecular Stetter
S.3.1.2		RINGS = NO	0	0					0	Intermolecular
S.3.4.1	 O R1 H EWG R2 EWG O EWG R1 EWG R2	RINGS = YES	0	0					0	Intramolecular
S.3.4.2		RINGS = NO	12	12	81	14	90	6	100	Intermolecular Stetter

● FGS = ALDEHYDE ● FGS = KETONE ● PROP = EWG ● HS = 2 ● HS = 1 ● HS = 0 ● HS = 1 ; EWGS = 1 ● SAME=1

Key: ● Superb ● Excellent ● Very good ● Good ● Fair ● Poor ● Very poor ● No support

Table 48 **Enantioselective Stetter reaction rules (part 5) showing: Addition of aldehydes to di- and tri-substituted Michael acceptors forming acyclic and cyclic adducts. Column key: see Table 46.**

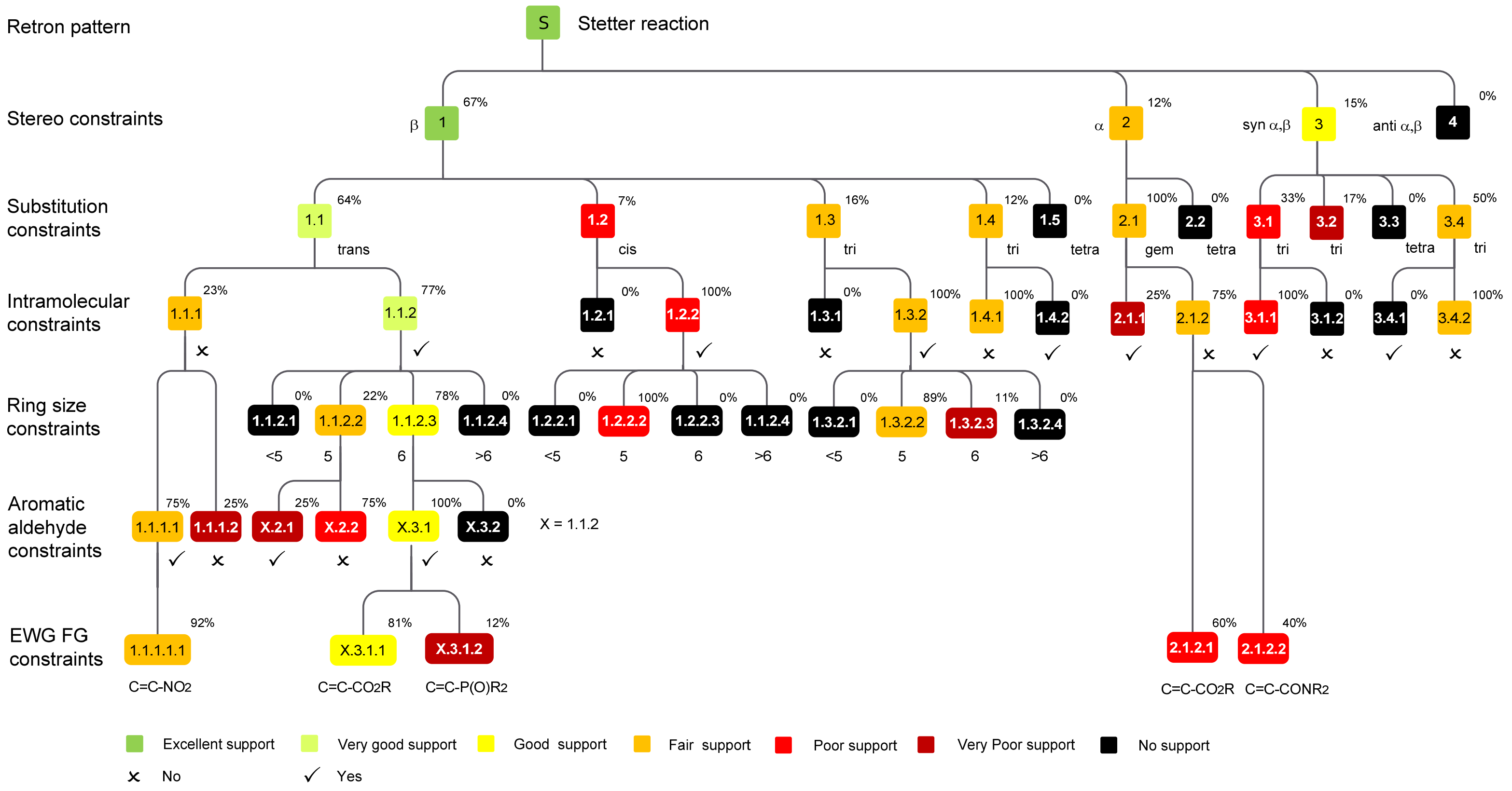


Figure 113 **The rule hierarchy for the enantioselective Stetter reaction.**

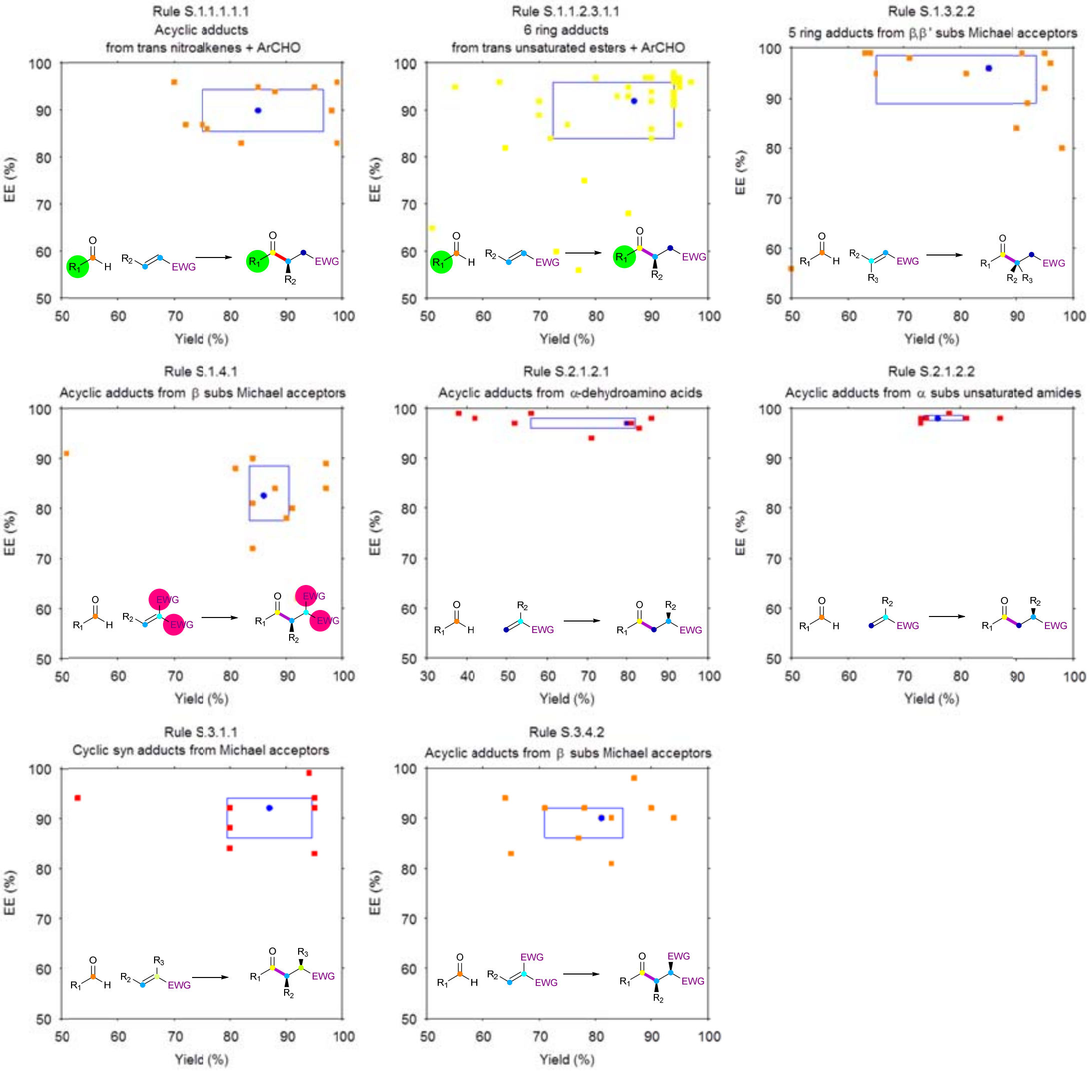


Figure 114 Yield and enantioselectivity scatter plots for the highest reputation Stetter rules.

Results and Discussion

The counts of total reaction examples and examples with good *% ee* values are shown for each rule in Table 44 through to Table 48. Figure 113 summarises the complete Stetter rule hierarchy and it is colour coded to indicate the enantioselective reputation for each rule.

The addition of aldehydes to electron-deficient-alkenes (rule A) is represented by an aggregate total of 412 examples of which only 164 have useful *% ee* values. The limited scope of the reaction is evident from both the low example reputations of many terminal rules and also the large number of unsupported rules. The results confirm the observation that intramolecular Stetter reactions are confined to the formation of 5 or 6 membered rings.[292] The best supported rule is the intramolecular reaction of aryl aldehydes with substituents containing *trans*-α,β unsaturated esters to form 6 membered rings (rule S.1.1.2.3.1.1 – 34 examples, *% ee* - median: 93%, IQR: 84 – 96%). Other notable highlights, albeit with only modest example reputations, include the enantioselective formation of:

- Quaternary chiral centres (rule S.1.3.2.2 – 16 examples, *% ee* - median: 96%, IQR: 88 – 98%).
- Acyclic adducts from aryl aldehydes and nitroalkenes (rule S.1.1.1.1.1 – 11 examples, *% ee* - median: 90%, IQR: 86 – 95%).
- α-Amino acid derivatives from α-dehydroamino acids (rule S.2.1.2.1 – 9 examples, *% ee* - median: 97%, IQR: 96 – 97%).
- Acyclic adducts from 1,1 disubstituted α,β unsaturated amides (rule S.2.1.2.2 – 6 examples, *% ee* - median: 98%, IQR: 97 – 98%).

The yield and enantioselectivity plots (Figure 114) show that reactions of aldehydes with 1,1 disubstituted alkenes appear to have excellent enantioselectivity with superb reliability when compared to rules with superior reputations. However caution must be applied to this observation due to the low example reputation.

The rule hierarchy is not exclusive to the Stetter reaction as rhodium catalysed hydroacylation reaction examples appear as the only representatives for rules S.2.1.1 and S.2.1.2.2. These examples are confined to 1,1 disubstituted electron-deficient alkenes but in general the method has a broader alkene scope as it is also effective with unfunctionalised alkenes.[294, 295] The hydroacylation of unfunctionalised alkenes can be accommodated in the transform hierarchy by introducing a new root rule containing no functional group constraints and this rule is accompanied by the introduction of a new daughter rule that specifies that all alkene substituent are not electron-withdrawing (*via* the EWGS = 0 constraint applied to both alkene atoms) to establish a separate branch. This approach is identical to that already demonstrated with the alkene epoxidation transform.

Substrate Controlled Stereoselective Transforms

Any reaction which creates a new stereogenic center may proceed in a diastereoselective fashion if a pre-existing stereogenic centre has influence on the creation of the new centre.[ddd] The influence of the pre-existing centre may contribute a steric or conformational bias such that one of the two (or more) diastereomers is selectively generated. As expected, the greatest influence is exerted by immediate adjacent centres but in certain reactions control can be exerted from more remote positions.[296, 297] Stereoselective control is often easier in cyclic substrates due to the comparatively limited conformations available. Diastereoselective control is also exerted in substrate directed reactions where a functional group associated with a pre-existing stereogenic centre controls the approach of the reactive species to a reactive diastereotopic face.[298]

Predictive models for determining the predominant diastereomer have been developed for a range of reaction types. This includes: the Zimmerman-Traxler closed transition state model for the aldol and similar reactions; [267] the Felkin-Ahn and Polar Felkin-Ahn models for nucleophilic addition to carbonyls with α-stereogenic centres; [299, 300, 301] the Cram chelation model for addition to α-heteroatom substituted carbonyls; [302] the Reetz chelation model for nucleophilic addition to β-heteroatom substituted carbonyls; [303] and the Houk model for addition to alkenes with α-stereogenic centres; [304] amongst many others.

Many predictive models require ranking substituents by comparative steric bulk. This ranking may be applied to substituents connected to a chiral centre or arranged around alkene double bond and commonly represented by the symbols R_S, R_M, R_L for small, medium and large steric bulk. A method to reliably calculate this ranking has not been successfully implemented and this currently limits the scope of diastereoselective rules that can be constructed using the methodology described in this thesis. An implementation where R_S is hydrogen, R_M is carbon with no α branching and R_L is carbon with α branching present (e.g. i-Pr, t-Bu, cyclohexyl, phenyl *etc*.) proved to be too limiting to be useful in generating fully inclusive reaction example sets from the CIRX database. A discussion on how this limitation may be overcome is presented in the final chapter.

Diastereoselectivity rules based on models influenced by electronic effects are within the scope of the current rule constraints language. These rule generally incorporate a small substituent

[ddd] Any non-stereospecific reaction that introduces two or more stereocentres may also exhibit diastereoselectivity.

(limited to hydrogen), an electron-withdrawing substituent, and a carbon substituent at the pre-existing stereocentre or they identify functional groups responsible for directed stereoselectivity. Two classes of diastereoselective transforms conforming to these model subsets are presented in the following sections: directed epoxidation of allylic alcohols; and nucleophilic addition to aldehydes with α polar substituents, specifically alkyl and enolate addition.

Diastereoselective Directed Epoxidation of Allylic Alcohols

Predictive Models

The alcohol functional group in chiral allylic alcohols can influence the diastereoselective outcome epoxidation reactions when using a peroxy acid reagent such as m-CPBA. The diastereoselective control involves the alcohol group directing the peroxy acid to a specific face of the cyclic alkene via a hydrogen bonded "spiro-butterfly" transition state (Figure 115).[305] The predominant diastereomer is dependent on ring size where rings of size 5 or 6 favour the *syn* hydroxy epoxide product while medium rings may favour the *anti*-hydroxy epoxide product.[306]

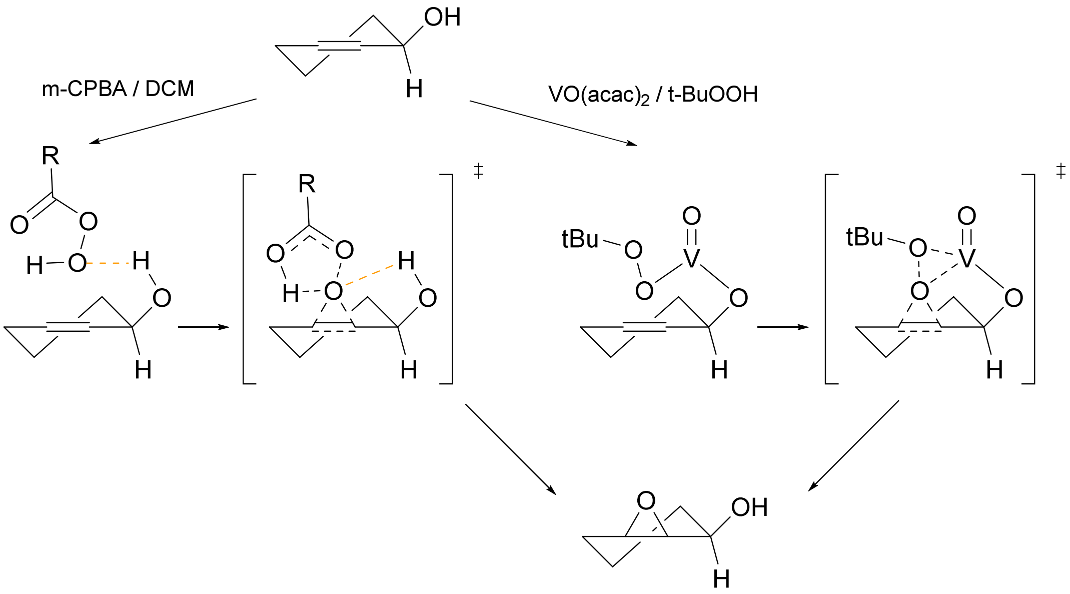


Figure 115 **The mechanisms and transitions states of directed peracid epoxidation of cyclic alkenes and transition-metal catalysed alkyl peroxide epoxidation of cyclic alkenes.**

Functional group directed diastereoselectivity can also be accomplished with the catalytic t-BuOOH / VO(acac)$_2$ or t-BuOOH / Ti(Oi-Pr)$_4$ systems via a transition-metal chelation complex that tethers the substrate to the oxidising reagent. The diastereoselectivity of the product appears to be ring size independent and consistently produces the *syn* hydroxy epoxide product.[306]

The conformational flexibility of acyclic allylic alcohols presents a tougher challenge for achieving diastereoselective control. The ability to bias the stereoselectivity is dependent on the location and types of substituent and the reagents used to oxidise the alkene.

The Sharpless model for *peroxy acid* epoxidation of secondary acyclic allylic alcohols[307, 308] predicts that the optimum steric conditions for good diastereoselective control are provided when there is a substituent *cis* to the chiral alcohol substituent that can provide steric differentiation between the alcohol rotamers. The model is geometrically related to the generalised Houk model for alkene face selectivity controlled by an α-chiral centre[304] but differs in the details of torsional angle selection on the α-chiral substituent bond. The model sets the dihedral angle between the alcohol C-O and the alkene C=C to ~120° to satisfy the steric and mechanistic requirements for the peroxy acid O-O bond to orientate *trans*-antiperiplanar to the alkene C=C so that a lone pair on the distal peroxy oxygen can maximise its interaction with the alkene π* orbital. The O-O bond is simultaneously directed to the C=C face by hydrogen bonding between the hydroxyl hydrogen and the other distal oxygen lone pair. Figure 116 shows the two reactive conformers RC3 and RC4 that can achieve this arrangement. The preferred lowest energy conformer RC4 has the smallest alcohol substituent eclipsing the *cis* alkene substituent.

The Sharpless model for *transition-metal catalysed* epoxidation model sets the dihedral angle between the alcohol C-O and the alkene C=C to ~50° to achieve the *trans*-antiperiplanar mechanistic requirements for the metal coordinated peroxide O-O bond. The lowest energy

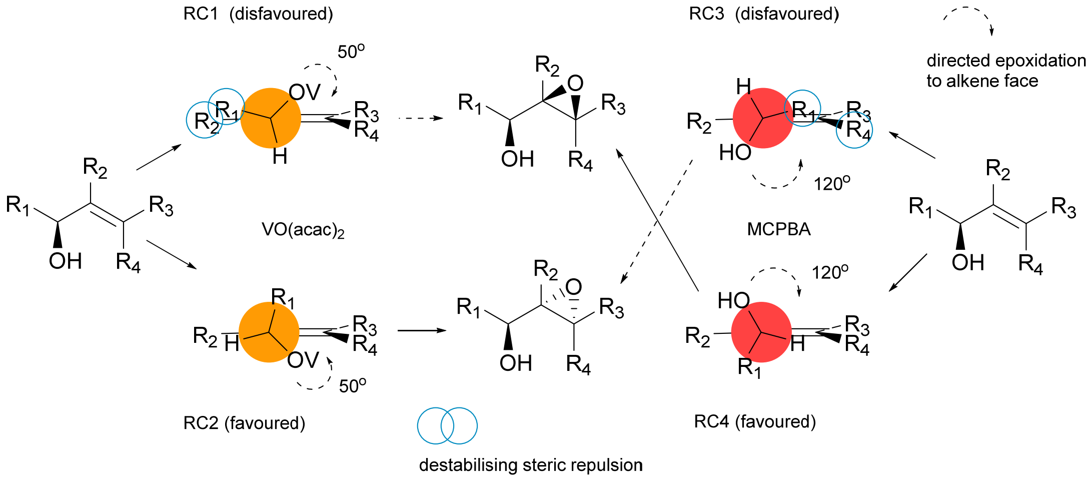


Figure 116 **Diastereoselective predictive models for peroxy acid and peroxide/transition-metal mediated epoxidation of secondary acyclic allylic alcohols.**

reactive conformer RC2 is preferred over conformer RC1 as the least steric repulsion is offered when the smallest alcohol substituent is eclipsed with the *gem*-alkene substituent.

Rules for Cyclic Allylic Alcohols

Alkene epoxidation rules E.1.1.1.4, E.1.1.3.4, E.1.1.4.4 and E.1.1.6.4 model the epoxidation of *cyclic* allylic secondary chiral alcohols where the preferred product is the *syn*-α,β-epoxy alcohol (Table 49: *vide infra*). Rule E.1.1.1.4 is further subdivided into rules E.1.1.1.4.1 and E.1.1.1.4.2 to explore the effect of ring size on the domination of the *syn* diastereomer over the *anti* diastereomer. The rule set permutes the alkene substitution patterns under the constraint that a *cis*-substituent is always present in accordance with the embedment of the rules onto small and medium rings.

Alkene epoxidation rules E.1.1.1.3, E.1.1.1.4.3, E.1.1.1.4.3, E.1.1.3.3, E.1.1.4.3 and E.1.1.6.3 model the epoxidation of cyclic allylic secondary chiral alcohols where the preferred product is the *anti*-α,β-epoxy alcohol (Table 50: *vide infra*). These rules are the diastereomer counterparts of the *syn* epoxy alcohol rule set.

The key features of the rules are the inclusion of a stereodefined environment at the alcohol group which is retained in the product. The inclusion of one hydrogen substituent at this centre precludes tertiary alcohols from consideration. The SAMERING constraint is used to ensure that all the relevant atoms are part of the same cycle. The specialised rules E.1.1.1.4.1 and E.1.1.1.4.2 add the constraints "RINGS=5,6" and "RINGS>6" on the ring bonds to subdivide the E.1.1.1.4 example set into small and medium/large rings. This ring size subdivision is repeated with the counterpart rule E.1.1.1.3. The selection of medium rings containing a *trans*-alkene is not considered. The "EWGS=0" constraints on the alkene and epoxide carbons eliminate Michael acceptor examples that may use non-directing nucleophilic epoxidation conditions. The alkene carbons also carry a "TOPICITY=DIASTEREO" constraint to eliminate C_2 symmetric allylic alcohol examples as in this case the alkene faces are homotopic and consequently non-stereoselective (Figure 117).

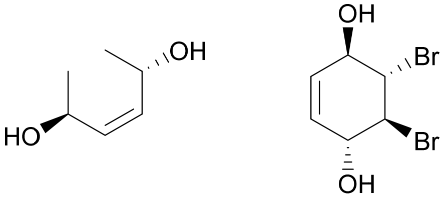


Figure 117 **C_2 symmetric allylic alcohols with homotopic C=C bonds.**

Rules for Acyclic Allylic Alcohols

Alkene epoxidation rules E.1.1.1.1, E.1.1.2.1, E.1.1.3.1, E.1.1.4.1, E.1.1.5.1 and E.1.1.6.1 model the epoxidation of non-terminal acyclic secondary chiral alcohols where the preferred product is the *syn*-α,β-epoxy alcohol. (Table 51: *vide infra*). The six rules enumerate all the substituent patterns of the allylic alcohol that generate two new chiral centres.

Alkene epoxidation rules E.1.1.1.2, E.1.1.2.2, E.1.1.3.2, E.1.1.4.2, E.1.1.5.2 and E.1.1.6.2 model the epoxidation of non-terminal acyclic secondary chiral alcohols where the preferred product is the *anti*-α,β-epoxy alcohol. (Table 52: *vide infra*). These rules are the diastereomeric counterparts of the non-terminal *syn* epoxy alcohol rule set.

The terminal alkene rules E.1.2.1.1 and E.1.2.3.1 model the epoxidation of terminal acyclic secondary chiral alcohols where the preferred product is the *syn*-α,β-epoxy alcohol with one new chiral centre. Rules E.1.2.1.2 and E.1.2.3.2 are the counterparts that generate *anti*-α,β-epoxy alcohols (Table 53: *vide infra*).

All rules use the acyclic “RINGS=NO” constraint on the bond linking the chiral secondary alcohol to the alkene to ensure that the bond is rotatable. All other non-ring constraints are applied in the same manner as described for the cyclic allylic alcohol rules. All substitution patterns are defined to probe the influence of non-hydrogen substituents to enable comparison to the predictive models.

Results and Discussion

The counts of total reaction examples and examples with *% de* values are shown for each rule in Table 49 through to Table 53. The reaction reputation, likely reaction yields and likely diastereomeric excess values with reliability estimates are colour coded to indicate the score rating (*vide supra*).

The *syn*-epoxidation of cyclic allylic alcohols is well represented for small cyclic alkenes (rule E.1.1.1.4.1 – 57 examples). Manual inspection of the matched reaction examples within this set showed that *syn*-epoxidation in small rings are dominated by the peroxy acid method (46 examples) with a smaller occurrence of the transition-metal catalysed method (9 examples). A smaller set of examples for the *anti*-epoxidation product was observed (rule E.1.1.1.3.1 - 12 examples), with a significant number of the epoxide products formed via a bromohydrin intermediate (using aqueous NBS and base).

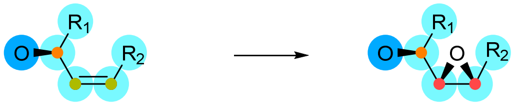
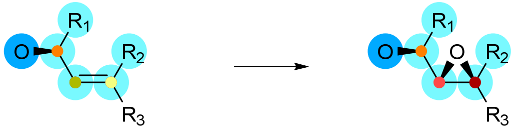
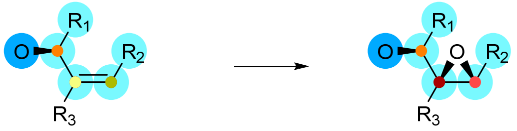
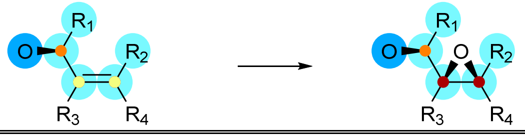
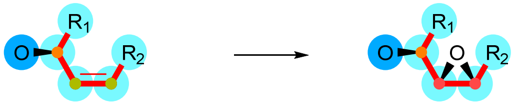
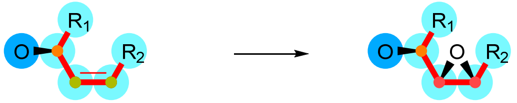
Support for the formation of *syn*- or *anti*-alkoxy epoxides in medium or large rings was very low (7 and 5 examples) although there was evidence that peroxy acids favour the formation of the *anti*-alkoxy epoxide with 4 out of 5 examples compared to only 1 out of 7 examples for the *syn*-alkoxy epoxide product. Support for substituted cyclic alkenes was also low with a total of just 12 *syn*-alkoxy epoxide and 4 *anti*-alkoxy epoxide examples. A notable feature of the cyclic allylic alcohol example sets was the very low levels of reporting of % *de* values (5 from 93 examples) making the use of the diastereoselective scoring parameter ineffective for estimating likely diastereoselectivity for the epoxidation transform.

The *syn*-epoxidation of non-terminal acyclic allylic alcohols has some representation for *cis*- and *trans*-alkenes (8 and 7 examples) and the preference is for the peroxy acid mediated method (8 examples compared to 3 examples of transition-metal catalysis). Other substitution patterns are largely unsupported. Support for the *anti*-adducts of *trans*-allylic alcohols is absent with only *cis*-allylic alcohols and *gem-trans*-allylic alcohols represented (9 and 12 examples) where the preference is for transition-metal catalysis (15 examples compared to 2 peroxy acid examples).

The *syn*-epoxidation of *terminal* allylic alcohols has low representation (rules E.1.2.1.1 and E.1.2.3.1) with a total of 7 examples largely mediated with peroxy acids (4 examples). The *anti*-epoxidation of terminal allylic alcohols is better presented (rules E.1.2.1.2 and E.1.2.3.2) with a total of 32 examples largely mediated by transition-metal catalysis (17 examples). The diastereoselectivity rating of the latter is very good with excellent reliability (from 11 examples).

The acyclic allylic alcohol example sets are better supported with reported diastereoselective ratios (30 from 83 examples) making the use of the *% de* scoring parameter more reliable especially for *anti*-alkoxy epoxides formed from terminal allylic alcohols.

Overall the examples selected for the epoxidation rules support a strong preference for functional group directed diastereoselective control using either peroxy acid or peroxide/transition-metal catalysed methods. The preference for *syn* selectivity in small rings is adequately supported. Anti diastereoselectivity in small rings is weakly indicated via non-directed methods using halohydrin intermediates. The effect of medium/large rings on diastereoselectivity is not strongly observed due to a lack of sufficient supporting examples. The acyclic epoxidation rules are only well supported for terminal allylic alcohols but good overall support is evident for the Sharpless model of *syn* and *anti* diastereoselectivity for peroxy acids and peroxide/transition-metal catalysts. The enumeration of the available substitution patterns for both the cyclic and acyclic rules enables the retrosynthesis executive to prune out unsupported cases (see next chapter).

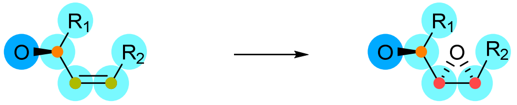
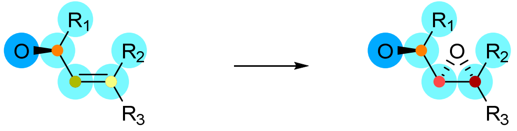
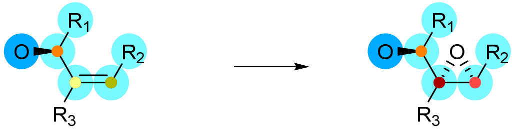
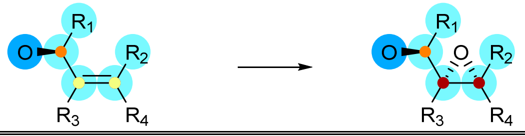
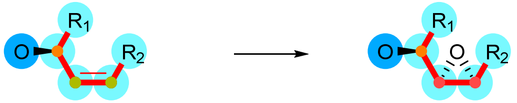
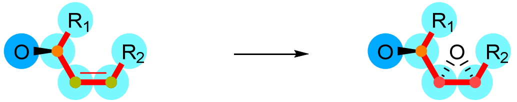
Rule identifier	Rule Pattern	A	B	C	D	E	F	Notes
E.1.1.1.4	 ● SAMERING=1	60	5	80	22			1. m-CPBA (40) 2. ROOH / VO(acac)$_2$ (6) 3. Other/unknown (14)
E.1.1.3.4	 O R$_1$ R$_2$ R$_3$ O R$_1$ R$_2$ R$_3$ ● FGS=ALCOHOL;HS=1 ● HS=1	6	0	93	3			1. m-CPBA (4) 2. ROOH / VO(acac)$_2$ (1) 3. Other/unknown (1)
E.1.1.4.4	 O R$_1$ R$_2$ R$_3$ O R$_1$ R$_2$ R$_3$ ● HS=0;EWGS=0;TOPICITY=DIASTEREO ● HS=1;EWGS=0;TOPICITY=DIASTEREO	3	0					1. m-CPBA (2) 2. Other/unknown (1)
E.1.1.6.4	 O R$_1$ R$_2$ R$_3$ R$_4$ O R$_1$ R$_2$ R$_3$ R$_4$ ● HS=0;EWGS=0 ● HS=1;EWGS=0	3	0					1. ROOH / VO(acac)$_2$ (2) 2. Other/unknown (1)
E.1.1.1.4.1	 — RINGS = 5, 6	57	5	80	23			1. m-CPBA (39) 2. ROOH / VO(acac)$_2$ (4) 3. Other/unknown (14)
E.1.1.1.4.2	 — RINGS > 6	7	0	84	21			1. m-CPBA (1) 2. ROOH / VO(acac)$_2$ (2) 3. Other/unknown (4)

Key: ● Superb ● Excellent ● Very good ● Good ● Fair ● Poor ● Very poor ● No support

Table 49 Directed diastereoselective epoxidation of allylic alcohols (part 1) showing: cyclic alkenes converted to *syn*-epoxy alcohols.

Columns:

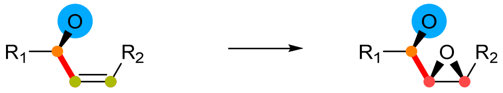
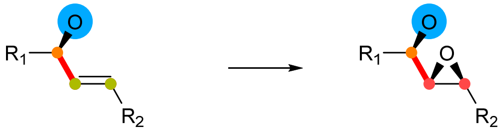
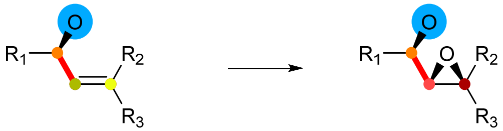
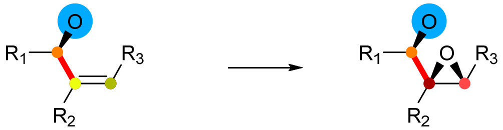
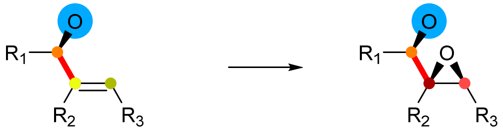
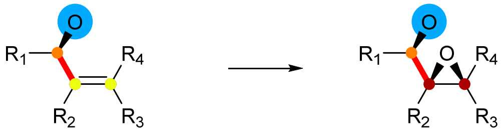
A – Number of examples	**B – Number of examples with *% de* values**
C – Estimated yield (median %)	**D – Reliability of the estimated yield (IQR spread in %)**
E – Estimated *% de* value (median %)	**F – Reliability of the estimated *% de* value (IQR spread in %)**

Rule identifier	Rule Pattern	A	B	C	D	E	F	Notes
E.1.1.1.3	 ● SAMERING=1	17	0	73	32			1. m-CPBA (8) 2. NBS / H_2O / t-BuOK (4) – via bromohydrin 3. Other/unknown (5)
E.1.1.3.3	 ● FGS=ALCOHOL;HS=1 ● HS=1	4	0					1. m-CPBA (3) 2. ROOH / VO(acac)$_2$ (1)
E.1.1.4.3	 ● HS=0;EWGS=0;TOPICITY=DIASTEREO ● HS=1;EWGS=0;TOPICITY=DIASTEREO	1	0					
E.1.1.6.3	 ● HS=0;EWGS=0 ● HS=1;EWGS=0	1	1					
E.1.1.1.3.1	 — RINGS = 5, 6	12	0	72	25			1. m-CPBA (4) 2. NBS / H_2O / t-BuOK (4) 3. Other/unknown (4)
E.1.1.1.3.2	 — RINGS > 6	5	0					1. m-CPBA (4) 2. Other/unknown (1)

Key: ● Superb ● Excellent ● Very good ● Good ● Fair ● Poor ● Very poor ● No support

Table 50 Directed diastereoselective epoxidation of allylic alcohols (part 2) showing: cyclic alkenes converted to *anti*-epoxy alcohols.

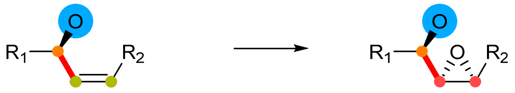
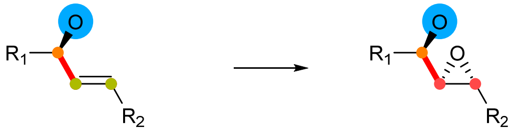
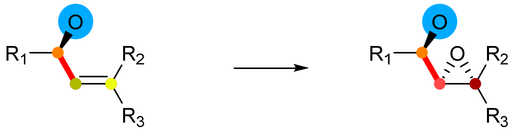
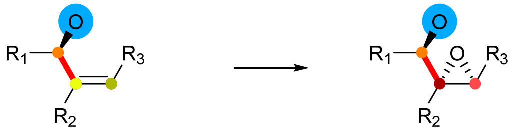
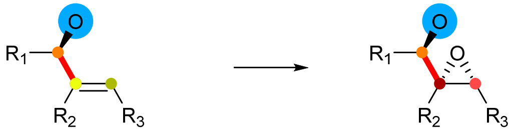
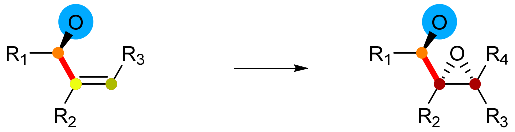
Columns: **A – Number of examples** **B – Number of examples with *% de* values**
C – Estimated yield (median %) **D – Reliability of the estimated yield (IQR spread in %)**
E – Estimated *% de* value (median %) **F – Reliability of the estimated *% de* value (IQR spread in %)**

Rule identifier	Rule Pattern	A	B	C	D	E	F	Notes
E.1.1.1.1		8	4	85	15			1. m-CPBA (5) 2. ROOH / Mo(CO)$_6$ (1) 3. Other/unknown (2)
E.1.1.2.1	 — RINGS = NO	7	2	70	20			1. m-CPBA (3) 2. ROOH / Ti(O-iPr)$_4$ (2) 3. Other/unknown (2)
E.1.1.3.1	 ● FGS=ALCOHOL;HS=1 ● HS=1	2	1					1. m-CPBA (1) 2. Other/unknown (1)
E.1.1.4.1	 ● HS=0;EWGS=0;TOPICITY=DIASTEREO ● HS=1;EWGS=0;TOPICITY=DIASTEREO	0	0					
E.1.1.5.1	 ● HS=0;EWGS=0 ● HS=1;EWGS=0	3	0					1. ROOH / Ti(O-iPr)$_4$ (2) 2. m-CPBA (1)
E.1.1.6.1		0	0					

Key: ● Superb ● Excellent ● Very good ● Good ● Fair ● Poor ● Very poor ● No support

Table 51 Directed diastereoselective epoxidation of allylic alcohols (part 3) showing: acyclic alkenes converted to *syn*-epoxy alcohols.

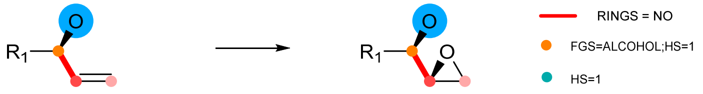
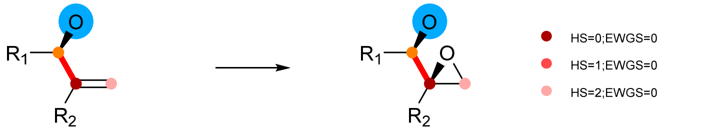
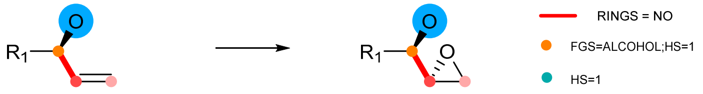
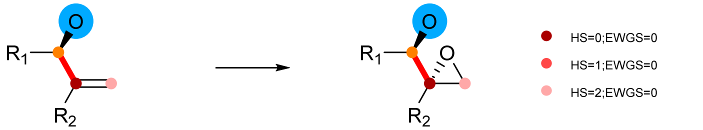
Columns: A – Number of examples B – Number of examples with *% de* values
C – Estimated yield (median %) D – Reliability of the estimated yield (IQR spread in %)
E – Estimated *% de* value (median %) F – Reliability of the estimated *% de* value (IQR spread in %)

Rule identifier	Rule Pattern	A	B	C	D	E	F	Notes
E.1.1.1.2		1	0					
E.1.1.2.2	 — RINGS = NO	9	5	73	37			1. ROOH / Ti(O-iPr)$_4$ (5) 2. ROOH / VO(acac)$_2$ (1) 3. m-CPBA (1) 4. Other/unknown (2)
E.1.1.3.2	 ● FGS=ALCOHOL;HS=1 ● HS=1	0	0					
E.1.1.4.2	 ● HS=0;EWGS=0;TOPICITY=DIASTEREO ● HS=1;EWGS=0;TOPICITY=DIASTEREO	0	0					
E.1.1.5.2	 ● HS=0;EWGS=0 ● HS=1;EWGS=0	12	8	84	9	90	10	1. ROOH / Ti(O-iPr)$_4$ (7) 2. ROOH / VO(acac)$_2$ (2) 3. m-CPBA (1) 4. Other/unknown (2)
E.1.1.6.2		0	0					

Key: ● Superb ● Excellent ● Very good ● Good ● Fair ● Poor ● Very poor ● No support

Table 52 Directed diastereoselective epoxidation of allylic alcohols (part 4) showing: acyclic alkenes converted to *anti*-epoxy alcohols.

Columns: A – Number of examples
B – Number of examples with *% de* values
C – Estimated yield (median %)
D – Reliability of the estimated yield (IQR spread in %)
E – Estimated *% de* value (median %)
F – Reliability of the estimated *% de* value (IQR spread in %)

Rule identifier	Rule Pattern	A	B	C	D	E	F	Notes
E.1.2.1.1		3	0					1. Other/unknown (3)
E.1.2.3.1	 O R1 R2 O HS=0;EWGS=0 HS=1;EWGS=0 HS=2;EWGS=0	4	0					1. m-CPBA (4)
E.1.2.1.2		15	6	65	33	90	9	1. ROOH / Ti(O-iPr)$_4$ (7) 2. ROOH / VO(acac)$_2$ (3) 3. Other/unknown (5)
E.1.2.3.2	 O R1 R2 O HS=0;EWGS=0 HS=1;EWGS=0 HS=2;EWGS=0	17	7	80	22	90	6	1. ROOH / VO(acac)$_2$ (4) 2. ROOH / Ti(O-iPr)$_4$ (3) 3. M-CPBA (1) 4. Other/unknown (9)

Key: ● Superb ● Excellent ● Very good ● Good ● Fair ● Poor ● Very poor ● No support

Table 53 Directed diastereoselective epoxidation of allylic alcohols (part 5) showing: terminal acyclic alkenes converted to *syn*- and *anti*-epoxy alcohols.

Columns: A – Number of examples
C – Estimated yield (median %)
E – Estimated *% de* value (median %)

B – Number of examples with *% de* values
D – Reliability of the estimated yield (IQR spread in %)
F – Reliability of the estimated *% de* value (IQR spread in %)

Diastereoselective Nucleophilic Addition to α-Chiral Aldehydes

Predictive Models

The development and refinement of predictive models for the stereoselective nucleophilic addition to α-substituted aldehydes and ketones has been the subject of on-going research over the last 60 years beginning with models developed by Cram[309] and Cornforth.[310] The current accepted models for nucleophilic addition to aldehydes and acyclic ketones is due to developments by Felkin[299, 300] and Anh.[311, 301] The key features of the Felkin-Anh (FA) model are: the preferential selection of a reaction-like reactive conformer where the α-substituents are fully staggered to minimise steric repulsion with the carbonyl centre; the substituent with the best acceptor σ*orbital is aligned 90° to the C=O bond; the nucleophile approaches the carbonyl bond along the Bürgi-Dunitz trajectory (95° to 105°);[312, 313] and selection of the carbonyl face offering the least destabilising steric repulsion to the incoming nucleophile. Figure 118 shows the two lowest energy reactive conformations RC1 and RC2 that compete to form products. In the absence of an electron-withdrawing substituent the model predicts that the bulkiest substituent occupies the orthogonal substituent position.

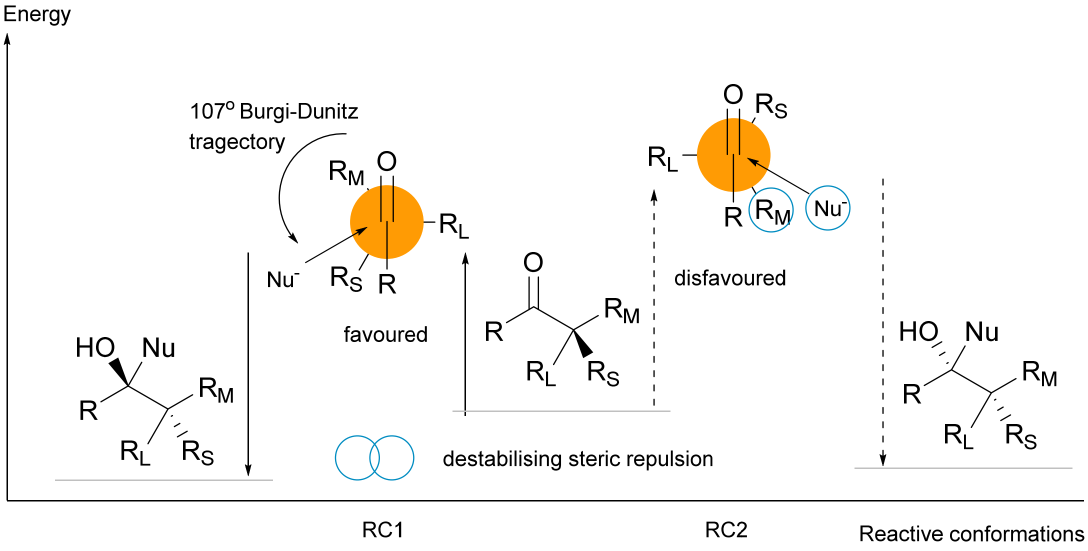


Figure 118 The Felkin-Anh (FA) model for nucleophilic addition to α-chiral carbonyls with non-polar α-substituents. Substituents sizes are $R_L > R_M > R_S$.

The polar Felkin-Anh model (PFA) specifically addresses the presence of an electron-withdrawing substituent (C-Z) at the α-chiral centre by invoking transition state stabilisation through hyperconjugation of the $\sigma^*_{(C\text{-}Z)}$ + $\pi^*_{(C=O)}$ anti-bonding orbitals as the dominant factor. Hyperconjugation lowers the energy of the aldehyde LUMO and is minimal when the C-Z bond is orthogonal to the C=O bond. Thus the lowest energy LUMOs are achieved in pre-transition state conformers RC1 and RC2 (Figure 119). Reactive conformer RC2 offers the least steric hindrance to the approaching nucleophile and hence determines the preferred product stereoisomer.

The FA and PFA models predict the stereochemical outcome in many simple situations but the reality for more complex cases is that many of the potential reactive conformers possess similar energies in which many secondary interactions in more remote regions of the molecule tip the balance one way or the other. This makes the development of a single overall model an extremely difficult undertaking.

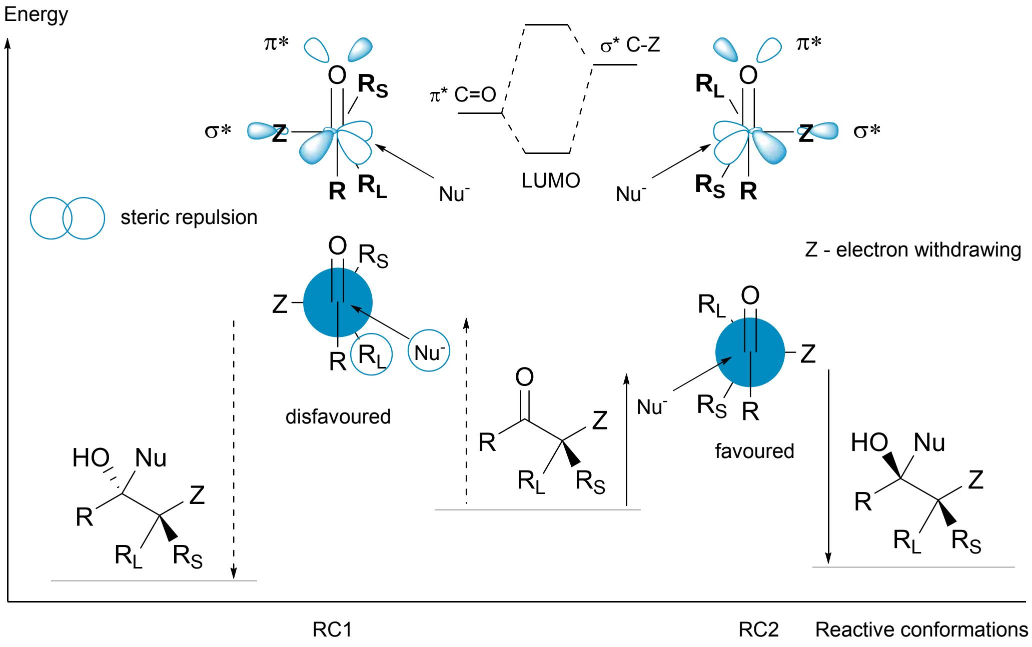


Figure 119 The Polar Felkin-Anh (PFA) model for nucleophilic addition to α-chiral carbonyls with an electron-withdrawing α-substituent.

Recently Evans has resurrected and updated[eee] an older model devised by Cornforth that proposed dipole minimisation as the dominant mechanism of stereoselection in carbonyls containing strong polar α-substituents.[314] The Cornforth-Evans (CE) model predicts the same outcome as the PFA model for nucleophiles creating one new stereogenic centre (Figure 120). However the two models are differentiated in more complex cases where two or more new stereogenic centres are created. For example the CE model has proved successful in predicting the stereochemical outcome of enolboron mediated aldol reactions where the standard PFA

[eee] Evans added the Felkin staggered conformation and the Bϋrgi-Dunitz nucleophile attack trajectory to the Cornforth model. Dipole and steric minimisation occurs at ~150° and not at the fully opposed 180°$_{(C\text{-}Z,\ C=O)}$ dihedral angle of the original Cornforth model.

model has presented difficulties (*vide infra*).[315] Theoretical studies suggest that the PFA model is favoured when Z is less electronegative (-SR, -PR_2) whereas the CE model better predicts the lowest energy rotamer when Z is strongly electronegative (-F, -Cl, -OR). The PFA and CE models are also sensitive to the nature of the nucleophile *via* the strength of the $Nu_{(HOMO)} \rightarrow \sigma^*_{(C\text{-}Z)}$ interaction. For example calculations indicate that enolborane nucleophiles are weakly stabilised by the PFA hyperconjugation mechanism compared to CE electrostatic stabilisation.[314]

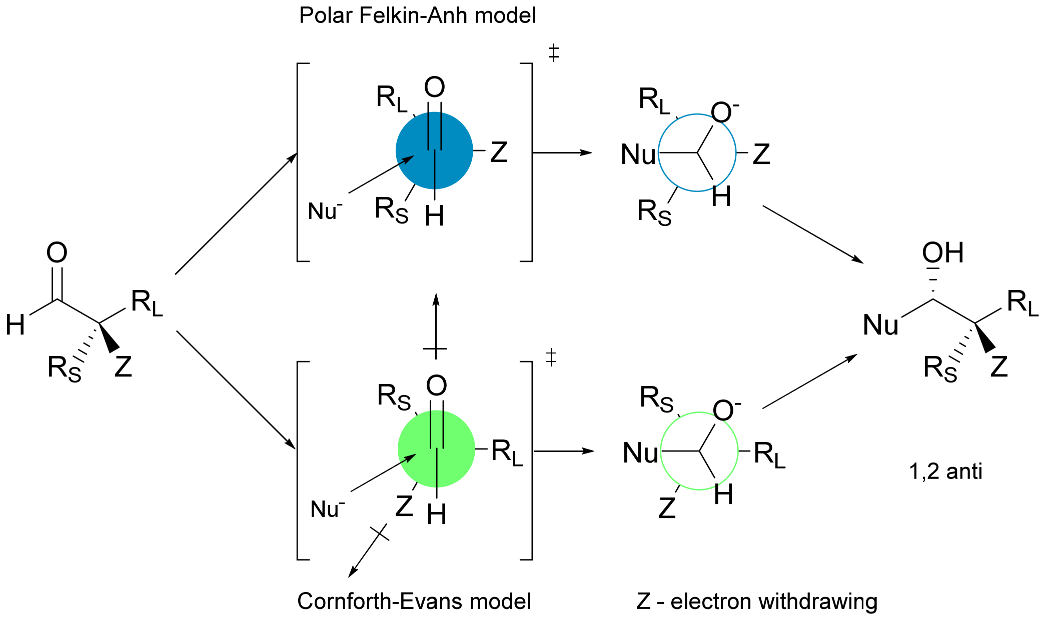


Figure 120 The favoured rotamers due to the Polar Felkin-Anh and Cornforth-Evans (CE) electrostatic models for nucleophilic addition to α-heteroatom substituted aldehydes.

The stereochemical outcome of nucleophilic addition to chiral aldehydes and ketones with an α–donor atom is predicted with a high degree of certainty by the Cram Chelation (CC) model when reacting under *chelating* conditions (Figure 121).[302] The presence of Lewis acidic reagents capable of bidentate coordination with the carbonyl oxygen and the α-donor atom locks the conformation of the carbonyl compound such that the donor atom eclipses the carbonyl oxygen. Under chelating conditions the ratio of the diastereomeric products is determined in part by the steric differentiation between the R_S and R_L substituents partially shielding each face of the C=O bond and the position of the equilibrium between non-chelated and chelated aldehyde populations. The degree of chelation is controlled by the Lewis acidity of the coordinating metal, the reaction temperature, the nature of the substituents on the α–donor atom, and the competitive ability of the solvent to coordinate with the metal.[303] The CC and PFA models place the smallest substituent R_S on opposing sides of the C=O bond allowing stereochemical control to be exerted to produce either 1,2-*syn* or 1,2-*anti* diastereomer adducts under chelating and non-chelating conditions respectively.

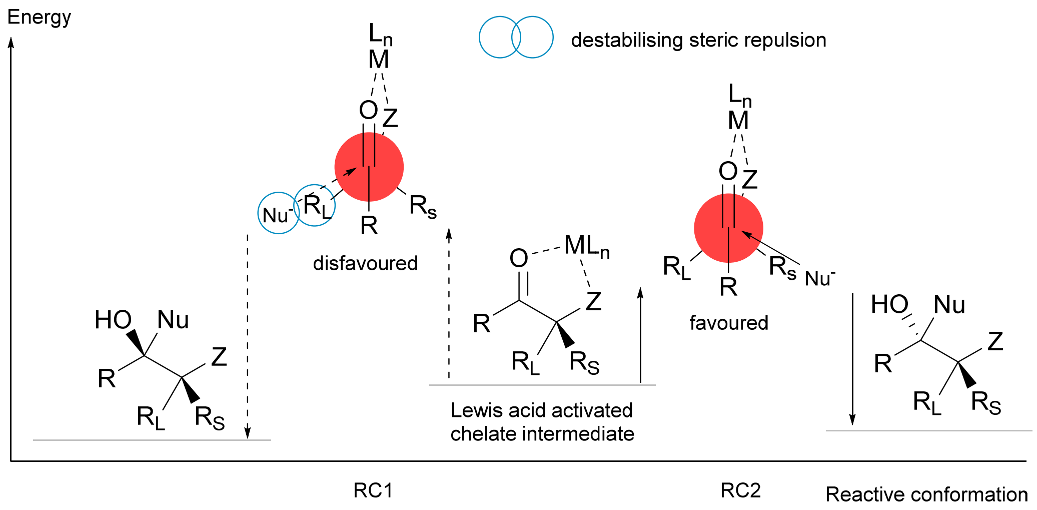


Figure 121 The Cram Chelation (CC) model for nucleophilic addition to α-chiral carbonyls with a donor α-substituent.

Stereoselective control in nucleophilic additions to carbonyls can be extended to the assembly of adducts in which multiple contiguous stereocentres are constructed. For example substituted achiral enolate additions to chiral α-substituted aldehydes produce adducts containing three contiguous stereocentres. Composite models combining the Zimmerman-Traxler and Felkin-Anh or (modified) Cornforth models have been used to predict the dominant product diastereomer in these aldol reactions. Reactions proceeding via *Z*-enolates reacting with α-heteroatom substituted aldehydes produce mixtures of 2,3-*syn,* 3,4-*syn*[fff] and 2,3-*syn,* 3,4-*anti* adducts while those proceeding via *E*-enolates produce mixtures of 2,3-*anti,* 3,4-*syn* and 2,3-*anti,* 3,4-*anti* adducts (see Figure 122 and Figure 123).[315] The presence of the stereogenic centre α to the aldehyde results in differentiation between the Re and Si faces of the aldehyde which in turn dictates which of two alternative Zimmerman-Traxler arrangements is preferred. Differences in *syn*-pentane steric interactions between the enolate substituents and the substituents of the *dominant* rotamer of the aldehyde determine which reactive conformation has the lowest energy in each Zimmerman-Traxler arrangement and this determines the dominant product when reacting under kinetic conditions.[316]

[fff] See Figure 122 for the reference atom numbering.

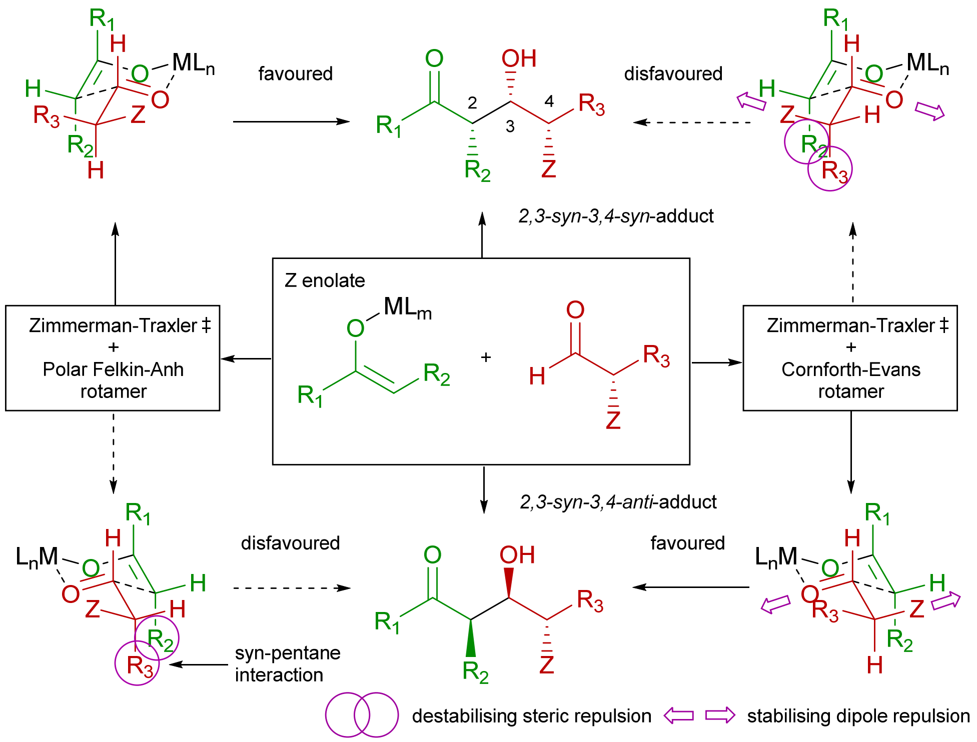


Figure 122 The Polar Felkin-Anh and Cornforth-Evans model outcomes for the aldol reaction of *Z*-enolates with α-heteroatom substituted aldehydes.

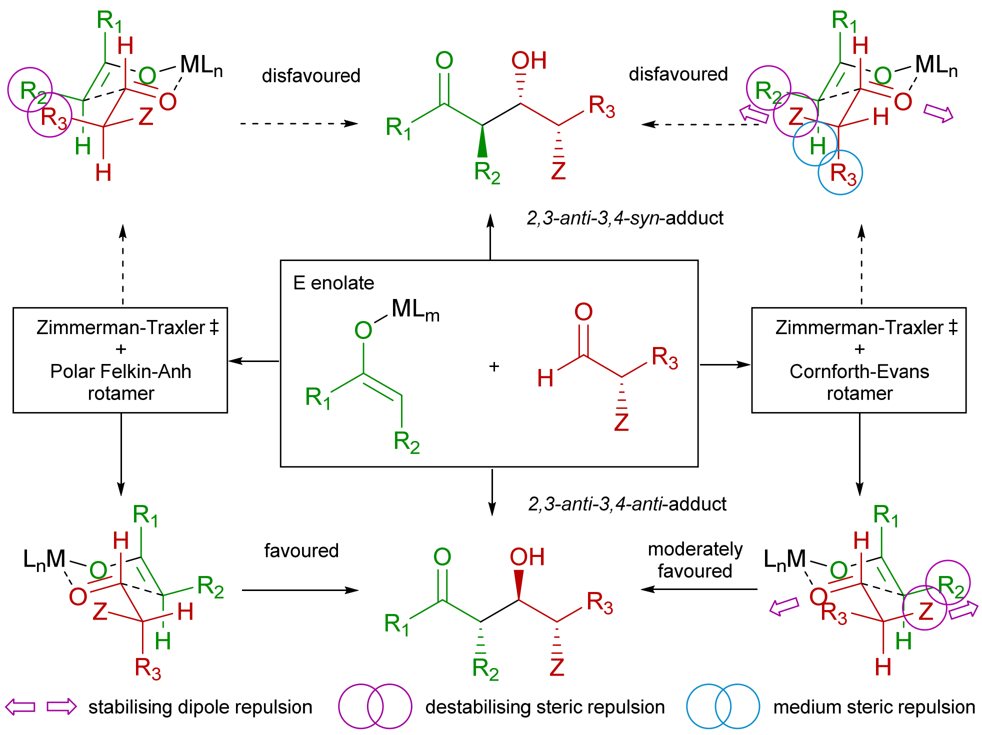


Figure 123 The Polar Felkin-Anh and Cornforth-Evans model outcomes for the aldol reaction of *E*-enolates with α-heteroatom substituted aldehydes.

Evans has observed that the outcomes of the reaction of boron and lithium *Z*- and *E*-enolates with α-alkoxy aldehydes are inconsistent with the hyperconjugation controlled PFA model but are better modelled by invoking the dipole controlled CE model. The PFA model predicts that *E*-enolates induce *better* 3,4-*anti* selectivity compared to *Z*-enolates whereas the CE model predicts the observed outcome, that *Z*-enolates induce *better* 3,4-*anti* selectivity compared to *E*-enolates. [315]

The general versatility of the aldol reaction is such that many alternative methods of stereo control are available. Methods employing enzymes or organocatalysts such as (*R*)- or (*S*)-proline have stereochemical outcomes determined by mechanisms that do not involve the Zimmerman-Traxler chair-like transition state (*e.g.* the hydrogen-bond driven Houk-List proline catalysis model[317, 318]) and so may lead to different diastereoselective preferences. For example the use of chelation control and/or metals that prefer octahedral (as opposed to tetrahedral) coordination spheres in conjunction with chiral *N*-acyloxazolidinones or *N*-acylthiazolidinethiones auxiliaries result in a reactive boat transition state that inverts the normal stereochemical preference.[268, 285]

Enhanced diastereoselective effects occur when *both* reacting partners carry chiral centres *via* the mechanism of double stereo differentiation (DSD).[245] Enhanced diastereoselectivity occurs when a pairing of enantiomers of the two chiral substrates are sterically matched and working in concert in at least *one* available transition state. Conversely the steric interactions of mismatched enantiomer pairs work against each other in *all* available transition states, and this may significantly reduce the observed diastereoselectivity between the alternative product stereoisomers (see Figure 124). Clever use can be made of double stereo differentiation, for example it can be exploited in the reaction of *achiral* enolate precursors with *chiral* aldehydes to control which diastereomer is produced *via* the introduction of a chiral auxiliary or the use of transient chiral reagents/catalysts used to form a *chiral* intermediate enolate after which the *temporary* chiral agent is removed.[319]

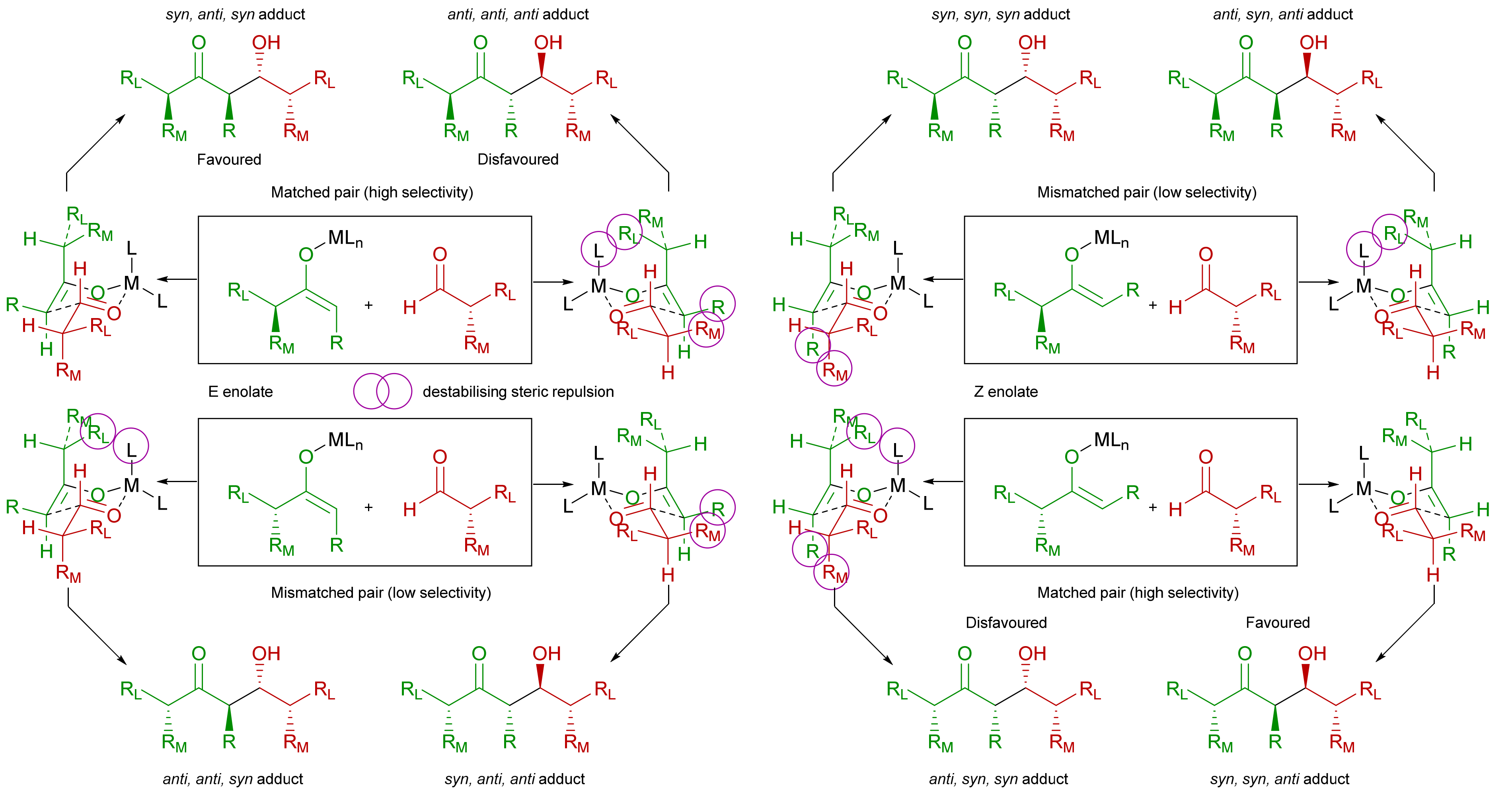


Figure 124 **Matched and mismatched selectivity models combining Zimmerman-Traxler, Felkin-Anh and Houk components showing predicted outcomes for the aldol reaction between non-hetero α-chiral ketones and non-heteroatom substituted α-chiral aldehydes *via* preformed *E*- and *Z*-enolate intermediates.**

Rules for Alkyl Addition to α-Heteroatom Substituted Aldehydes

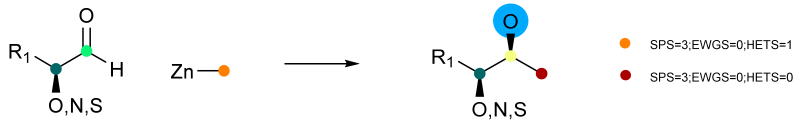
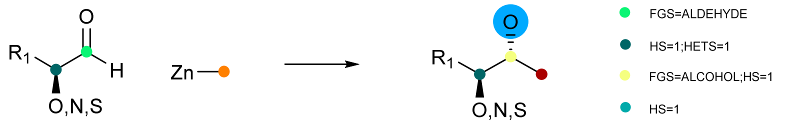
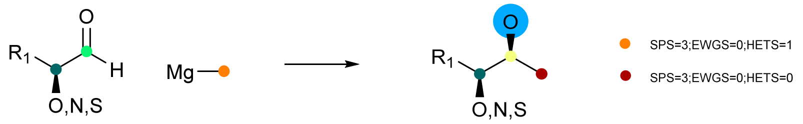
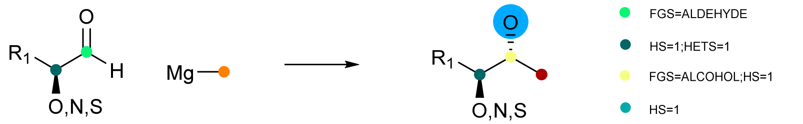
Rules RZC.1.1.2.1 and RMC.1.1.2.1 model the generation of 1,2-*syn* adducts via the nucleophilic addition of alkylzinc and alkyl Grignard reagents to α-heteroatom substituted aldehydes. The corresponding 1,2-*anti* adducts are handled by rules RZC.1.1.2.2 and RMC.1.1.2.2 (see Table 54: *vide infra*). The alkyl nature of the nucleophile is set via the hybridisation constraint SPS=3 (sp^3) on the atom connected to the metal. The α-chiral aldehydes are limited to those containing one hydrogen substituent, an electron-withdrawing atom with lone pairs (O,N,S) and an additional carbon substituent. The use of the HS=1;HETS=1 constraint expression on the α-chiral carbon atom ensures that substituent R_1 must connect to a carbon atom. These rules are designed to determine the level of example support for chelation and non-chelation controlled stereoselectivity with selected alkyl metal reagents capable of chelation.

Rules for Enolate Addition to α-Heteroatom Substituted Aldehydes

Rules A.2.3.1.1.6 and A.2.3.1.1.7 (Table 55: *vide infra*) model the generation of 3,4-*syn* and 3,4-*anti* adducts via the nucleophilic addition of unsubstituted enolates to α-heteroatom substituted aldehydes. The rules are designed to extend the aldol transform (*vide supra*) and they inherit established atom and bond constraints from the parent rule A.2.3.1.1 while adding an unchanged α-chiral centre to the aldehyde substructure. The α-chiral aldehydes are limited to those containing one hydrogen substituent, an electron-withdrawing atom (O,N,S) and a carbon substituent.

In a similar manner, rules A.2.1.2.1.12 and A.2.1.2.1.13 (Table 55: *vide infra*) model the addition of *E*-enolate precursors to α-heteroatom substituted aldehydes leading to the 2,3-*anti* 3,4-*syn* and 2,3-*anti* 3,4-*anti* adducts. Rules A.2.1.1.1.14 and A.2.1.1.1.14 model the addition of *Z*- enolate precursors to α-heteroatom substituted aldehydes leading to the 2,3-*syn* 3,4-*syn* and 2,3-*syn* 3,4-*anti* adducts.

The supposed intermediacy of *E*- or *Z*-enolates is inferred from the relative stereochemistry of the 2,3 chiral centres found in the product substructure *assuming* a Zimmerman-Traxler transition state.[272] The diastereoselective aldol rules are designed to determine the level of example support for the generation of each of the four possible stereoisomers without concern for the mechanism that gives rise to them.

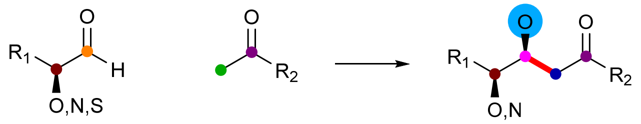
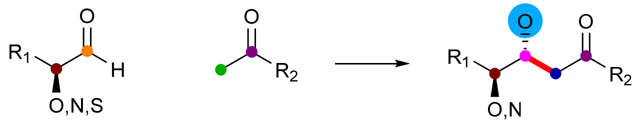
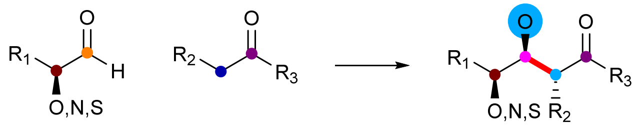
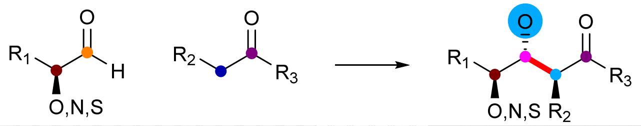
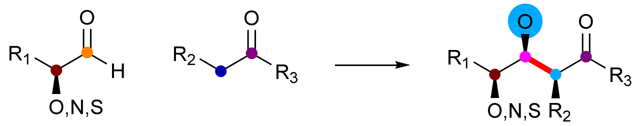
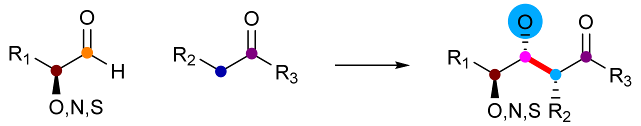
Rule identifier	Rule Pattern	A	B	C	D	E	F	Notes
RZC.1.1.2.1	 O R1 H O,N,S Zn O R1 O,N,S SPS=3;EWGS=0;HETS=1 SPS=3;EWGS=0;HETS=0	28	21	69	27	90	6	Cram Chelation control → 1,2-*syn* adduct
RZC.1.1.2.2	 O R1 H O,N,S Zn O R1 O,N,S FGS=ALDEHYDE HS=1;HETS=1 FGS=ALCOHOL;HS=1 HS=1	4	3	■	■	■	■	Polar Felkin-Anh control → 1,2-*anti* adduct
RMC.1.1.2.1	 O R1 H O,N,S Mg O R1 O,N,S SPS=3;EWGS=0;HETS=1 SPS=3;EWGS=0;HETS=0	48	33	62	21	56	64	Cram Chelation control → 1,2-*syn* adduct
RMC.1.1.2.2	 O R1 H O,N,S Mg O R1 O,N,S FGS=ALDEHYDE HS=1;HETS=1 FGS=ALCOHOL;HS=1 HS=1	32	19	74	22	96	15	Polar Felkin-Anh control → 1,2-*anti* adduct

Key: ● Superb ● Excellent ● Very good ● Good ● Fair ● Poor ● Very poor ● No support

Table 54 Rules for distereoselective alkylzinc and alkyl Grignard mediated additions to α-hetero aldehydes under chelation and non-chelation conditions.

Columns:

A – Number of examples	**B – Number of examples with *% de* values**
C – Estimated yield (median %)	**D – Reliability of the estimated yield (IQR spread in %)**
E – Estimated *% de* value (median %)	**F – Reliability of the estimated *% de* value (IQR spread in %)**

Rule identifier	Rule Pattern	A	B	C	D	E	F	Notes
A.2.3.1.1.6	 ● HS = 1; HETS=1 ● FGS = ALDEHYDE ● FGS = ALCOHOL; HS = 1 ● PROP = EWG	27	15	75	29	94	12	3,4-*syn* adduct Enolate types: Organocatalyst/enzyme - 12; B - 5; Li – 3
A.2.3.1.1.7	 ● HS = 3 ● HS = 2 ● HS = 1	44	20	70	20	92	11	3,4-*anti* adduct Enolate types: Li –16; B – 12; organocatalyst/enzyme - 8
A.2.1.2.1.12	 ● HS = 1; HETS=1 ● FGS = ALDEHYDE	15	8	60	18	82	50	2,3-*anti* 3,4-*syn* adduct Enolate types: B – 5; Li – 5; Ti – 1; organocatalyst/enzyme – 6
A.2.1.2.1.13	 ● FGS = ALCOHOL; HS = 1 ● PROP = EWG	31	11	61	25	96	3	2,3-*anti* 3,4-*anti* adduct Enolate types: B – 6; Li – 5; Ti – 1; organocatalyst/enzyme – 9
A.2.1.1.1.14	 — RINGS = NO	22	5	70	29			2,3-*syn* 3,4-*syn* adduct Enolate types: B – 6; Ti – 4; organocatalyst/enzyme – 6
A.2.1.1.1.15	 ● HS = 2 ● HS = 1	52	11	70	26	94	42	2,3-syn 3,4-*anti* adduct Enolate types: B – 12; Ti – 10; Li – 8; organocatalyst/enzyme - 8

Key: ● Superb ● Excellent ● Very good ● Good ● Fair ● Poor ● Very poor ● No support

Table 55 Rules for distereoselective unsubstituted and substituted enolate addition to α-hetero aldehydes.

Columns:

A – Number of examples

B – Number of examples with *% de* values

C – Estimated yield (median %)

D – Reliability of the estimated yield (IQR spread in %)

E – Estimated *% de* value (median %)

F – Reliability of the estimated *% de* value (IQR spread in %)

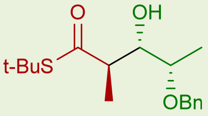
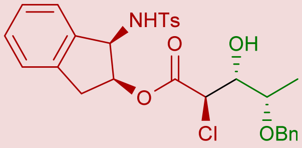
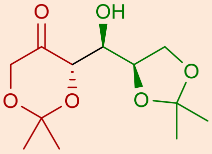
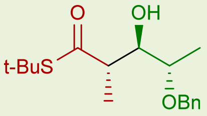
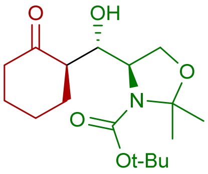
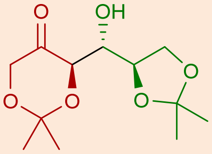
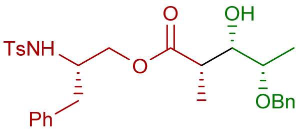
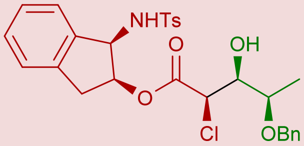
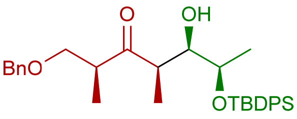
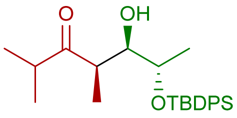
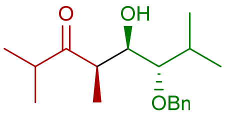
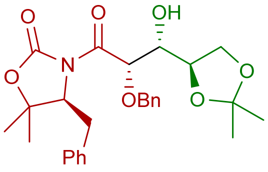
Rule A.2.1.2.1.12	2,3-*anti*-3,4-*syn* adduct	3 of 15 examples
 A Bis((1R,3S,4S)-3-menthylmethyl)BBr / base/Et$_2$O Yield 60% (98% de; 100% ee)[95]	 B From (S)-BnOCH(Me)CHO using 1) TiCl$_4$/base 2) TiCl$_4$/RCHO Yield 82% (98% de; 100% ee)[96]	 C Cy$_2$BCl/base/DCM Yield 59% (74% de)[97]
Rule A.2.1.2.1.13	**2,3-*anti*-3,4-*anti* adduct**	**3 of 31 examples**
 D Bis((1S,3R,4R)-3-menthylmethyl)BBr / base/Et$_2$O Yield 64% (98% de; 100% ee)[95]	 E D-proline/CHCl$_3$/DMSO Yield 81% (98% de)[98]	 F D-proline/DMF Yield 76% (96% de; 98% ee)[99]
Rule A.2.1.1.1.14	**2,3-*syn*-3,4-*syn* adduct**	**3 of 22 examples**
 G TiCl$_4$/base/DCM Yield 84% (98% de)[100]	 H From (R)-BnOCH(Me)CHO using 1) TiCl$_4$/base 2) TiCl$_4$/RCHO Yield 94% (98% de; 100% ee)[96]	 J TiCl$_3$OiPr/base/DCM Yield 90% (44% de)[101]
Rule A.2.1.1.1.14	**2,3-*syn*-3,4-*anti* adduct**	**3 of 52 examples**
 K TiCl$_4$/base/DCM Yield 78% (49% de)[102]	 L TMS$_2$NH/BuLi/THF Yield 70% (98% de)[92]	 M Et$_2$BOTf/base/DCM Yield 73% (98% de)[103]

Table 56 **Example reactions for the addition of E/Z substituted enolates to α-hetero aldehydes producing each of the four diastereomer arrangements. Related diastereomeric compounds are highlighted with matching background colours.**

Results and Discussion

Alkyl Addition to α-Heteroatom Substituted Aldehydes

The reaction example counts, yields and stereoselectivity statistics for rules describing alkyl addition to α-heteroatom substituted aldehydes using zinc and Grignard reagents are presented in Table 54.

The reaction example support for the chelation controlled synthesis of 1,2-*syn* alkyl adducts from alkylzinc reagents is good, with very low support for non-chelation control. Stereoselectivity under chelation controlled conditions is generally high and very reliable (RZC.1.1.2.1, median de: 90%; spread: 6% from 21 examples) although median yield and yield reliability is modest. Grignard alkyl addition can be used to generate both 1,2-*syn* and 1,2-*anti* products. Under non-chelating conditions the support statistics indicate that the 1,2-*anti* adducts are produced with very good stereoselectivity and good reliability (RMC.1.1.2.2, median de: 96%; spread: 15% from 19 examples). However the 1,2-*syn* chelation controlled route offers poor stereoselectivity with very poor reliability (RMC.1.1.2.1, median de: 56%; spread: 64% from 33 examples) indicating that Grignard reagents are generally a poor choice for chelation control.

The switch from chelation to non-chelation control is determined by the steric bulk of the substituent attached to the α-heteroatom. Small groups such as PMB, Bn, MOM and Me favour chelation control while bulky silyoxy substituents generally frustrate the ability of magnesium (but not zinc) to chelate to the α-heteroatom. [320, 321]

These scoring statistics provide good evidence for the retrosynthesis route scoring module to curtail the application of rule RMC.1.1.2.1 due to very poor example support and rule RZC.1.1.2.2 due to very poor stereoselectivity statistics. Within this rule set there are good complementary solutions for solving both 1,2-*syn* and 1,2-*anti* selectivity. The level of *% de* value support for calculating the stereoselectivity statistics is good on this occasion.

Enolate Addition to α-Heteroatom Substituted Aldehydes

The reaction example count, yield and stereoselectivity statistics for the enolate addition to α-hetero aldehydes are presented in Table 55.

Rules A.2.3.1.1.6 and A.2.3.1.1.7 cover the formation of 3,4-syn and 3,4-anti aldol adducts from unsubstituted enolate (or equivalent) precursors. Both rules are well supported by example reactions, although predicted yields are generally modest. The example sets show good diastereoselectivity statistics (3,4-*syn*, median de: 94%; spread: 12% from 15 examples and 3,4-*anti*, median de: 92%; spread: 11% from 20 examples). A manual survey of the reaction reagents, catalysts and conditions show that the 2,3-*syn* adduct examples are dominated by enzymatic and organo-catalytic (proline based) methods whereas the 3,4-*anti* adduct examples are dominated by the methods utilising boron and lithium enolates.[a]

The four rules A.2.1.2.1.12, A.2.1.2.1.13, A.2.1.1.1.14 and A.2.1.1.1.15 select example reactions that cover all possible diastereomeric outcomes of reacting substituted enolate (or equivalent) precursors with α-hetero aldehydes (Table 55: *vide supra*). The observed example reputation for each diastereomer is generally good and all represent viable retrosynthetic disconnections. Notably the stereoisomers that exhibit the 3,4-*anti* configuration have twice as much support over the 3,4-*syn* stereoisomers. This observation may *cautiously* lend some support to Evans' claim that polar effects dominate stereoselection in reactions that proceed through a Zimmerman-Traxler transition state. [315]

The stereoselectivity statistics for the rules are problematic in terms of providing good reliability indicators (% *de* spread up to 50%). The exception is the 2,3-*anti*-3,4-*anti* stereoisomer which has excellent diastereoselective statistics with an unusually high reliability indicator (rule A.2.1.2.1.13, median de: 96%; spread: 3% from 11 examples). This result would be commensurate with the predicted high selectivity over the competing 2,3-*anti*-3,4-syn stereoisomer *assuming* control is exclusively *via* the *E*-enolate Zimmerman-Traxler/PFA model (see Figure 123: *vide supra*). However a manual survey revealed that only 12 of the 31 examples were reacted via boron or metal enolates while 9 were reacted using organocatalysts or enzymes (the remainder could not be classified) leaving it unclear why the stereoselective statistics are outstanding when compared to the remaining members of the set. In contrast the other stereo configuration rules have reasonable diastereoselectivity estimates but *extremely* poor reliability indicators.

A survey was conducted over the 120 examples covered by the rule set to evaluate and comment on the diversity of stereo control strategies employed to achieve each stereo configuration. A small pertinent sample of the matched reaction examples is presented in Table 56 (*vide supra*).

Diastereomer pairs A and D are formed via the use of double stereo differentiation using antipodal chiral boron *E*-enolates reacted separately with the same enantiopure α-alkoxyaldehyde to selectively form either the mismatched 2,3-*anti*-3,4-*syn* adduct (A) or the matched 2,3-*anti*-3,4-*anti* adduct (D), with both exhibiting high *% de* and *% ee*.[319] As expected,

[a] See the "notes" column in Table 55 for the count of each class of Lewis acid.

the combined Zimmerman-Traxler/PFA model predicts the observed outcomes. These two examples demonstrate the use of the double stereo differentiation strategy, *via* chiral reagents, to selectively control the formation of either 3,4-*syn* or 3,4-*anti* adducts.

The formation of diastereomeric compounds B and H exploit double chelation and chiral auxiliary tactics, in combination with the double stereo differentiation strategy, to select either 2,3-*anti* or 2,3-*syn* adducts with high *% de* and *% ee*. An (O,N) cyclic chelated titanium *Z*-enolate pre-formed from the enantiopure cis-1-tosylamido-2-indanyl-chloroacetate ester is reacted separately with either of a pair of antipodal bidentate chiral α-alkoxyaldehydes, pre-activated as cyclic titanium chelates. This generates the mismatched 2,3-*anti*-3,4-*syn* adduct (B) *via* the (*S*)-alkoxyaldehyde or the matched 2,3-*syn*-3,4-*syn* adduct (H) *via* the (*R*)-alkoxyaldehyde.[322] The use of the Cram-chelation model correctly determines the 2,3-*anti* or 2,3-*syn* outcomes as the *pre-chelated* reactants are unable to participate in a regular Zimmerman-Traxler arrangement.

Finally diastereomeric compounds C and F are generated by exploiting the different stereoisomer preferences exhibited by either E boron enolates reagents,[323] or D-proline organo-catalysis,[324] due to the differing reaction mechanisms and transition states of these reactions.

Conclusions

A methodology has been described for constructing stereoselective retrosynthesis transforms as sets of atom-atom mapped reaction drawings describing substructures of the desired reaction conversion in varying levels of environmental detail around the reacting centre. A key feature is the ability to precisely specify the stereochemical modifications undergone in the reaction such that enantioselective and diastereoselective transforms can be accurately described. These drawings are organised into a hierarchy to aid incremental rule development. The environmental details around the reaction core are described using a combination of non-reacting substituent substructures and constraint annotations written in an enhanced version of the PATRAN language.

The general approach is to begin with the minimum retron substructure and add levels to the hierarchy that separate examples by functionality type (EWG, EDG), stereochemistry, substitution pattern, followed by reaction specific features such as atom hybridisation, substituent aromaticity, specific functional groups, and inter/intramolecular variants amongst many others.

The use of a large reaction database such as CIRX has proven largely effective in mapping out the known stereochemical scope of each transformation. Each transformation is scored using a set of ratings based on example count (reputation), yield, and stereoselectivity. These quantities are treated for the robust removal of outliers by selecting the median value as the representative scoring parameter with the interquartile range used as a measure of the reliability of the parameter. The scoring parameters are further reduced by assigning them to empirically designed nominal values based on 7 point Likert scales to denote ratings ranging from "excellent" to "very poor" making them easy for chemists to compare alternative routes.

In general the transform scoring approach assumes that the consulted reaction databases have a sufficiently comprehensive abstraction policy trending towards maximum coverage of the literature to improve knowledge of each reaction's scope. A small number of sampling tests were performed to validate aspects of this assumption. The CIRX database appears to abstract most reactions from each selected paper, including the reactions with reported 0% yields. A cautionary note is that only one actively maintained database with the necessary stereoselectivity data was made available for this research project. There are very few database choices available beyond the CIRX database. Unfortunately the larger and more comprehensive Reaxys and CAS Scifinder databases do not carry the required enantioselectivity and diastereoselective data.[b]

Limitations to the approach include the general lack of reported diastereoselectivity data in the CIRX database as only 3% of example reactions report a *% de* value. Manual sampling of the reaction examples selected by the diastereoselective transforms established that this absence is largely due to a common practice of not reporting *% de* values in the primary literature. This contrasts with the more reliable reporting of *% ee* values for enantioselective reactions. It is presumed that the physical ease of separation and disposal of unwanted diastereomers is behind the absence of the *% de* data.

The current inability to compute the comparative sizes of substituents attached to a stereogenic centre limits the scope for designing effective diastereoselective rules via R_S, R_M, and R_L query atoms. Rules limited to stereogenic centres with attached hydrogen, heteroatom and carbon substituents were demonstrated and proved effective.

[b] Reaxys records that % *ee* values are available in the source paper but does not abstract the values into the database.

The design approach used for developing enantioselective transforms and diastereoselective transform was demonstrated in detail for selected reaction types. The selection of supporting example reactions was performed against the combined CIRX and REFLIB databases. Published stereoselectivity models for each transform class were reviewed and the matched reaction example statistics compared to the model expectations. The reputation and selectivity statistics in general supported many of the model predictions. However, attempts to correlate models and selectivity statistics are limited when multiple reaction methods and mechanisms are available for the overall transformation (e.g. the aldol reaction). The details of the reaction conditions and the nature of reactive intermediates are not accounted for by the rules as there is no mechanism to filter examples beyond substructure matching. For example, the use of transmetalation of Grignard reagents with transition-metals to modulate reactivity and/or alter reaction mechanism is transparent to the rules, and consequently may distort attempts to correlate selectivity statistics to a mechanism model.

The transform rule methodology is not confined to stereoselective chemistry. Transforms containing hierarchies of rules for describing *stereospecific* reactions are also amenable to the method. For example, reactions undergoing the S_N2 mechanism or stereospecific vinylic coupling are readily handled directly in the stereochemical reaction drawing notation. In these cases *% ee* and *% de* ratings are largely irrelevant and are ignored.

The current set of transforms developed by the method is listed in Table 57. Experience has shown that a transform consisting of around 50 rules can be constructed within a day if good

Alkylation of aldehydes	Alkynylation of aldehydes	Aldol reaction
Alkylation of ketones	Alkynylation of ketones	Mukiyama aldol
Aziridation of alkenes	Alkene reduction	Henry reaction
Epoxidation of alkenes	Imine reduction	Mannich reaction
Dihydroxylation of alkenes	Ketone reduction	Bayliss-Hillman
Metal conjugate addition	Heck coupling (vinyl)	Aza-Bayliss-Hillman
Boronate conjugate addition	Hiyama coupling (vinyl)	Rauhaut-Currier
Carbanion conjugate addition	Kumada coupling (vinyl)	Stetter reaction
Carbenoid cyclopropanation	Negishi coupling (vinyl)	Diels-Alder
Epoxide opening – C nucleophiles	Stille coupling (vinyl)	Claisen rearrangement
Epoxide opening – N nucleophiles	Suzuki coupling (vinyl)	Cope rearrangement
Epoxide opening – O nucleophiles	Sulphide/sulphoxide oxidation	
Epoxide opening – S nucleophiles	Mitsunobu reaction	

Table 57 Implemented stereoselective and stereospecific transforms.

review papers are at hand to help plan the structure of the constraints hierarchy. The addition of a rule cloning tool greatly speeded up the process of building hierarchies as each new rule only requires minor modification to a cloned parent rule. There are further opportunities to increase the efficiency of the rule building process. For example, enumerating all possible substitution patterns around stereogenic centres can be a tedious process when done manually and should be automated.

Chapter 6

One-Step Stereoselective Retrosynthesis

"In solving a problem of this sort, the grand thing is to be able to reason backward"

- Sherlock Holmes in "A Study in Scarlet".

Introduction

This chapter describes the application of stereoselective transforms in one-step retrosynthetic analyses of some selected target molecules. A number of programs were created to build the retrosynthesiser application. First, an off-line program that automatically constructs rule hierarchies from individual rules is described. Each assembled rule hierarchy represents a reaction transform and is stored in a database for access by the main retrosynthesis module. Second, the main retrosynthesis program is composed of four main modules: a procedure that identifies all matching retrons in the target molecule using the stored transform hierarchies; a 'transformer' that constructs the precursor molecules by applying the matched transform rules to the target molecule; a simple scoring module that collects rule and precursor derived parameters that are used to rank alternative retrosynthetic paths; and a problem assessment module that reverses the application of the transform on the generated precursor to identify potential regioselectivity issues. The chapter concludes with a discussion of future directions with particular focus on adapting the work described in this thesis to the application stereochemical strategies to score and possibly direct a deeper stereoselective retrosynthetic plan.

Rule Hierarchy Construction

In the previous chapter the formulation of a transform as a hierarchy of rules was described in conceptual terms as a means to subdivide sets of rules that progressively describe more detailed stereo controlling environments. The operation of the retrosynthesiser described in this chapter relies on a concrete representation of the hierarchy to guide the retron search (*vide infra*). This required an algorithm to be devised to calculate and store the rule ancestor and descendent relationships.

The hierarchy building algorithm is divided into two phases. The first phase supports the interactive addition and removal of drawn rules via the retrosynthesis workbench application (*vide supra*) by updating and storing pairs of rule identifiers in an indexed ancestor-descendant table. The second phase is executed when the retrosynthesiser requests the set of immediate daughter rules for a given rule. This two phase mechanism ensures that rule addition, modification and removal and the extraction of immediate daughter rules are efficient operations requiring a simple indexed search for those identifier pairs that refer to the rule of interest (Figure 125).

The first phase is executed during rule insertion or rule modification and proceeds in the following manner. The characteristic reaction core (CRC) code for the new rule is calculated from its reaction hyperstructure and all existing rules with identical CRC codes are retrieved[iii]. Each retrieved rule is matched to the new rule using a CSP solver to determine if the rule hyperstructures are related in the following ways: as a descendent of the new rule (*via* subgraph isomorphism); as an ancestor of the new rule (*via* supergraph isomorphism); or as neither a descendent nor an ancestor. To accomplish this two CSP solvers are configured for subgraph isomorphism. One CSP solver matches the new rule to the existing rule to find substructure solutions and the other CSP solver reverses this arrangement to find superstructure solutions.

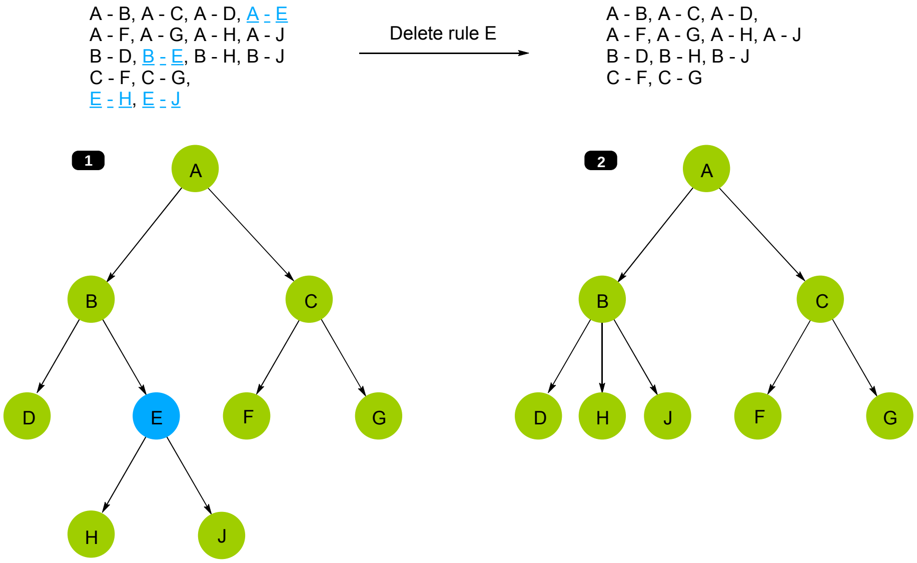


Figure 125 The representation of rule hierarchies *via* an indexed ancestor - descendant table. Removing rule E from rule hierarchy (1) generates a new rule hierarchy (2) after deleting all ancestors - descendant pairs containing a reference to rule E.

[iii] All rules with an identical CRC code belong to a single transform represented by a tree of such rules.

The special case where a rule solves as both an ancestor and a descendant of each another is flagged as an error and the new rule is rejected as a duplicate of an existing rule. A PATRAN constraints function is registered with the CSP solver to determine if a PATRAN expression associated with an atom or bond is equal to, or a subset of, the compared expression. The additional constraints function is required so that rule graphs can be directly compared.

If the pair of rules do not have a lineal relationship, a third CSP solver is configured for solving graph isomorphism to determine if the rule retrons are identical. This is accomplished by comparing the rule product graphs which acts as the rule retron. A pair of distinct rules may have identical retrons if the retrosynthetic target can be derived from different precursor stereoisomers (*e.g.* from *cis*- or *trans*-alkene precursors). All rules with common retrons are used by the retrosynthesiser to consider each route to assess which precursor stereoisomer ranks best based on example reputation and target molecule scoring information.

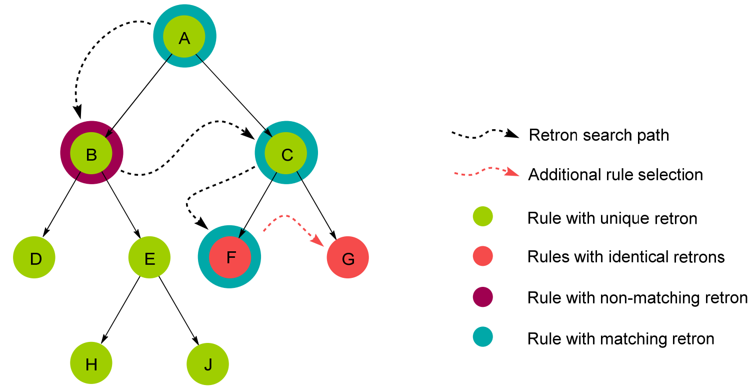

A
B
C
D
E
F
G
H
J
Retron search path
Additional rule selection
Rule with unique retron
Rules with identical retrons
Rule with non-matching retron
Rule with matching retron

Figure 126 The rule to target retron search procedure. The search starts at root rule A, descending to the next level each time a retron match is made or traversing to the right if a non-match. The search is successful when the matched terminal rule F is found. Rule G is also selected as it shares a common retron with F.

The rule ancestor and descendant relationships and retron equality relationships are recorded in the database on completion of the rule comparison. Two additional pieces of information about each rule relationship are recorded. First, the mapping of the hyperatom[jjj] pairs between two the rules is extracted from the CSP solver solution and expanded to include all symmetry equivalent atoms and the map is stored in the database. This mapping information is used by the retrosynthesiser for optimising the search for matching rule retrons (*vide infra*). Second, those transform rules with no ancestors are marked as root rules and those with no descendants are marked as terminal rules. The retrosynthesiser begins all retron searches from transform root

[jjj] Hyperatoms are the graph nodes in the reaction hyperstructure graph.

rules and is successful if it can traverse the hierarchy and match a terminal rule to the target molecule (Figure 126).

The second phase of rule hierarchy construction is daughter rule retrieval which proceeds in the following manner. An indexed lookup is performed to retrieve all ancestor-descendant pairs where the rule of interest is the ancestor rule. The corresponding descendant rules identifiers are inserted into a daughter rule set which is immediately scanned and an indexed lookup of each referenced rule as the ancestor rule is performed in same manner as for the original rule. The retrieved descendant rules from each of these subsequent searches are removed from the daughter rule set. After completion of the scan the set contains only the immediate daughter rules. The inverse of this procedure gathers the immediate parents of any given rule.

A One-Step Retrosynthesiser

An experimental one step retrosynthesiser module was created to exhaustively apply all stored transforms to a target molecule to generate sets of plausible precursors. These precursors are scored by reputation, estimated yield and estimated stereoselectivity. In addition functional group tolerance is assessed based on evidence gathered from the corresponding example reactions. Additional perception of features such as *meso* or C_2 symmetry of the generated precursor molecule and the stereotopic properties of the precursor reaction site are made available for scoring the route. Structure simplifying parameters based on the types of bond and stereocenter alterations are also calculated (*vide infra*) to support route scoring. The user can choose which parameters take priority. For example, priority can be given to retrosynthetic disconnections that avoid regioselectivity and diastereoselectivity problems or those that simplify precursors by maximising stereocentre clearance and/or bond cleavage in addition to choosing routes with best example reputation or predicted stereoselectivity.

The user interface of the one-step retrosynthesiser is shown in Figure 127. The target molecule is either drawn using the Marvin Sketch tool or is loaded from a disk file. The retrosynthetic analysis is then initiated by a button press and the results are displayed visually as an inverted retrosynthetic tree showing the precursor molecules and details of the applied rules. These details include yield and enantiomeric excess estimates with associated reliability values and the rule reputation as the number of examples. The rule information button displays an additional page detailing: tolerated and reacted functional groups; structure simplification metrics such as the numbers and types of removed, modified or transposed stereocentres; and the number of made, broken or modified bonds. Internally the retrosynthesis tree is represented as a bipartite directed graph composed of alternating molecule and applied rule nodes. [325]

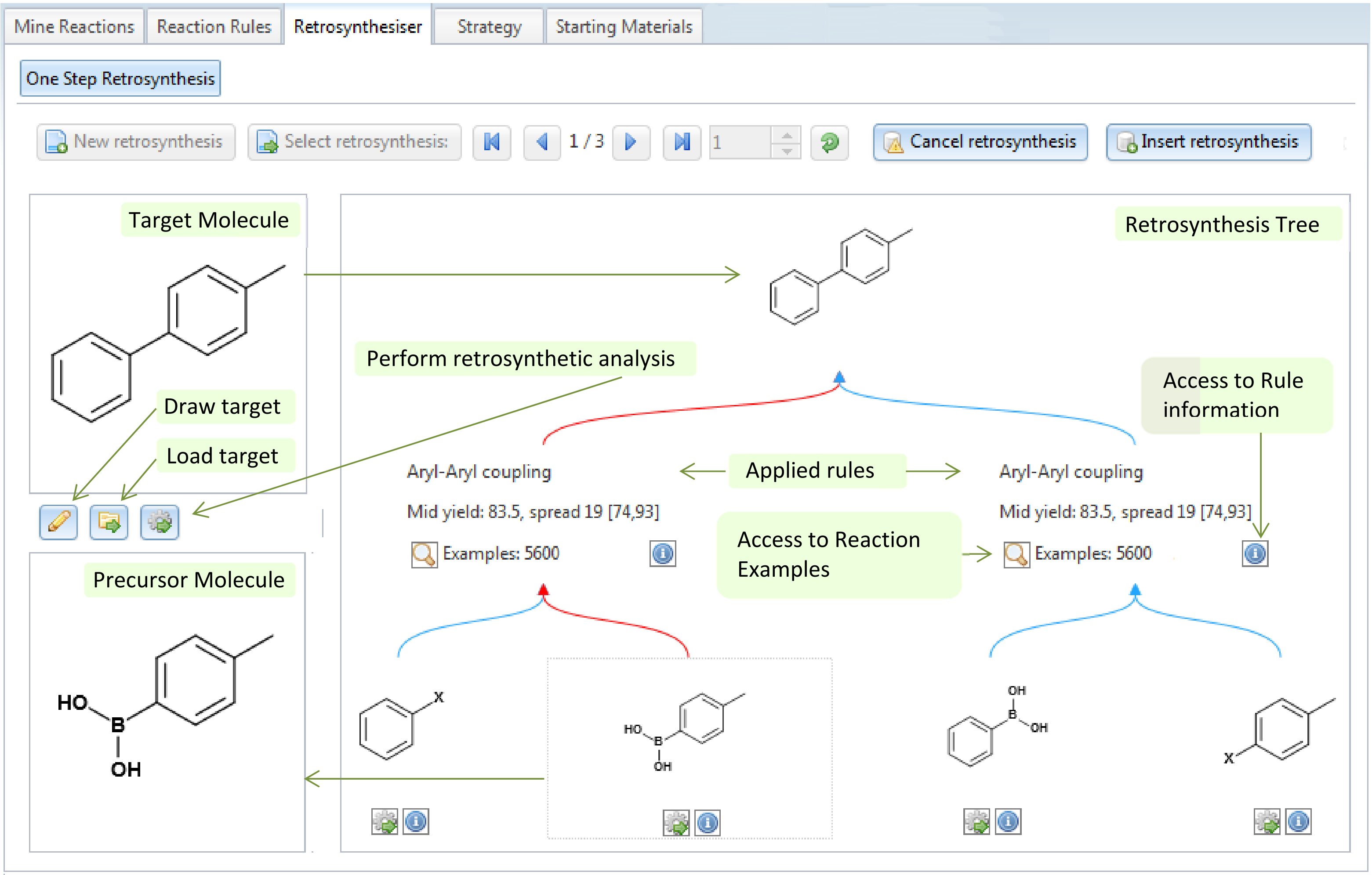


Figure 127 The one-step retrosynthesiser user interface.

Rule matching

Rules are matched to target molecules in order to generate precursor molecules. The transform matching process is conducted in the following manner. A CSP solver is configured for solving subgraph isomorphism with isomorph free matching enabled to prevent the generation of symmetrical (duplicate) solutions. The rule product graph (the retron) is used as the CSP query and the target molecule serves as the CSP target. The CSP solver is set to solve for both stereo exact solutions and enantiomeric solutions. This allows each stereoselective rule to match either enantiomer of a target molecule without requiring both stereochemical forms of each rule to be explicitly represented in the transform rule hierarchy. To accommodate this rule optimisation, the CSP solution records the stereo match parity so that precursor generation is able to generate and draw the correct stereoisomer from the application of a single form of the stereo rule.

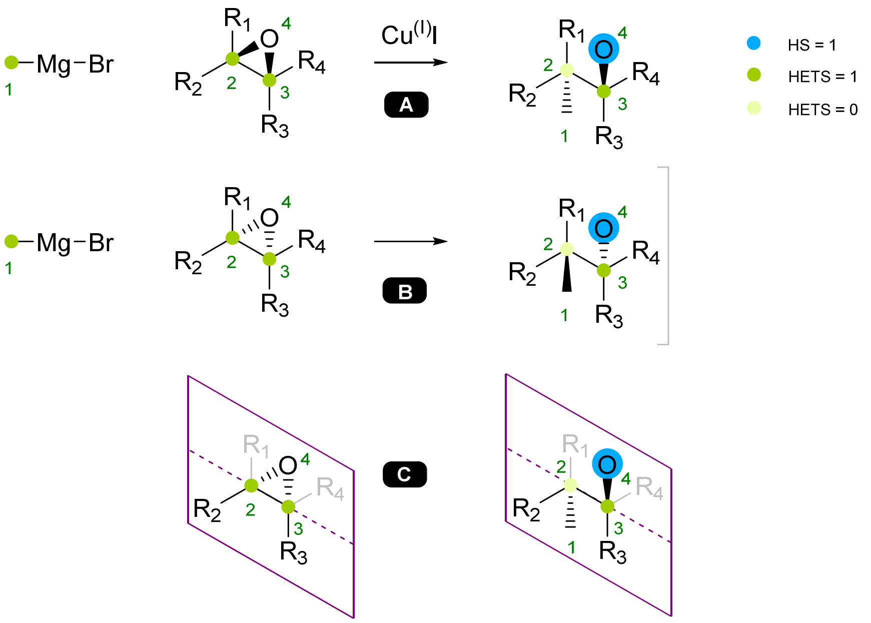


Figure 128 A transform rule uses both the exact form of the rule (A) and its mirror image (B) to discover a matching stereochemical retron in a target molecule. The exception is when the rule retron structure and it precursor dual (C) contain corresponding reflection planes, in which case only the exact form of the rule is used. This prevents the generation of pairs of duplicate retrosynthetic routes.

If the rule retron and its precursor contain corresponding planes of symmetry, the CSP solver is configured only for exact stereochemical solutions to prevent pairs of duplicate precursor routes (Figure 128). The existence of the dual reflection plane is obtained by symmetry perception performed directly on the rule hyperstructure graph (*vide supra*).

All root rules of all transforms are exhaustively matched to the target molecule to locate all non-duplicate retron locations using the precursor generating procedure shown in Figure 129 (Procedure A). When a minimum retron match is located a more detailed search of the rule hierarchy is made to find the maximum (best) retron match using the recursive retron matching procedure shown in Figure 130 (Procedure B). This recursive procedure is executed in the

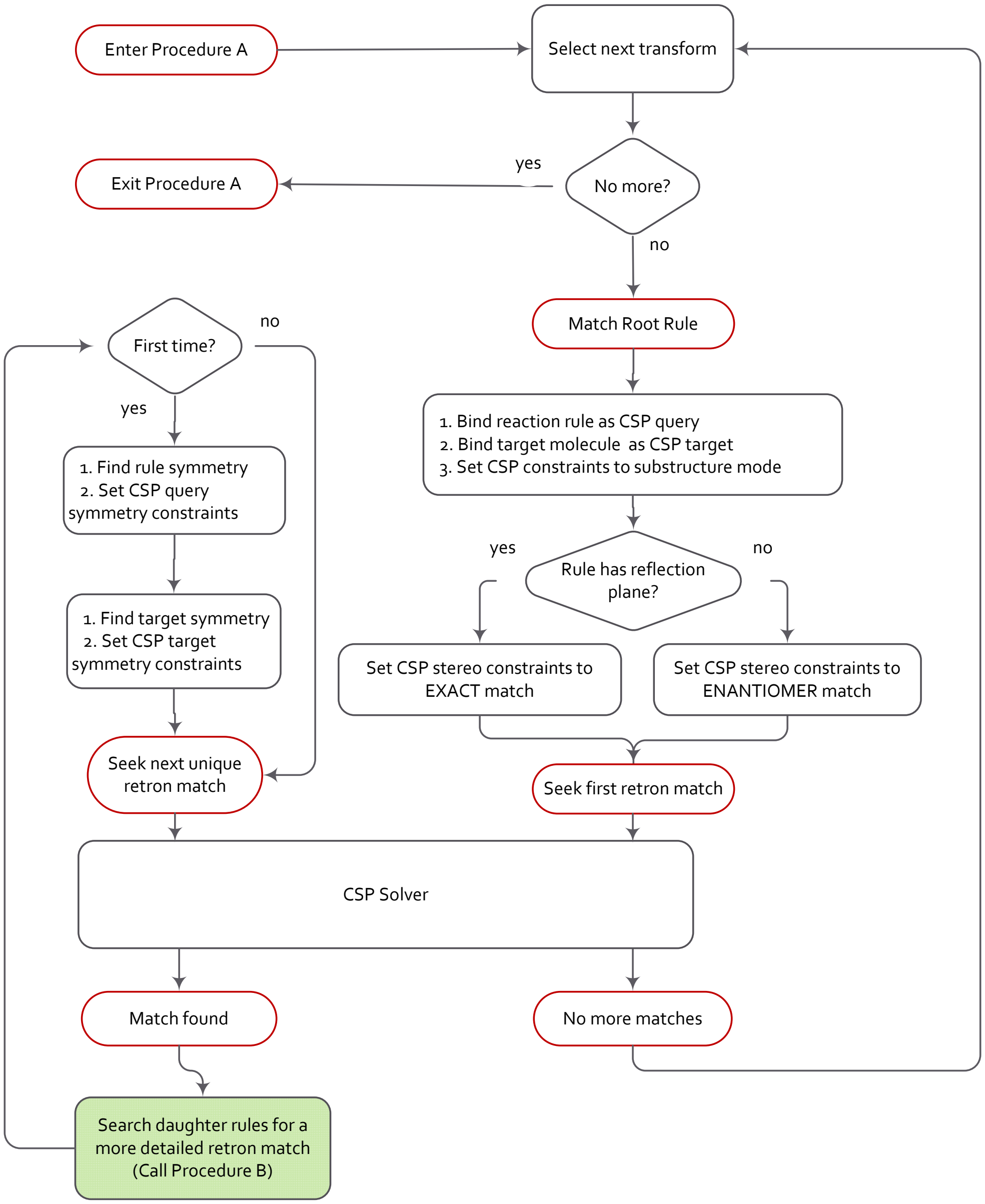

Enter Procedure A
Select next transform
yes
Exit Procedure A
No more?
no
Match Root Rule
First time?
no
yes
1. Find rule symmetry
2. Set CSP query symmetry constraints
1. Bind reaction rule as CSP query
2. Bind target molecule as CSP target
3. Set CSP constraints to substructure mode
yes
Rule has reflection plane?
no
1. Find target symmetry
2. Set CSP target symmetry constraints
Set CSP stereo constraints to EXACT match
Set CSP stereo constraints to ENANTIOMER match
Seek next unique retron match
Seek first retron match
CSP Solver
Match found
No more matches
Search daughter rules for a more detailed retron match (Call Procedure B)

Figure 129 The precursor generation flow chart for Procedure A: the outer loop exhaustively matches root retrons to the target molecule to find all unique retron locations.

following manner. Each daughter rule of the matched root rule is tested against the target in turn until a match is made. To make this efficient the result of the root rule match and the stored *root-rule-to-daughter-rule* atom map is used to pre-calculate the mapping of the retron site to the daughter rule. This is communicated to the CSP solver by editing the atom candidate constraints to ensure the daughter rule is matched onto the *same* retron location. In this

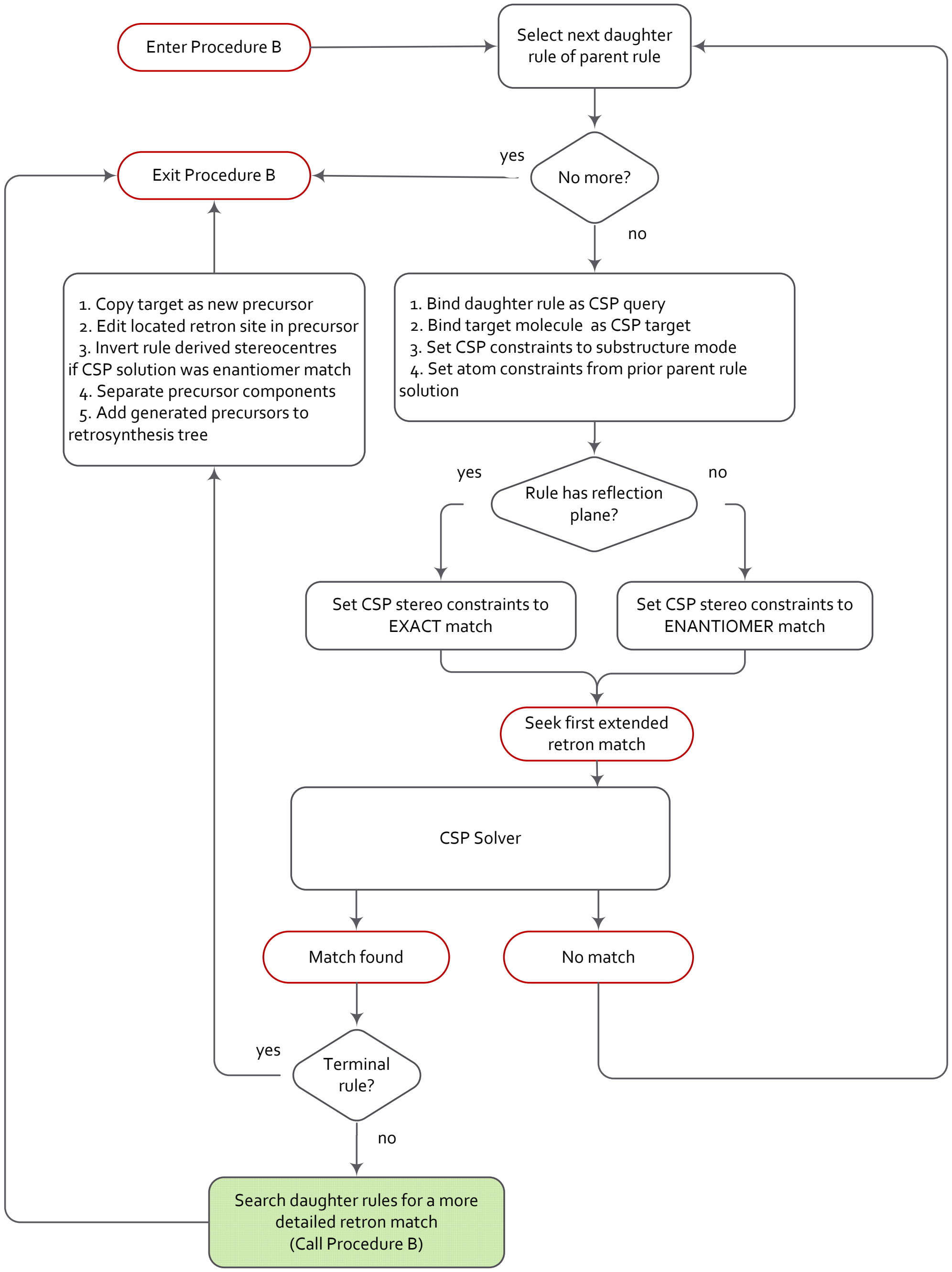

Enter Procedure B
Select next daughter rule of parent rule
Exit Procedure B
yes
No more?
no
1. Copy target as new precursor
2. Edit located retron site in precursor
3. Invert rule derived stereocentres if CSP solution was enantiomer match
4. Separate precursor components
5. Add generated precursors to retrosynthesis tree
1. Bind daughter rule as CSP query
2. Bind target molecule as CSP target
3. Set CSP constraints to substructure mode
4. Set atom constraints from prior parent rule solution
yes
Rule has reflection plane?
no
Set CSP stereo constraints to EXACT match
Set CSP stereo constraints to ENANTIOMER match
Seek first extended retron match
CSP Solver
Match found
No match
yes
Terminal rule?
no
Search daughter rules for a more detailed retron match
(Call Procedure B)

Figure 130 The precursor generation flow chart for Procedure B: the inner recursive search for the best retron match at a target molecule site originally found by the root retron.

manner each daughter rule is able to efficiently test additional constraints and extend the retron site (if needed) without any unnecessary backtrack searching.

If a daughter rule is matched, the above procedure is invoked recursively with the daughter rule taking the place of the root rule until either the search fails to find another match or a *terminal* matching rule is found. The rule walk executed by Procedure A with Procedure B is summarised in Figure 126 (*vide supra*).

When a terminal rule is located then all rules that are known to share the same retron are also retrieved and these are used to generate alternative precursors which are added to the retrosynthesis tree. The details of precursor generation are described in the next section.

If no matching daughter rule is found then the application of the current transformation to the current retron site is abandoned. After the above recursive processing is complete the search re-enters Procedure A to find the next matching site for the current root rule or, if no further matches are found, processing moves to consider the next root rule. This completes the transform matching phase.

Precursor Generation

The generation of stereochemically correct precursors is performed by executing structure edit commands derived directly from the matching transform rule. These edit commands are pre-computed when the rule is added to the database by finding the differences between the rule retron structure and the rule precursor structure. The edit commands are most conveniently derived directly from the rule hyperstructure representation. The generated retrosynthetic edit command list is applied to matched target structures in the following order:

1. *Atom creation commands*: created atoms are those that exist in the rule precursor structure but not in the rule retron structure. They usually correspond to leaving group atoms.
2. *Bond creation commands*: created bond correspond to those that exist in the precursor structure but not in the rule retron structure. They are the bonds found within leaving groups and the bonds that join the leaving group to the main precursor structure. Bonds involved in rearrangement reactions that are broken in the synthetic direction also appear in this list.
3. *Atom modification commands*: these correspond to changes in charge (or radical) values between the retron and precursor structure.
4. *Bond modification commands*: these correspond to changes in bond order between the retron and precursor structure.

5. *Bond removal commands*: these correspond to bonds that exist in the rule retron structure but not the rule precursor structure. These are usually bonds within moieties that are introduced by reagents in the synthetic direction. Bonds involved in rearrangement or intramolecular cyclisation reactions that are formed in the synthetic direction also appear in this list.
6. *Atom removal commands*: These are usually atoms within moieties that are introduced by reagents applied in the synthetic direction.

Following the application of the structure edit commands the next stage involves rearranging the positions of the atoms to match the layout of the rule precursor structure, using the latter as a template. Scaling, rotation and translation transforms are calculated and applied in stages to each precursor atom to place the atom into a stereochemically correct position according to the rule precursor template. This is performed in the following steps:

1. A rule bond is chosen as a reference bond. A reasonable heuristic is to choose one of the central bonds in the precursor pattern. A scaling and rotation transform is computed that aligns the rule precursor to match the orientation of the corresponding bond in the newly generated precursor.
2. All substituents attached to the atoms at the end of the reference bond in the new precursor molecule are realigned (rotated and translated) to match the angles of the R_n appendages in the transformed rule precursor template. This action ensures that the stereochemistry represented in the new precursor will correspond to the stereochemistry dictated by the rule.
3. If an open valence in the rule was not represented by an R_n substituent, then standard angles are chosen for the realigned precursor substituents based on bond order and number of attachments (e.g. 120 degrees is chosen for sp^3 or sp^2 atoms with 2 or 3 drawn attachments).[kkk]
4. All newly introduced substituents originating from the rule precursor pattern are scaled and orientated using the transform generated in step 1. This ensures that the new substituent bonds lengths are scaled to match the bonds of the target molecule.
5. The final alignment step reorients and scales the whole precursor molecule to align the largest substituent to the corresponding substituent in the target molecule. This

[kkk] The bond angles at open valence sites are not critical as these should never appear at stereogenic centres whereas open attachment sites at stereogenic centres are always represented with R_n atoms to retain the correct stereochemical identity.

minimises the number of atoms reoriented between target and precursor and provides a more aesthetically pleasing result. At this stage the layout of the precursor molecule is complete.

6. All stereodefined bonds (i.e. those drawn as wedged or hashed) that were within the retron site are cleared to standard bonds types.
7. The corresponding stereo bond symbols in the rule precursor template are applied to the corresponding bonds in the generated precursor molecule under the following conditions: if the retron matched with exact stereo parity then the bond symbols are copied *as-is* from the rule precursor template; if the retron matched with inverted stereo parity (an enantiomer match), then the inverse bond symbols are copied (*i.e.* a

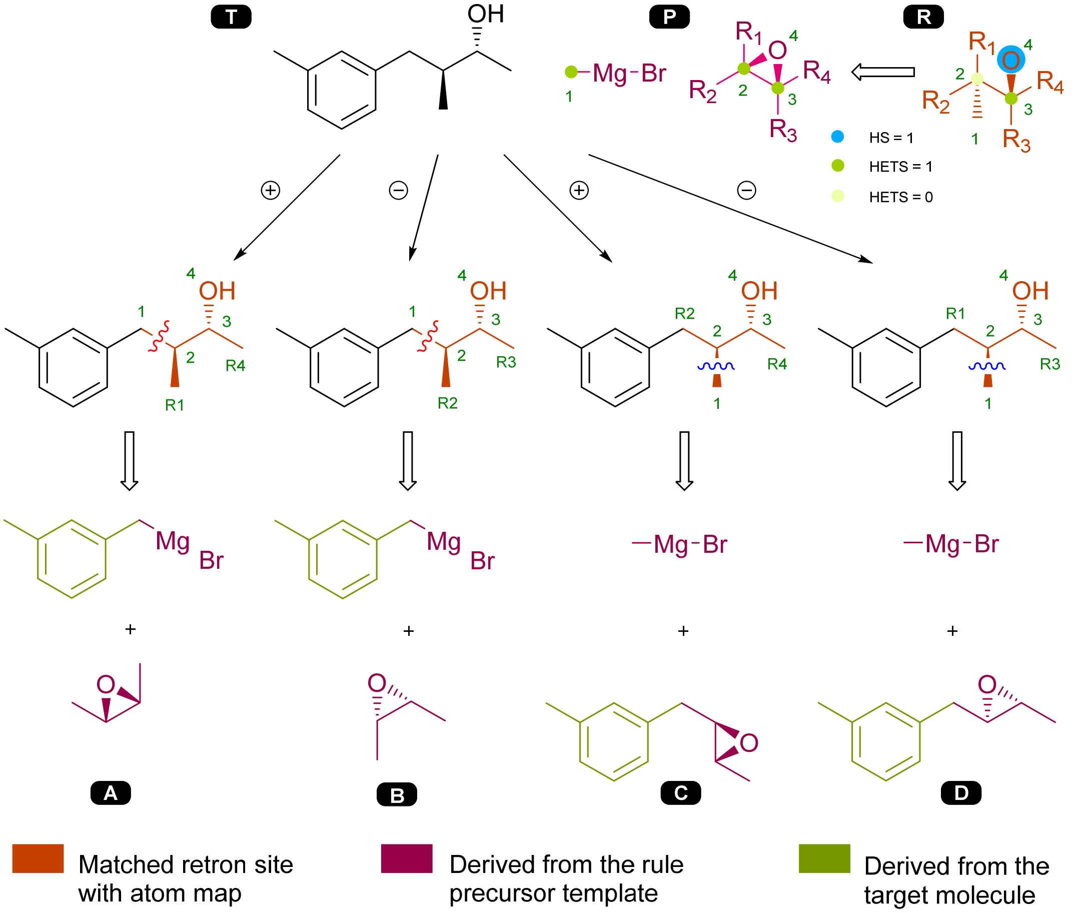
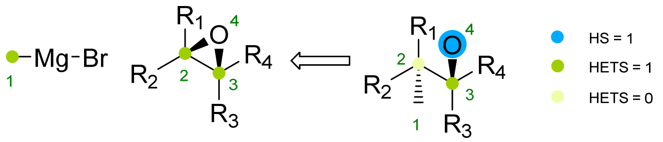

T
OH
P
Mg-Br
1
R1
O
4
R4
R2
2
3
R3
R
HS = 1
HETS = 1
HETS = 0
⊕
⊖
⊕
⊖
A
B
C
D
Matched retron site with atom map
Derived from the rule precursor template
Derived from the target molecule

Figure 131 **The use of the rule precursor pattern P as a template to generate alternative sets of precursors from target T. The stereochemistry of the precursors is derived directly from the stereochemistry of the rule precursor structure in combination with the stereochemical parity of the retron. A ⊕ parity match retains the template stereochemistry while a ⊖ parity match inverts the template stereochemistry. The alternative mappings of the retron to target T generate the pathways A to D.**

wedge for a hash and a hash for a wedge).

The outcome of the application of these precursor generation steps is shown in Figure 131 using a stereospecific epoxide ring opening rule on the target secondary alcohol T. The generation of precursors B and D illustrate that a stereo inverted retron match (⊖) requires an inversion of the stereo bond symbols copied from the rule precursor template P.[III] The generation of precursors A and C illustrate that an exact stereo match (⊕) retains the stereo bond symbols copied from P.



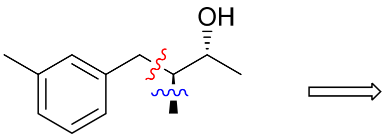
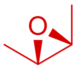
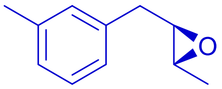
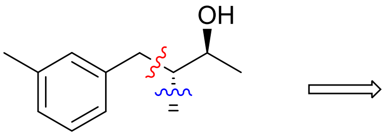
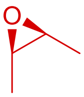
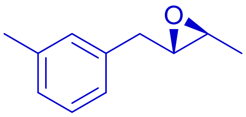
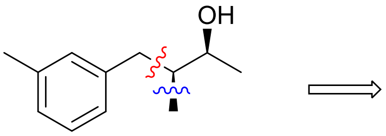
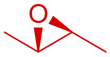
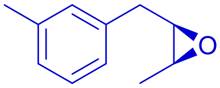
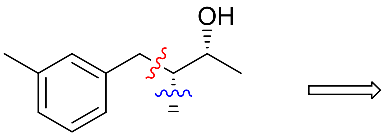
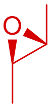
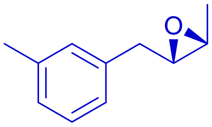
Target alcohol	Epoxide Precursor 1	Epoxide Precursor 2
 T	 TP1	 TP2
 U	 UP1	 UP2
 V	 VP1	 VP2
 W	 WP1	 WP2

Table 58 **The application of a epoxide ring opening transform rule to a set of secondary alcohol stereoisomer targets (T, U, V and W). The targets T and U are enantiomers of each other as are targets V and W. Two alternative disconnections exist for each target. The accompanying Grignard precursors are omitted.**

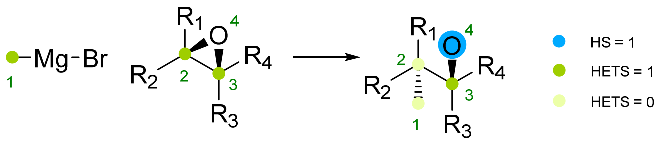
[III] In normal operation the stereo inverted retron matches (marked with ⊖) leading to precursors B and D are supressed as the rule has dual reflective symmetry in patterns R and P. Precursors A and B are consequently identical due to this reflection plane, as are precursors C and D.

Table 58 illustrates the application of a stereospecific epoxide ring opening rule to a full set of enantiopure secondary alcohol stereoisomers targets (T, U, V and W). Each target molecule has two alternative disconnections that lead to alternative epoxide and Grignard precursors. Using the existing symmetry perception routines (*vide supra*), precursors TP1 and UP1 are recognised as *meso* compounds, VP1 and WP1 as C_2 symmetric compounds (and as enantiomers of each other) and the epoxide precursors TP2, UP2, VP2 and WP2 are recognised as non-symmetric molecules.

Recognising Regioselectivity Problems

> *"It's a poor sort of memory that only works backwards"*
>
> – the White Queen in "Alice's adventures in Wonderland & through the looking glass".

Symmetrical functional groups, such as epoxides, present a regioselectivity problem as the proposed retrosynthetic disconnection may be the least reactive synthetic pathway. Multiple occurrences of the same functional group also introduce potential regioselectivity issues. The ability to perceive which pathway is favoured is dependent on a number of factors including



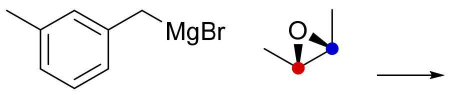
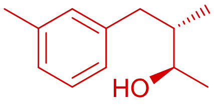
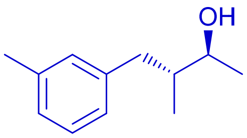
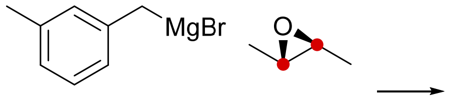
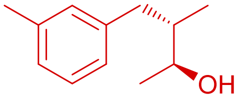
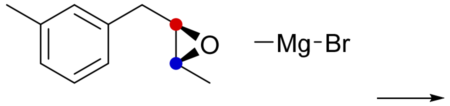
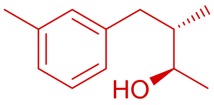
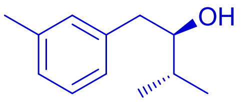
Precursors	Alcohol Product 1	Alcohol Product 2	R
 TP1	 T	 U	×
 VP1	 V		×
 TP2	 T	 X	✓

Table 59 Detection of regioselectivity issues by the reverse application of the precursor generating transform. Column R indicates that regioselectivity issues exist. Nucleophilic attack at colour coded atoms leads to the corresponding colour coded product.

knowledge of the comparative reactivity of completing functional groups, knowledge of reaction conditions leading to the recognition of the reaction mechanism and knowledge of the steric and electronic effects of substituents around the reaction centre. Using the combined effects of these factors to elucidate regioselective outcomes is currently beyond the capabilities of the research software implemented for this study.

A first order approach for the basic recognition of regioselectivity problems was implemented by applying the inverse of the matched transformation to each generated precursor and analysing the resulting (re)generated target molecules. Transform inversion is accomplished by constructing the inverted hyperstructure for each transform rule were all bonds marked as 'made' are relabelled 'broken' (and *vice versa*) and atom and bond modification property values are reversed (e.g. the hyper-bond order 'single-to-double' becomes 'double-to-single').

Table 59 presents the results of applying the inverse of the epoxidation transform used to generate the precursors set out in Table 58. When alternative products are generated these are pairwise compared using a CSP solver configured for full structure stereoisomer matching. The expected outcomes are: an enantiomer match; a diastereomer match; or a mismatch.[mmm] The latter result indicates that a potential regioselectivity problem may exist while the enantiomeric and diastereomeric results indicate potential stereoselectivity problems.

Three epoxides (TP1, VP1 and TP2) were selected from Table 58 to illustrate the outcomes. The *meso* epoxide TP1 generates two alternative *enantiomeric* products indicating that a desymmetrisation strategy is required to reach the desired target molecule.[76, 326, 327] The C_2 symmetric epoxide VP1 generates only the desired target molecule illustrating a key advantage that C_2 molecular symmetry imparts in eliminating both stereoselective and regioselective issues.[328, 329] The non-symmetric epoxide TP2 generates two non-stereoisomeric products confirming that there is a potential regioselectivity issue and this is considered when scoring the route.

Experimental Design

Six simple target molecules (Figure 132) were selected for examining the retrosynthetic disconnection capabilities of the one-step synthesiser application. The target molecules are relatively non-complex and were chosen with two factors in mind: that there is reasonable

[mmm] An exact stereo match is never expected because the (inverse) precursor generator is set for isomorph free generation.

likelihood that some of the proposed reactions could be found in the literature for route validation purposes; and given that the number of available stereoselective transforms is currently small, maximising the types of recognised disconnections. The expectation was that at least three or four alternative stereoselective disconnections would be suggested for each target using the transforms at hand. The targets were divided into three groups in which substituents or stereoisomers are varied to test how the transforms respond with respect to the evidence-based or computed scoring parameters.

No attempt has been made to combine the scoring parameters into a single ranking value and each parameter is presented separately. The expectation is that once integrated into the ARChem program, the scoring parameters will be used to rank the alternative routes based on parameter priorities and ordering preferences set by the user.

A useful ranking priority is given to retrosynthetic disconnections that clear the maximum number of stereocentres while simultaneously cleaving bonds in the target structure.[117] A lower priority is given to transforms that clear a maximum number of stereocentres without simultaneous molecular cleavage. Lowest priority is given to disconnections that clear stereocentres but leave neighbouring stereocentres unmodified as this presents additional diastereoselective complexity in the synthetic direction. A general method to predict which diastereomer is favoured has not been developed so the existence of a diastereotopic centre in the reaction site is used to flag that the required formation of the desired stereo configuration is uncertain. This situation is avoided when the reacting site only has enantiotopic centres. In this case a ranking is chosen using the evidence-based enantioselectivity rating of the transformation.

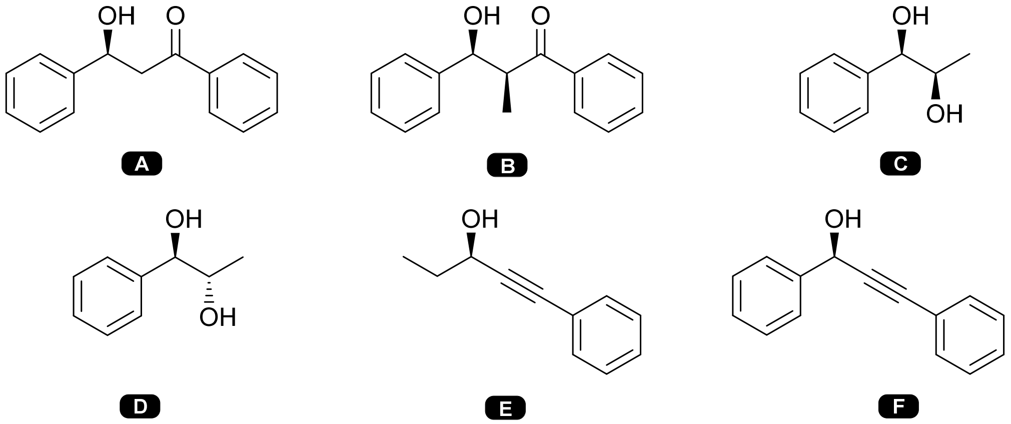


Figure 132 Experimental target molecules consisting of β-hydroxyl ketones (A, B), 1,2 diols (C,D) and propargyl alcohols (E,F).

The ranking approach is illustrated in Figure 133. Disconnection A in β-hydroxy ketones is highest ranked as it removes two adjacent stereocentres and cleaves the interconnecting bond. Disconnection B ranks lower than A as it removes one stereocentre and cleaves one bond but leaves one stereocentre intact, thus introducing a diastereoselectivity problem. Disconnections C and D are equivalent in rank, and rank lower than B as they retain a diastereoselectivity problem but do not cleave the structure. This ranking approach is shown in Figure 133 for 1,2 diols (disconnection E ranks higher than F) and propargyl alcohols (disconnections G and H are equivalent but rank higher than J).

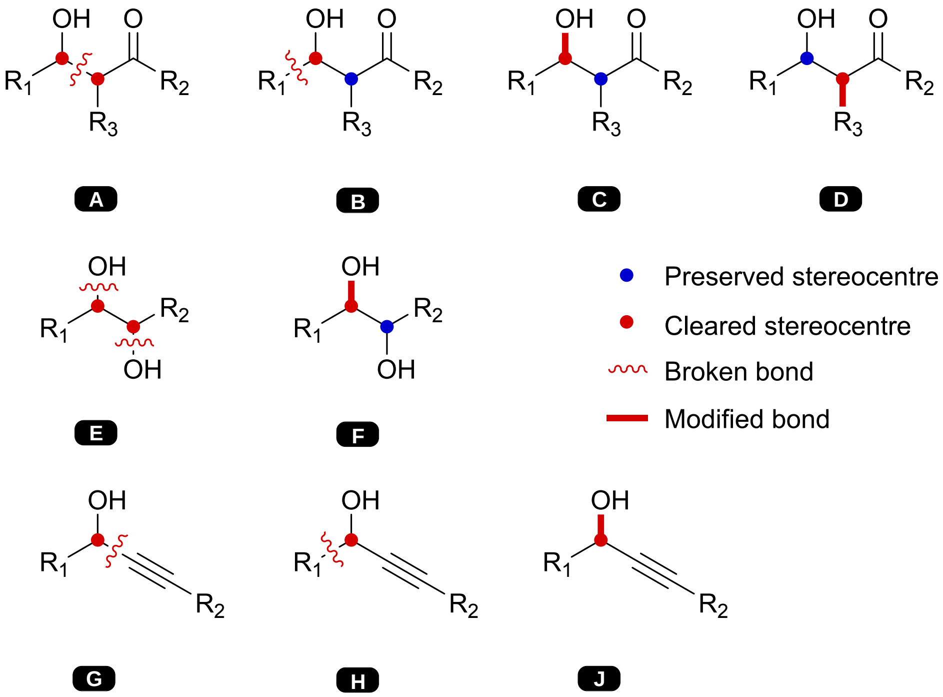


Figure 133 Stereo simplifying retrosynthetic disconnections that simultaneously break bonds and clear stereocentres found in β-hydroxy ketones, 1,2 diols and propargyl alcohols. Disconnections B, C, D and F have additional diastereoselective complexity and are less favourable choices.

The proposed disconnections and precursors generated by the retrosynthetic experiments are compared to exact examples located using the on-line Reaxys database[nnn]. A matched reaction provides positive validation that the transformation is feasible, but the lack of match is not interpreted as evidence that the proposed transformation is infeasible. Indeed the supporting evidence from similar reactions originally used to support the rule indicates a route is probable.

[nnn] Accessed at https://www.reaxys.com/reaxys/session.do

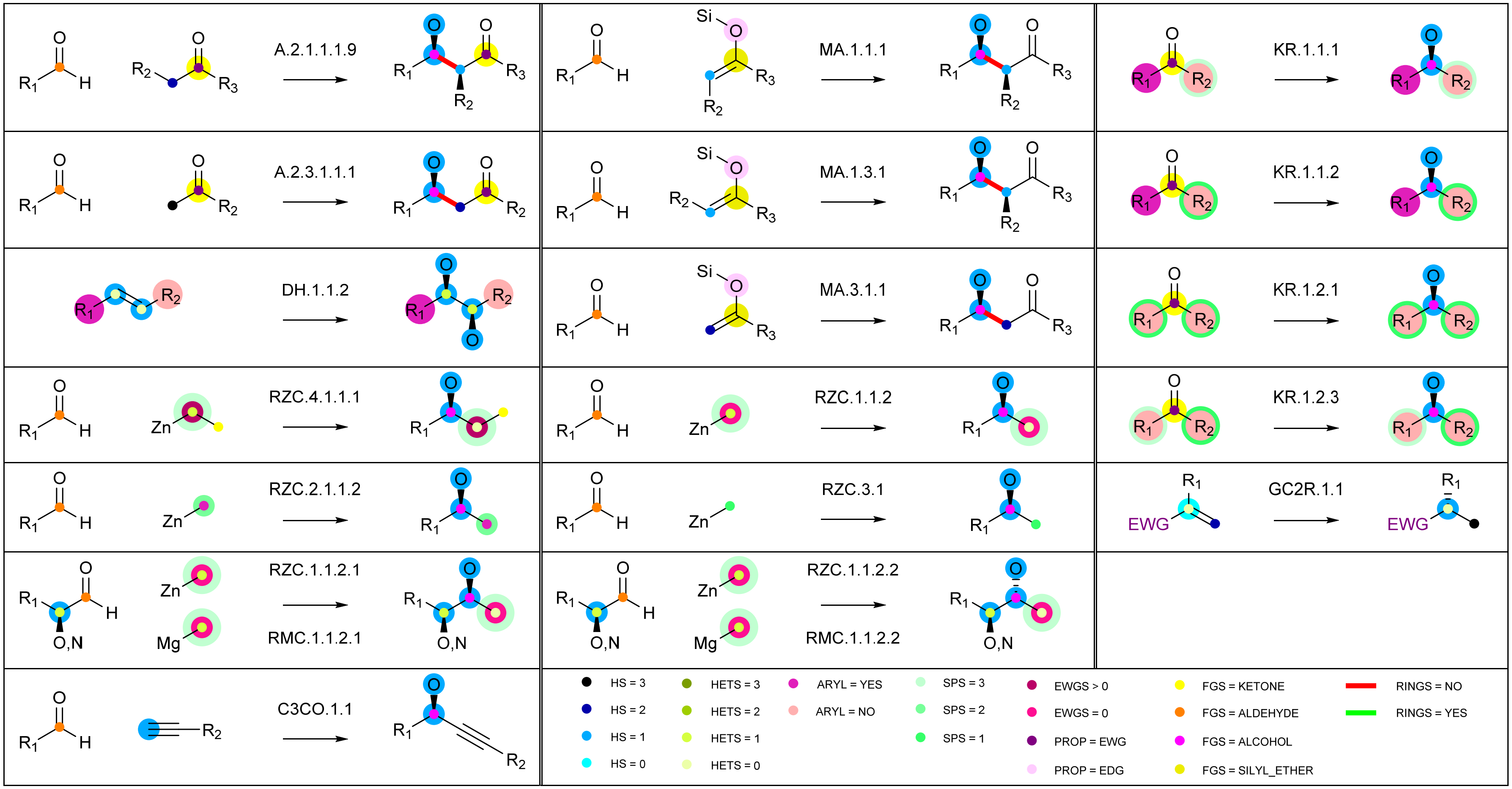


Table 60 Transform rules used to generate the precursors shown in Table 62, Table 63, Table 65, Table 66, Table 68 and Table 69.

Key:

RMC.x	**– alkyl/aryl-Mg addition to carbonyls**	**A.x**	**– direct aldol reaction**	**KR.x**	**– ketone reduction**
RZC.x	**– alkyl/aryl-Zn addition to carbonyls**	**MA.x**	**– Mukaiyama aldol reaction**	**DH.x**	**– dihydroxylation of C=C**
GC2R.x	**– reduction of gem substituted alkenes**	**C3CO.x**	**– alkynyl additions to carbonyls**		

Results and Discussion

Table 60 (*vide supra*) indexes and presents the details of all individual transform rules that were matched to the target molecules used in the experiments. It is presented as a cross reference source to aid in understanding the origin of the proposed retrosynthetic disconnections.

The results of experiments 1 to 6 are shown in Table 62 through to Table 69 (*vide infra*). Each table presents the proposed disconnections of the six selected target molecules (Figure 132: *vide supra*) using all the available transforms in the retrosynthetic analysis.

The measured and calculated scoring parameters presented in these tables are located in the labelled columns as follows:

A. Reaction reputation based on the total number of examples. Presented as a colour coded Likert value (representing one of: superb; excellent, very good; good; fair; poor; very poor; and no support)
B. Reaction reputation based on examples with known enantioselectivity. Presented as a colour coded Likert value.
C. The median yield of the reaction examples. Presented as a colour coded Likert value.
D. The reliability of the median yield. Presented as a colour coded Likert value.
E. The median enantiomeric excess value. Presented as a colour coded Likert value.
F. The reliability of the median enantiomeric excess value. Presented as a colour coded Likert value.
G. The stereo topicity of the reacting sp^2 centre, one of *homotopic*, *enantiotopic*, *diastereotopic*.
H. The number of carbon-carbon bonds broken in the retrosynthetic direction.
I. The number of stereocentres removed in the retrosynthetic direction.
J. Tolerated functional groups as lists of names and counts extracted from the supporting examples.
K. The transform ‘kill’ status based on lack of examples or the presence of interfering functional groups.
L. The presence or absence of regioselectivity issues. Determined by reversing the transform.
M. The proposed transformation matches a known reaction.

Parameters A to F and J are evidence-based parameters derived from matched reaction examples during transform rule indexing (*vide supra*). Parameters G to I are calculated by the reaction perception routines (*vide supra*) and are contributors for a structure simplification metric.[42] Proposed transformations are discarded (or marked for special processing) [ooo] under two conditions: either there are no example reactions for the applied rule; or the target structure contains functional groups for which there is no example evidence that the group is tolerated. This status is indicated by parameter K.

No account is made in the one-step retrosynthesiser for prioritising routes that lead to known starting materials. ARChem already supports starting material matching using suppliers databases and this aspect of route ranking will become available when this work fully integrated with ARChem.

β-Hydroxy Ketones

Experiments 1 and 2 compare one-step routes to a pair of β-hydroxy ketones that differ in substituent positions and the number of stereogenic centres (see Table 62 and Table 63: *vide infra*). Experiment 1 generates five routes to (3S)-3-hydroxy-1,3-diphenylpropan-1-one. Two of these routes are automaticaly discarded via the “kill” criteria applied to matched rules RZC.2.1.1.2 and RZC.4.1.1.1. Both these rules are part of the RZC transform concerned with organozinc addition to aldehydes or prochiral ketones. The RZC transform is broad in scope and covers the application of organozinc functionality adjacent to electron-withdrawing groups (such as Reformatsky reagents formed from α-halo esters). Rule RZC.4.1.1.1 establishes that α-halo ketones are not likely to be viable precursors for this method as no examples exist in the supporting databases.

The results of experiment 1 (synthesis of *(3S)-3-hydroxy-1,3-diphenylpropan-1-one*) are shown in Table 62 ranked by highest enantioselective reputation (column B) followed by highest likely enantioselectivity (column E) to place the ketone reduction route into the highest ranked position (route #1, rule KR.1.1.1). Additional support for the feasibility of the proposed FGI modification is provided by example evidence that the reduction can be conducted in the presence of other ketones. Issues regarding the relative reactivity of the two ketone groups are eliminated in this case as the two ketone groups are equivalent by symmetry. The two lower ranked routes (#2 and #3) are variants of the aldol reaction. A literature search via Reaxys indicated that the target molecule has been synthesised but only via the aldol routes with no evidence of the proposed higher scoring stereoselective ketone reduction route.

[ooo] For example this might include considering a protecting group strategy

Guiding a retrosynthesis plan towards simple starting materials requires choosing routes that emphasise a reduction in precursor complexity.[117] With this goal in mind, route ranking should prioritise disconnections that maximise stereocentre clearance, ideally with simultaneous carbon-carbon or carbon-hetero bond cleavage.[117] Consequently the primary route selection criteria should not necessarily favour reaction reputation, likely yield, stereoselectivity and reliability over structural simplification.

In accordance with the above goal, the results of experiment 2 (synthesis of *(2S,3S)-3-hydroxy-2-methyl-1,3-diphenylpropan-1-one*) shown in Table 63 are ranked to favour precursor simplification. The six proposed routes are sorted to choose the best routes which maximise the retrosynthetic clearance of stereocentres with concurrent disconnection of carbon-carbon bonds. This ranking favours routes #6, #7 and #8. A secondary ranking applied within this group prioritises enantioselectivity over example reputation. This ordering promotes the direct aldol disconnection as the best route, followed by the equivalent Mukaiyama aldols disconnections.

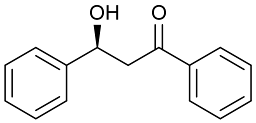
Ketone reduction (route #9) and hydrogenation of enones (route #1) are ranked below the aldol disconnections due to a lower number of cleared stereocentre and the absence of carbon-carbon bond disconnections marking these routes as less simplifying. In addition both these routes are automatically recognised by stereotopicity perception as presenting potential diastereoselectivity problems. The stereotopicity of the reacting centre may be used by the scoring module to rank diastereoselective routes below those that only need to address enantioselectivity.

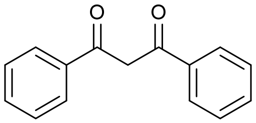
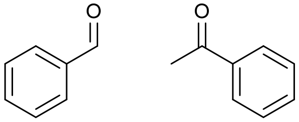
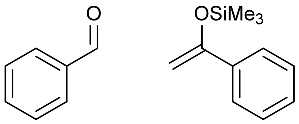
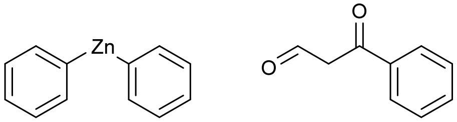
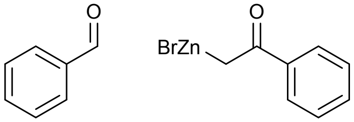
Route #	**Yield**	**% *ee***	**% *de***	**Reaction**	**Reference**
2	76	98		Direct *syn*-aldol	[330]
3	90 95	81 95		Mukiayama aldol	[331] [332]
8	98 90	97 88		Mukiayama *syn*-aldol	[333] [334]

Table 61 Known reactions matching the disconnections proposed in retrosynthesis experiments 1 and 2.

Zinc aryl addition to carbonyls (route #11) is automatically discounted by the scoring module as no evidence for the selective addition to an aldehyde in the presence of a ketone exists in the database (despite aldehydes being more reactive than ketones). Alternatively the route could be retained if the interfering ketone group is marked for subsequent processing by a protection/deprotection planning application. A tool capable of this type of analysis was developed for the Lhasa program.[335]

Table 61 lists known solutions for the proposed retrosynthetic steps. The literature routes employ both direct and Mukaiyama aldol reaction with the Mukaiyama aldol producing the best results.



Route #	Rule identifier	Generated precursors	A	B	C	D	E	F	Additional scoring parameters	K	L	M
1	*KR.1.1.1*		🟢	🟢	🟢	🟡	🟢	🟢	G. Reacting symmetrical carbonyl is enantiotopic H. 0 C-C bonds formed I. 1 stereocentre formed J. Tolerated FGs examples: KETONE (42)	×	×	×
2	*A.2.3.1.1.1*		🟢	🟢	🟡	🔴	🟨	🟡	G. Reacting carbonyl is enantiotopic H. 1 C-C bond formed I. 1 Stereocentre formed J. Tolerated FGs examples: n/a	×	×	✓
3	*MA.3.1.1*		🟢	🟨	🟨	🔴	🟨	🟡	G. Reacting carbonyl is enantiotopic H. 1 C-C bond formed I. 1 stereocentre formed J. Tolerated FGs examples: n/a	×	×	✓
4	*RZC.2.1.1.2*		🟨	🟨	🟨	🟡	🟡	🟨	G. Reacting carbonyl is enantiotopic H. 1 C-C bond formed. I. 1 Stereocentre is formed J. Tolerated FGs examples: KETONE (0)	✓	✓	×
5	*RZC.4.1.1.1*		⚫	⚫	⚫	⚫	⚫	⚫	No examples found of Reformatsky reagents formed from α haloketones.	✓	×	×

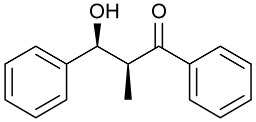
Rating Key: 🟩 Superb 🟢 Excellent 🟨 Very good 🟡 Good 🟠 Fair 🔴 Poor 🟤 Very poor ⚫ No support

Table 62 Experiment 1: A one-step retrosynthesis plan generated for (3S)-3-hydroxy-1,3-diphenylpropan-1-one.

A – reputation B – enantioselective reputation C – likely yield D – yield reliability

E – likely enantioselectivity F – enantioselectivity reliablity K – ✓ killed transform L – ✓ has a potential regioselectivity issue

M – ✓ is a known reaction (see Table 61)



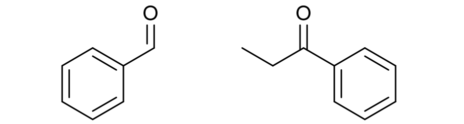
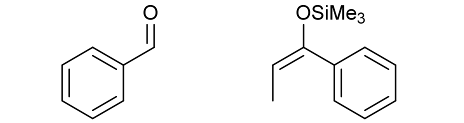
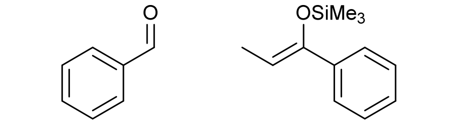
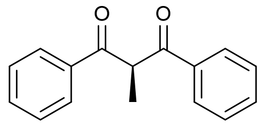
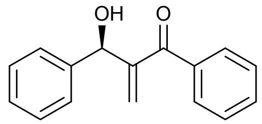
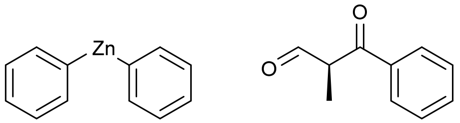
Route #	Rule identifier	Generated precursors	A	B	C	D	E	F	Additional scoring parameters	K	L	M
6	*A.2.1.1.1.9*		🟢	🟢	🟡	🔴	🟡	🟢	G. Reacting carbonyl is enantiotopic H. 1 C-C bond formed. I. 2 Stereocentres formed J. Tolerated FGs examples: n/a	×	×	×
7	*MA.1.1.1*		🟡	🟠	🟢	🟠	🟡	🟢	G. Reacting carbonyl is enantiotopic H. 1 C-C bond formed. I. 2 stereocentres formed J. Tolerated FGs examples: n/a	×	×	×
8	*MA.1.3.1*		🟡	🟡	🟡	🟡	🟠	🟡	G. Reacting carbonyl is enantiotopic H. 1 C-C bond formed. I. 2 stereocentres formed J. Tolerated FGs examples: n/a	×	×	✓
9	*KR.1.1.1*		🟢	🟢	🟢	🟡	🟢	🟢	G. Reacting carbonyl is diastereotopic *and* desymmetrised. H. 0 C-C bonds formed. I. 1 stereocentre formed J. Tolerated FGs examples: KETONE (55)	×	×	×
10	*GC2R.1.1*		🟢	🟡	🟢	🟢	🟡	🟢	G. Reacting alkene is diastereotopic. H. 0 C-C bonds formed. I. 1 Stereocentre is formed J. Tolerated FGs examples: ALCOHOL (52), KETONE (26)	×	×	×
11	*RZC.4.1.1.1*		🟡	🟡	🟡	🟡	🟡	🟡	G. Reacting carbonyl is diastereotopic H. 1 C-C bond formed. I. 1 Stereocentre is formed J. Tolerated FGs examples: KETONE (0)	✓	✓	×

Table 63 Experiment 2: A one-step retrosynthesis plan generated for (*2S,3S*)-3-hydroxy-2-methyl-1,3-diphenylpropan-1-one. Key – see Table 62 caption.

1,2-Diols

Experiments 3 and 4 compare one-step routes generated for two stereoisomeric 1,2 diols. The results for the one-step retrosynthesis plan for the *syn*-diol *(1R,2R)-1-phenylpropane-1,2-diol* are listed in Table 65 and retrosynthesis plan for the *anti*-diol *(1R,2S)-1-phenylpropane-1,2-diol* are listed in Table 66. The prosed routes are shown ranked by the number cleared stereocentres followed by the number of disconnected carbon-carbon bonds and then by stereoselective reputation and finally by likely stereoselectivity.

Both analyses offer seven routes involving three types of reaction: dihydroxylation of alkenes (routes #12 and #17); reduction of ketones to secondary alcohols (routes #13, #14, #18 and #19); and the alkylation of aldehydes (routes #15, #16, #21 and #21). Those involving the arylation of aldehydes have been omitted for space reasons, but are comparable to the listed alkylation routes. No other routes were offered due to the limited number of transforms available. The ranking criteria outlined above placed the dihydroxylation reaction at the top based on best precursor simplification. The remaining routes receive a lower ranking as each only clears one stereocentre and must deal with diastereoselective control. The alkylation/arylation routes are marked as either discarded or requiring functional group protection as the reaction example evidence suggests that the alcohol group is interfering.

The overall rating for the dihydroxylation of *trans*-alkyl/aryl alkenes (rule DH.1.1.2) to form the *syn*-diol is excellent with a very good estimated yield and superb estimated enantioselectivity and reliability. In contrast the dihydroxylation of *cis*-alkyl/aryl alkenes (rule DH.1.2.2) to form the *anti*-diol has only a fair rating for both enantioselectivity and reliability. These evidence-based estimations are in agreement with literature examples of the target reactions (see Table 64: the *syn*-diol from the trans-alkene is 99 %ee; the *trans*-diol from the cis-alkene 72% ee).

To increase the likelihood of improved stereoselectivity, possibly at the expense of better yields and with more difficult stereocontrol, a chemist can consider the alternative routes by altering the ranking criteria to favour stereoselectivity over precursor simplification. This demotes the *cis*-alkene dihydroxylation reaction below ketone reduction and aldehyde alkylation in experiments 3 and 4. It is noted that the matched ketone reduction rules KR.1.1.1 and KR.1.2.1 do not address the diastereoselectivity problem (see rule details in Table 60) and the high reputation and stereoselectivity ratings are consequently over optimistic. This is typical of underspecified rules. [ppp] Both rules list a significant number of examples with tolerated hydroxyl

[ppp] The transform needs further development to address common diastereoselectivity situations.

functional groups but this count is generalised and is not specific to adjacent hydroxyl groups. A literature survey for the corresponding transformation only found specialised solutions using enzyme catalysts.

The zinc and magnesium mediated alkylation transformations show better promise and are viable alternatives to the dihydroxylation route. The statistics record that no known example reactions tolerate the hydroxyl group indicating that functional group protection is a likely requirement. The matched rules RZC.1.1.2.1, RZC.1.1.2.2, RMC.1.1.2.2 and RMC.1.1.2.2 all directly address the diastereoselective issue introduced by the α-chiral centre meaning the rule rating can be treated as realistic. In experiment 3, chelation control using alkylzinc reagents leads to the *syn*-diol with likely very good diastereoselectivity and excellent reliability (route #15). The corresponding reaction using Grignard reagents (route #16) is rated to give poor diastereoselectivity with very poor reliability indicating this route should be rejected. In experiment 4, non-chelation conditions that promote polar Felkin-Ahn selectivity lead to the required *anti*-diol target. This outcome cannot be attained using zinc alkylating reagents (route #21) as zinc is a powerful chelating metal. The corresponding Grignard variant can be used and has a reasonable reputation with very good diastereoselectivity (route #20) when applied in the right conditions. Figure 134 shows a literature example of the reaction for a modification of route #20 where the hydroxyl group is protected as a silyl ether. The reaction achieved moderate yield and diastereoselectivity in favour of the desired modified target (5:1).[336]

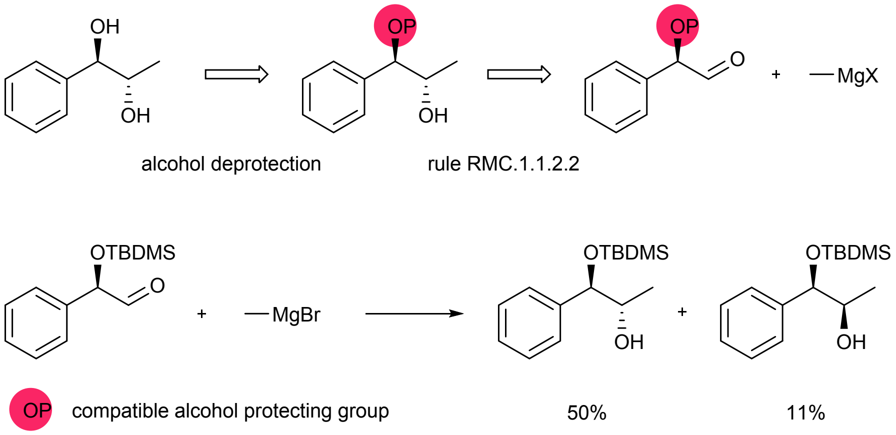
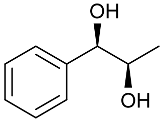


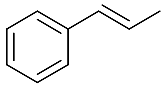
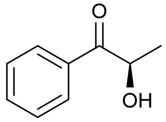
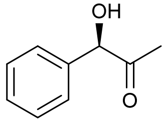
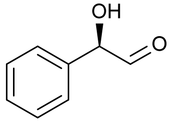
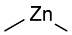
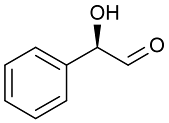
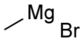
Figure 134 A feasible retrosynthetic route from *(1R,2S)-1-phenylpropane-1,2-diol* requiring alkyl addition to the protected α-hydroxy aldehyde under diastereoselective control. This has been achieved with reasonable diastereoselectivity and moderate yield using methyl magnesium bromide added to the silyl ether protected aldehyde under non-chelating conditions. [9]

Table 64 lists known solutions for the proposed retrosynthetic steps for synthesising the 1,2-diols.

Route #	Yield	% *ee*	% *de*	Reaction	Reference
12	99 99 99	92 99 97		Standard Sharpless dihydroxylation (*trans* alkene)	337 338 339
14	86		98	Enzymatic reduction of α-hydroxy ketones	340
18	85	72	98	Enzymatic reduction of α-hydroxy ketones	340
17	-	72		Modified Sharpless dihydroxylation (*cis* alkene)	341

Table 64 Known reactions matching the disconnections proposed in retrosynthesis experiments 3 and 4.

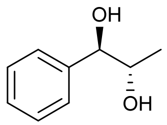


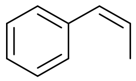
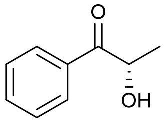
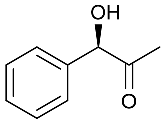
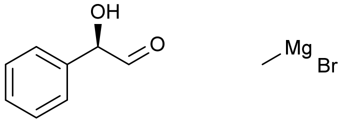
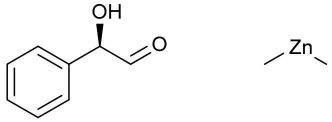
Route #	Rule identifier	Generated precursors	A	B	C	D	E	F	Additional scoring parameters	K	L	M
12	*DH.1.1.2*		🟩	🟨	🟨	🟡	🟢	🟢	G. Reacting alkene is enantiotopic H. 0 C-C bonds formed. I. 2 Stereocentres formed J. Tolerated FGs examples: n/a	×	×	✓
13	*KR.1.1.1*		🟩	🟩	🟩	🟡	🟩	🟩	G. Reacting carbonyl is diastereotopic. H. 0 C-C bonds formed. I. 1 stereocentre formed J. Tolerated FGs examples: ALCOHOL (112)	×	×	×
14	*KR.1.2.1*		🟩	🟩	🟨	🟠	🟩	🟩	G. Reacting carbonyl is diastereotopic. H. 0 C-C bonds formed. I. 1 Stereocentre is formed J. Tolerated FGs examples: ALCOHOL (499)	×	×	✓
15	*RZC.1.1.2.1*	 	🟡	🟡	🟠	🔴	🟨	🟩	G. Reacting carbonyl is diastereotopic. H. 1 C-C bond formed. I. 1 stereocentre formed J. Tolerated FGs examples: ALCOHOL (0)	✓P	×	×
16	*RMC.1.1.2.1*	 	🟡	🟡	🟠	🟠	🔴	🟤	G. Reacting carbonyl is diastereotopic. H. 1 C-C bond formed. I. 1 stereocentre formed J. Tolerated FGs examples: ALCOHOL (0)	✓P	×	×

Rating Key: 🟢 Superb 🟩 Excellent 🟨 Very good 🟡 Good 🟠 Fair 🔴 Poor 🟤 Very poor ⚫ No support

Table 65 Experiment 3: A one-step retrosynthesis plan generated for (1R,2R)-1-phenylpropane-1,2-diol.

A – reputation **B – sterieoselective reputation** **C – likely yield** **D – yield reliability**
E – likely stereoselectivity **F – stereoselectivity reliablity** **K – ✓P kill the transform or mark the target for protection**
L – ✓ has a potential regioselectivity issue **M – ✓ is a known reaction (see Table 64)**



Route #	Rule identifier	Generated precursors	A	B	C	D	E	F	Additional scoring parameters	K	L	M
17	*DH.1.2.2*		🟨	🟡	🟡	🟨	🟠	🟠	G. Reacting alkene is enantiotopic H. 0 C-C bonds formed. I. 2 stereocentre formed J. Tolerated FGs examples: n/a	✗	✗	✓
18	*KR.1.1.1*		🟢	🟢	🟢	🟡	🟢	🟢	G. Reacting carbonyl is diastereotopic. H. 0 C-C bonds formed. I. 1 Stereocentre is formed J. Tolerated FGs examples: ALCOHOL (499)	✗	✗	✓
19	*KR.1.2.1*		🟢	🟢	🟨	🟠	🟨	🟢	G. Reacting carbonyl is diastereotopic. H. 0 C-C bonds formed. I. 1 Stereocentres formed J. Tolerated FGs examples: ALCOHOL (112)	✗	✗	✗
20	*RMC.1.1.2.2*		🟡	🟠	🟡	🟠	🟢	🟨	K. Reacting carbonyl is diastereotopic. L. 1 C-C bond formed. M. 1 stereocentre formed N. Tolerated FGs examples: ALCOHOL (0)	✓P	✗	✗
21	*RZC.1.1.2.2*		🟤	🟤	⚫	⚫	⚫	⚫	G. Reacting carbonyl is diastereotopic. H. 1 C-C bond formed. I. 1 stereocentre formed J. Tolerated FGs examples: ALCOHOL (0)	✓P	✗	✗

Rating Key: 🟩 Superb 🟢 Excellent 🟨 Very good 🟡 Good 🟠 Fair 🔴 Poor 🟤 Very poor ⚫ No support

Table 66 Experiment 4: A one-step retrosynthesis plan generated for (1R,2S)-1-phenylpropane-1,2-diol.

A – reputation **B – stereoselective reputation** **C – likely yield** **D – yield reliability**
E – likely stereoselectivity **F – stereoselectivity reliablity** **K – ✓P kill the transform or mark the target for protection**
L – ✓ has a potential regioselectivity issue **M – ✓ is a known reaction (see Table 64)**

Propargyl Alcohols

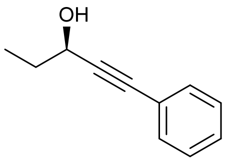
Experiments 5 and 6 present the one-step retrosynthetic analysis of two secondary propargyl alcohols which are differentiated by alkyl and aryl substituents attached to the alcohol group. The results shown in Table 68 and Table 69 are ranked to promote routes that maximise stereo-simplification with concurrent structural cleavage followed by reputation and enantioselectivity. Consequently the ketone alkynylation and zinc mediated alkylation and arylation routes (#22, #23. #26 and #27) all achieve top rank. These disconnections have excellent reputations, predicted yield with likely very good enantioselectivity and all offer practical synthetic routes (Table 67). The alkynylation routes have slightly better enantioselectivity reliability than alkylation or arylation.

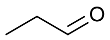
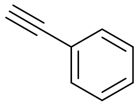
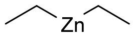
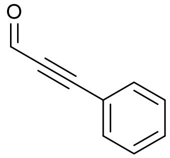
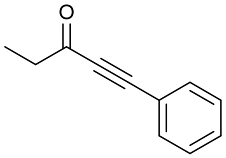
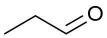
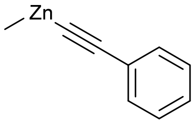
A survey of the alkynylation example reactions revealed that significant proportion is mediated via *in situ* generated mixed alkyl-alkynyl zinc reagents. This illustrates a minor problem with the reaction classification algorithm for ketone alkynylation reactions represented either as the alkyne reactant or the intermediate alkynyl-zinc reagent. In this case the reaction hyperstructures differ and are consequently treated as different reactions with different CRC values. This difference leads to two rules C3CO.1.1.1 and RZC.3.1 which ideally should be merged. The latter rule is supported by a very small number of explicit alkynyl-zinc examples reactions and appears as the lowest scoring routes (#25 and #29) in both experiments.

Route #	Yield	% *ee*	% *de*	Reaction	Reference
22	75	90		Zinc alkynylation of aldehydes	342
23	90	28		Zinc alkylation of propargyl aldehydes	343
24	95	46		Ketone reduction (using *(R)-alpine borane*)	344
26	90 97 95	90 90 96		Zinc alkynylation of aldehydes	345 346 347
27	93	95		Zinc arylation of propargyl aldehydes	348

Table 67 Known reactions matching the disconnections proposed in retrosynthesis experiments 5 and 6.

Routes with poor or very poor reputations that rank lower than highly rated alternative routes are prime candidates for removal. It is particularly important to prune out poor routes when run in unattended multistep retrosynthetic analyses operated by ARChem as otherwise the total number of evaluated routes rapidly becomes unmanageable. Poor rated routes should only be retained for evaluation when no better alternative is present.

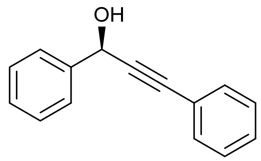


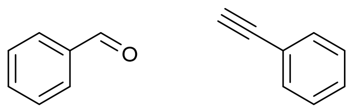
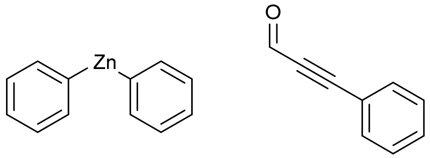
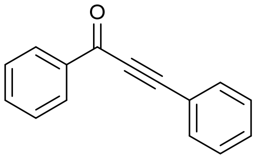
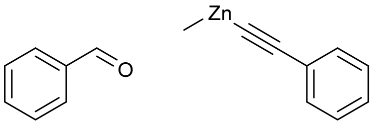
Route #	Rule identifier	Generated precursors	A	B	C	D	E	F	Additional scoring parameters	K	L	M
22	*C3CO.1.1*	 	Excellent	Excellent	Very good	Good	Very good	Very good	G. Reacting alkene is enantiotopic H. 1 C-C bond formed. I. 1 Stereocentre formed J. Tolerated FGs examples: ACETYLENE (370)	✗	✗	✓
23	*RZC.1.1.2*	 	Excellent	Excellent	Very good	Good	Very good	Good	G. Reacting carbonyl is enantiotopic. H. 1 C-C bonds formed. I. 1 stereocentre formed J. Tolerated FGs examples: ACETYLENE (28)	✗	✗	✓
24	*KR.1.2.3*		Excellent	Excellent	Very good	Good	Very good	Very good	G. Reacting carbonyl is enantiotopic H. 0 C-C bond formed. I. 1 Stereocentre is formed J. Tolerated FGs examples: ACETYLENE (214)	✗	✗	✓
25	*RZC.3.1*	 	Very poor	Very poor	Very good	Very poor	Fair	Poor	G. Reacting carbonyl is enantiotopic. H. 1 C-C bond formed. I. 1 stereocentre formed J. Tolerated FGs examples: ACETYLENE (2)	✗	✗	✗

Rating Key: ● Superb ● Excellent ● Very good ● Good ● Fair ● Poor ● Very poor ● No support

Table 68 Experiment 5: A one-step retrosynthesis plan generated for (3R)-1-phenylpent-1-yn-3-ol.

A – reputation **B – enantioselective reputation** **C – likely yield** **D – yield reliability**
E – likely enantioselectivity **F – enantioselectivity reliablity** **K – ✓ killed transform** **L – ✓ has a potential regioselectivity issue**
M – ✓ is a known reaction (see Table 67)



Route #	Rule identifier	Generated precursors	A	B	C	D	E	F	Additional scoring parameters	K	L	M
26	*C3CO.1.1*		🟩	🟩	🟨	🟡	🟨	🟨	G. Reacting aldehyde carbonyl is enantiotopic H. 1 C-C bond formed. I. 1 Stereocentre formed J. Tolerated FGs examples: ACETYLENE (370)	×	×	✓
27	*RZC.1.1.2*		🟩	🟩	🟨	🟡	🟨	🟡	G. Reacting aldehyde carbonyl is enantiotopic H. 1 C-C bond formed. I. 1 Stereocentre is formed J. Tolerated FGs examples: ACETYLENE (28)	×	×	✓
28	*KR.1.1.2*		🔴	🔴	🟨	🔴	🟡	🟡	G. Reacting ketone carbonyl is enantiotopic. H. 0 C-C bonds formed. I. 1 stereocentre formed J. Tolerated FGs examples: ACETYLENE (7)	×	×	×
29	*RZC.3.1*		🟤	🟤	🟨	🟤	🟠	🔴	G. Reacting aldehyde carbonyl is enantiotopic. H. 1 C-C bond formed. I. 1 stereocentre formed J. Tolerated FGs examples: ACETYLENE (2)	×	×	×

Rating Key: 🟢 Superb 🟩 Excellent 🟨 Very good 🟡 Good 🟠 Fair 🔴 Poor 🟤 Very poor ⚫ No support

Table 69 Experiment 6: A one-step retrosynthesis plan generated for (1R)-1,3-diphenylprop-2-yn-1-ol.

A – reputation **B – enantioselective reputation** **C – likely yield** **D – yield reliability**
E – likely enantioselectivity **F – enantioselectivity reliablity** **K – ✓ killed transform** **L – ✓ has a potential regioselectivity issue**
M – ✓ is a known reaction (see Table 67)

A notable difference between the two retrosynthetic analyses is apparent in routes #24 and #28. The latter's poor reputation suggests that reduction of aryl-alkynyl ketones has practical difficulties. Examples of all routes except route #28 were found in the literature although routes #23 and #24 suffered poor enantioselectivities.

Summary

The small scale one-step retrosynthesiser was used to generate potential solutions for a number of modest enantioselective and diastereoselective synthesis problems. The target substrates were deliberately kept simple so that the results of known routes could be compared to the evidence-based estimates applied by each rule. The use of multiple scoring parameters derived from example evidence and perceived properties of the reaction rule, target and precursors where used to rank the proposed routes. Prioritising precursor simplification over transform reputation and stereoselectivity often leads to the best routes. If poor stereoselectivity for the most simplifying routes is predicted then ranking the routes by best stereoselectivity offers an alternative approach that usually gives good results.

Many of the existing transforms can be improved by adding detailed rules for elucidating the degree of diastereoselective control available from example evidence. However this is currently hampered by an inability to evaluate steric factors. This issue is discussed in more detail in the following section.

Future Directions

> ***"Most people overestimate what they can do in one year and underestimate what they can do in ten"***
>
> **– Bill Gates in "The road ahead".**

This final section looks towards further developments of the methods described in this thesis.

Improved Transform Relevance

The current approach to transform design neglects the influence of reactions conditions on product outcomes especially with respect to issues such as regioselectivity. A case in point is the nucleophilic ring opening of asymmetric epoxides under either acidic or basic conditions. Acidic conditions favour nucleophilic attack at the carbon atom that best stabilises the cationic character developed in the transition state. Typically this is either at the most substituted carbon atom or at benzylic or allylic centres. Anionic nucleophiles formed under basic conditions attack under steric control at the least substituted carbon atom. Unmasking such regioselective effects requires that each reaction is classified by reagent class and solvent type to one of a set of prototypical conditions.[335, 234] Reaction condition constraints that reference the prototype conditions would then be used to further subdivide the reaction examples within each rule example set to improve example relevance and the estimated reputation and stereoselectivity scoring parameters.

Improved Stereoselective Transforms

Two significant problems were encountered when developing effective stereoselectivity rules. First was a general lack of available selectivity data and the second was an inability to estimate relative steric bulk of substituents when evaluating rules relying on R_S, R_M and R_L atom types to model the steric differentiation that determines stereoselectivity and product outcomes. The current set of implemented diastereoselective rules are limited to special cases where substituents are differentiated by non-steric factors.

The development of diastereoselective rules with effective scoring parameters was hampered on frequent occasions by the lack of sufficient supporting *% de* data in the CIRX database. For example the directed epoxidation of allylic alcohols (*vide supra*) achieved poor quantitative selectivity support even though sufficient reaction examples were available. This was attributed to a general lack of reporting of *% de* values in the primary literature compared to the more prevalent reporting of *% ee* values. For example 6% of the reactions in the CIRX database report *% ee* values whereas only 3% of reactions report *% de* values. The poor support is further compounded as diastereoselectivity data is spread over examples of a much broader set of reaction types compared to the situation with enantioselectivity data. The latter tends to cluster in the examples of a smaller group of established enantioselective reactions. It may be possible to overcome this deficiency by utilising computational models that attempt to derive good estimates of diastereoselectivity parameters from the steric and electronic properties of substituents.[349]

Recently computational methods have been successfully applied to correlate the steric parameters of catalyst ligands to the enantioselectivity of a reaction as a tool for rational catalyst design. This approach has been applied to a selection of asymmetric catalytic reactions.[350, 349] In these studies, empirically derived substituent steric values such as the Winstein-Holness (A-Value),[351] Taft[352, 353] and Charton[354, 355] parameters were poorly correlated except in the case of simple alkyl substituents.[349] Very good to excellent correlation between catalyst/ligand bulk and product enantioselectivity was found using more advanced computationally derived STERIMOL parameters[350] suggesting that this method may be transferable to the task of identifying the steric ranking of arbitrary R_S, R_M and R_L substituents in the reacting substrate. It may also be possible to estimate *% ee* or % de values in modelled reactions using this approach if sufficient example data is available. In the latter case it is essential that the reaction examples used to derive the selectivity estimation models are filtered to include only those using the same reagents or catalysts under matched conditions and have the necessary reported *% de/ee* values.

Verloop's STERIMOL method[356] uses CPK models[357] to calculate several distance parameters within substituents to attain a level of information content able to describe non-symmetric steric effects. The most useful STERIMOL parameters have been found to be L, the length of the

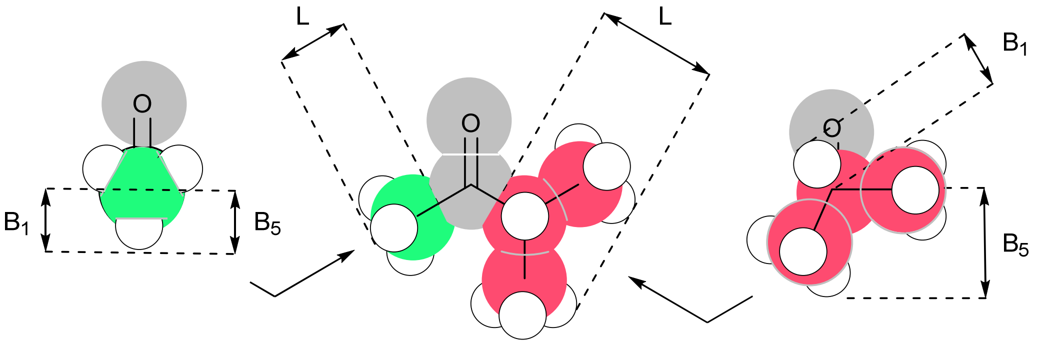


Figure 135 The STERIMOL distance parameters L, B_1 and B_5 for the two substituents of methyl isopropyl ketone.

principal substituent axis measured along the connecting bond, and two width parameters B_1 (proximal steric bulk) and B_5 (distal steric bulk) normal to the principal substituent axis. The B parameters measure the minimum and maximum extents of the substituent profile (Figure 135). In contrast, the better known Winstein-Holness, Taft and Charton parameters are single valued and necessarily correlate to a symmetrical spherical model of steric bulk in which rotation about the principal substituent bond is assumed to be faster than the chemical process used to measure the steric effect.[350] Typically these parameters only worked well in the catalyst design study for small or symmetrical substituents.[349]

Sigman *et al* have developed a model[a] to correlate the STERIMOL parameters of varied catalyst ligands to the measured enantioselective product ratios generated by the chiral catalysed Nozaki-Hiyama-Kishi (NHK) reaction. The Sigman model estimates the level of enantioselectivity achievable by a reaction using a specific catalyst and substrate by applying the following equation, where the constant and coefficient terms are determined by regression analysis conducted on a training set composed of a reactions in which the catalyst ligands are varied :

$$\Delta\Delta G^{\ddagger} = xB_1 + yB_5 + zL + c$$

Sigman postulates that the proximal steric bulk parameter B_1 dominates for reactions with well-ordered transition states as B_1 appears to correlate to the closest catalysis/substrate repulsive interactions due to branching at the ligand α atom. Reactions with less organised transition states incur larger contributions from the length (L) and distal steric bulk (B_5) parameters.

It is a reasonable assumption that steric rankings for reactant substituents can be derived using an analogous approach where the reaction conditions (including the catalyst) remain constant but the substrates are varied. Well represented enantioselective reactions such as the CBS reduction of ketones[358] are potentially good probes to attempt to model the effect of differences in substituent bulk on enantioselectivity. If reasonably accurate substituent bulk parameters can be obtained, rules containing R_S, R_M and R_L substituent constraints can be solved. Ranking by the proximal steric bulk parameter alone, as a first level approximation, may be sufficient to select qualifying examples for these reaction rules. A *relative* bulk ($\mathscr{B}$) parameter for determining substituent rank can be computed by the equation:

$$\mathscr{B} = B_1 + y'B_5 + z'L$$

[a] Additional cross terms with Hammett parameters can be added to model electron-donating/withdrawing effects.

Stereoselective Retrosynthetic Strategies

The use of structural patterns to locate strategic stereocentre relationships in a target molecule and convert these into retrosynthetic simplifying goals as part of an overall retrosynthetic plan is feasible.

Methods for constructing and applying goal directed plans to the retrosynthetic problem using the principles of *transform relevance* and *transform applicability* has been outlined over 40 years ago by Wipke (see chapter 1).[29] A transform is relevant if it can achieve specific strategic goals that simplify aspects of the target molecule but otherwise is not applicable as there is no direct retron match. Applicability can only be achieved by measuring the difference between the current target and the relevant transform and using these differences to generate tactical sub-goals. The applications of transforms that solve the sub-goals make it possible to promote the *relevant* transform to an *applicable* transform to achieve the key disconnections.

In his book "The Logic of Chemical Synthesis",[117] E. J. Corey describes an armoury of reusable strategies for exercising stereochemical control in the design of multistep syntheses. In the retrosynthetic direction these stereoselective goals are translated into the reduction of structural complexity by the systematic removal of stereocentres. This entails repeated identification of *clearable* and *non-clearable* stereocentres at each step of the retrosynthetic plan under the guidance of known stereo-simplifying transforms. Non-clearable stereocentres are identified as those that cannot be eliminated under stereocontrol using known transforms and are instead derived *via* starting materials selected from the chiral pool. Clearable stereocentres on the other hand can be subjected to a number strategies ranging from the control of stereocentres in polycyclic and monocyclic systems through to acyclic situations where local steric bias is judged sufficient to control stereoselectivity (see previous section).

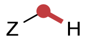
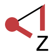
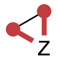
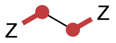
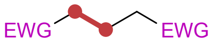
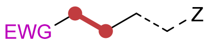
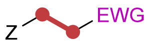
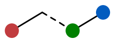
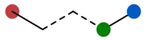
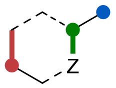
Corey's strategy rules can be ranked by the degree of complexity reduction each exhibits. A small selection of the most useful strategies is summarised in the following list:

1. Reduce stereo complexity and the size of the target molecule by the *simultaneous* retrosynthetic disconnection of C-C bonds and the concurrent removal of adjacent stereocentres.
2. Clear stereocentres by the removal of functionalised substituents to generate stereotopic functionality such as C=C, C=O, C=N.
3. Simplify stereo complexity by *converting* a 1,n stereo relationship to a 1,2 stereo relationship by a functional group transposition if it can be followed by a strategy that

clears the generated 1,2 stereocentre relationship (*i.e.* follow strategy 3 with either strategy 1 or strategy 2).

4. Separate remote stereocentres by a molecular disconnection on a bond path between those stereocentres.
5. *… and so on*

A number of these strategies can be expressed as sets of annotated structural patterns[b] in a manner similar to those used to describe transform rules or for recognising functional groups (*vide supra*). These patterns define the required feature relationships and define simplification

	Strategy Pattern	Strategy type	Simplification Goals	Relevant transforms
1		2	1 stereo removal	Hydrogenation of C=O, C=N
2		2	1 stereo removal	Epoxidation, aziridination
3		2	1,2 stereo removal	Epoxidation, aziridination
4		2	1,2 stereo removal	Dihydroxylation, amino hydroxylation, halolactonisation
5		1	1,2 stereo removal + bond cleavage	Conjugate addition
6		1	1,2 stereo removal + bond cleavage	FGI + conjugate addition
7		1	1,2 stereo removal + bond cleavage	Aldol, Henry, aza-aldol
8		3	1,4 → 1,2 stereo transposition	Claisen, Cope and analogues
9		3	1,5 → 1,2 stereo transposition	Claisen, Cope and analogues
10		3	1,5 → 1,2 stereo transposition	Claisen / aza-Claisen / Eschenmoser / Johnson / Ireland Claisen

🔵	GOAL = PRESERVE_STEREO	🔴	GOAL = REMOVE_STEREO	🟩	GOAL = CREATE_BOND
🟢	GOAL = CREATE_STEREO			🟥	GOAL = BREAK_BOND

Table 70 The formulation of some stereoselective strategy patterns used to set retrosynthetic goals. The strategy type numbers refer to the strategies listed immediately above.

goals via atom and bond annotations. Those strategies that are not amenable to direct pattern representation could be handled by executable scripts that identify the required features programmatically and then define appropriate simplification goals.

[b] These patterns could be referred to as “strategons” as matching them to a target molecule identifies the locations of strategic goals.

Table 70 lists a small selection of proposed strategy patterns specifically chosen to illustrate the generation of goal plans in the example analyses shown in Figure 136 and Figure 138. The key features of these patterns are they are generalised structural fragments concerned only with the relative locations of important features such as stereocentres, electronic features and heteroatoms. Specific bond orders, atom types and functional groups are largely ignored to allow for the discovery of strategic relevance in broad sets of transforms. The patterns are annotated with goal directives such as CREATE_BOND, BREAK_BOND, REMOVE_STEREO, PRESERVE_STEREO *etc*. These goal directives are transferred to the matched target molecule to be acted on by the retrosynthesis executive. Each strategy pattern would be matched to sets of transform rules in an off-line process to create a cross-reference table between strategies and relevant transforms. Relevant transforms would be identified by matching strategy patterns to the target molecule and looking up the cross-reference table for the transforms that achieve the goals set by the strategy.

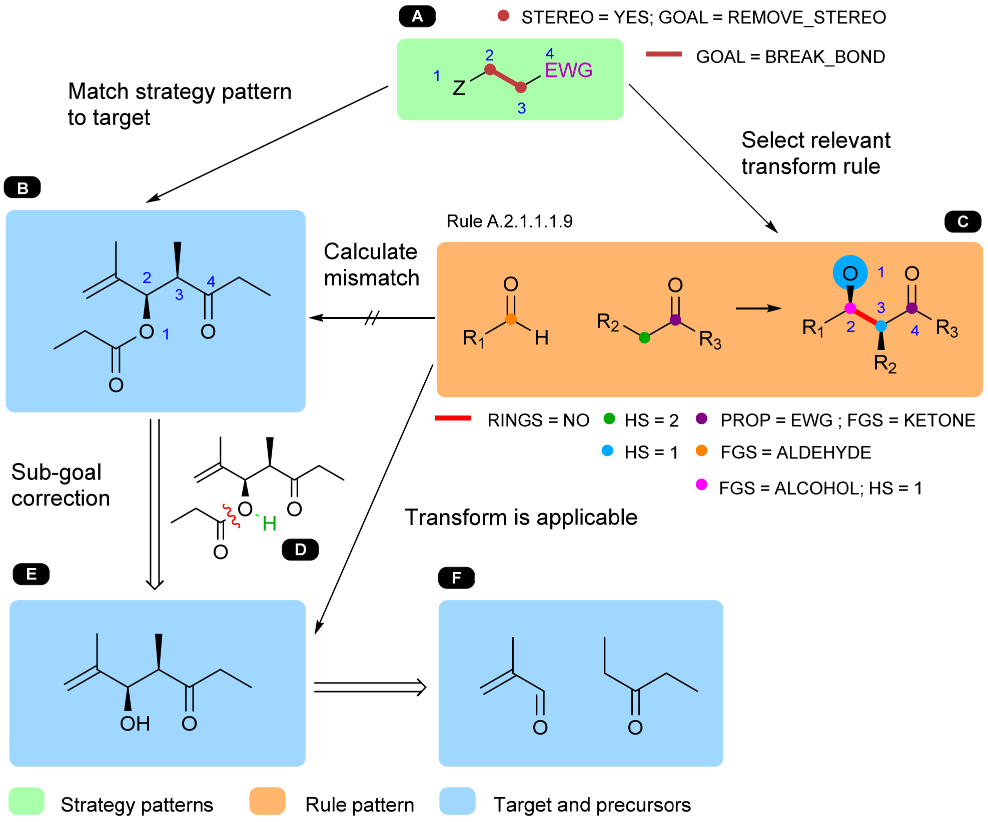


Figure 136 A strategy pattern A identifies simplifying goals in a target molecule B and selects relevant transforms for consideration. The difference graph D generated from the target and a selected transform C is used to set sub-goals which, if achievable, generates intermediate E. The corrected precursor can now be converted to precursors F by the applicable transform C.

Figure 136 illustrates the necessary processing steps. Strategy pattern A is matched to target molecule B and the mapping between them is noted. Rules that are cross referenced to the strategy pattern contain the required features and also achieve the set goals.

For example, one of the relevant rules (C, an aldol transformation) is selected and the mapping between it and the strategy pattern is used to find the differences between the target and rule. A direct comparison of the rule atoms and bonds to those in the target molecule establishes that atom position 1 is mismatched and that an ester C-O bond must be broken and a hydrogen atom must be added to the oxygen atom to eliminate the differences. These differences are translated into site specific sub-goals (BREAK_BOND, ADD_HYDROGEN) which must be solved by a transform search to generate precursor E. The transform C is now applicable to precursor E and is used to generate the achiral precursors F.

Deeper strategic analysis can be achieved by casting the target molecule as a skeletal graph containing only the key features retained in strategy patterns (Figure 137). A recursive application of strategic patterns to the skeletal target graph followed by direct manipulation of the skeleton according to the matched strategy goals (without reference to specific transforms) leads to a succession of stereo simplified precursor skeletons until no further goals can be identified. This technique generates a series of partial plans represented by a logically structured goal list containing 'AND/OR' nodes.[29] The 'AND' nodes, representing concurrent goals[c], are created for non-overlapping strategic pattern matches while 'OR' nodes, representing alternative goals, are added for overlapping matches.

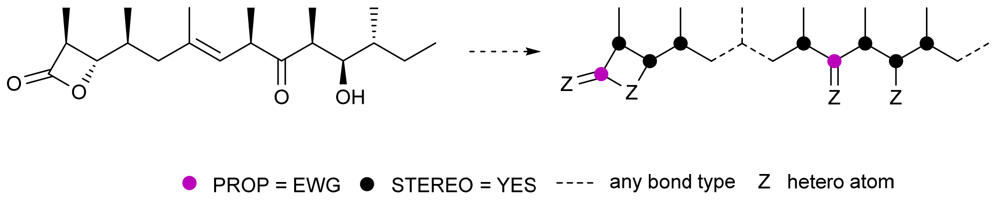


Figure 137 The abstraction of key stereo, electronic and heteroatom relationships in Ebelactone-A by the use of a skeletal graph.

[c] Concurrent goals can be executed in any order as the order of goal satisfaction is independent of the order of goal discovery.

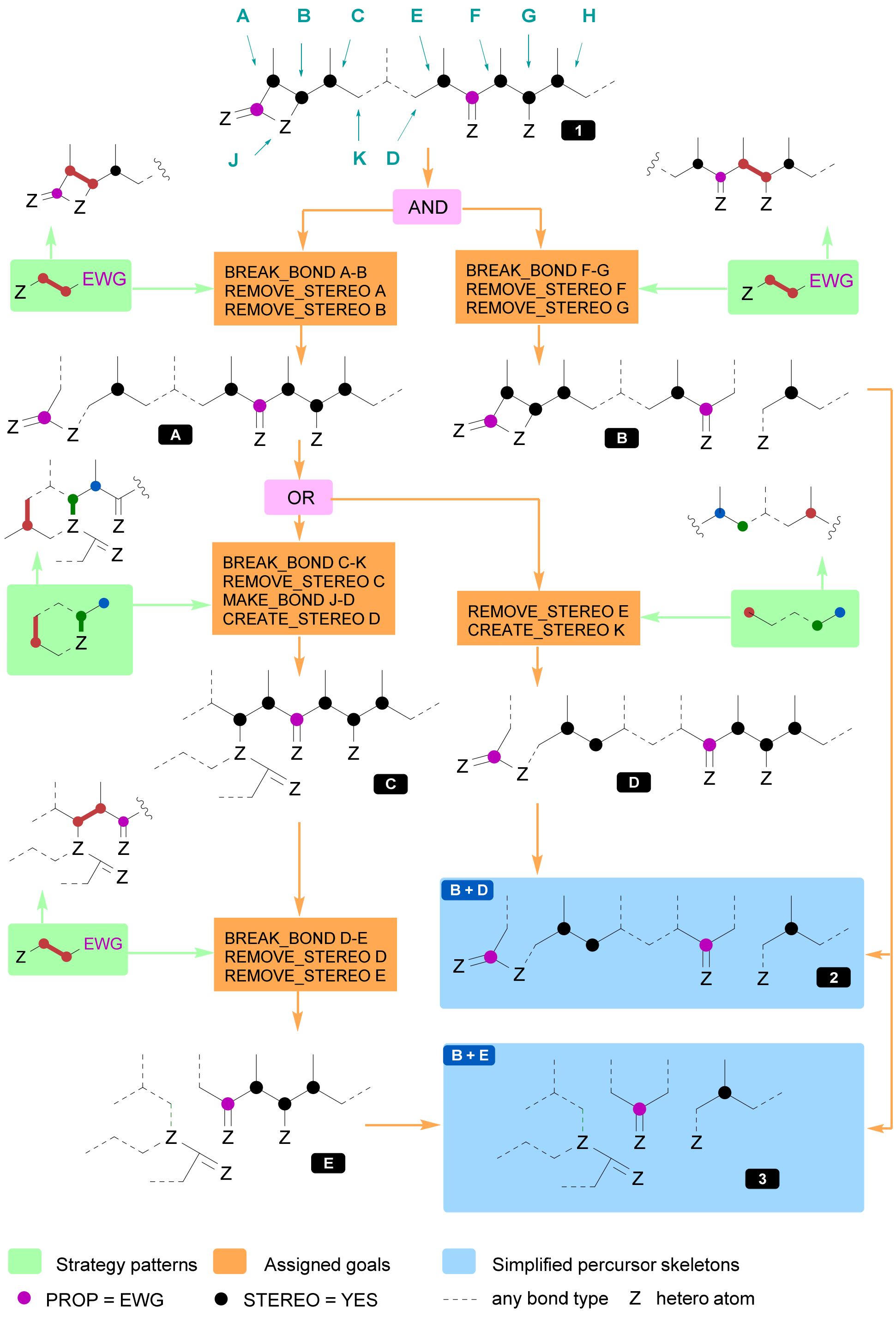


Figure 138 A simplification plan for the stereo controlled synthesis of Ebelactone-A (1) developed by recursively matching a limited set of strategy patterns to the Ebelactone-A skeleton graph and editing the skeleton according to the simplification goals (A, B, C, D and E). The most productive plan eliminates all but one stereocentre (outcome 3).

The potential to achieve non-trivial multistep plans is illustrated in Figure 138 for the β-lactone enzyme inhibitor Ebelactone-A [359, 360] using the strategy patterns selected from Table 70. Starting from the skeletal form of the target molecule (Figure 137), the analysis identifies two independent bond cleavages which achieve 1,2 stereo clearances (goals A *and* B using strategy 1) and two alternative 1,5 to 1,2 stereo transpositions (goal C *or* D using strategy 3), only one of which sets up a further recognisable bond cleavage which again achieves a 1,2 stereo clearance (goal E). The plan achieving the best combinations of goals (A-C-E and B) clears six of the seven stereocentres and cleaves the skeleton into three balanced pieces. The alternative combined plan (A-D and B) clears all but three stereocentres but has a lower disconnective convergence rating so is ranked below the best plan. [1]

Paterson and Hulme's enantioselective synthesis of (-)-Ebelactone A and B [359] was achieved using a similar strategy (Figure 139). The key strategic difference in their route was the early retrosynthetic removal of the stereocentre labelled H which resisted removal in the algorithmic

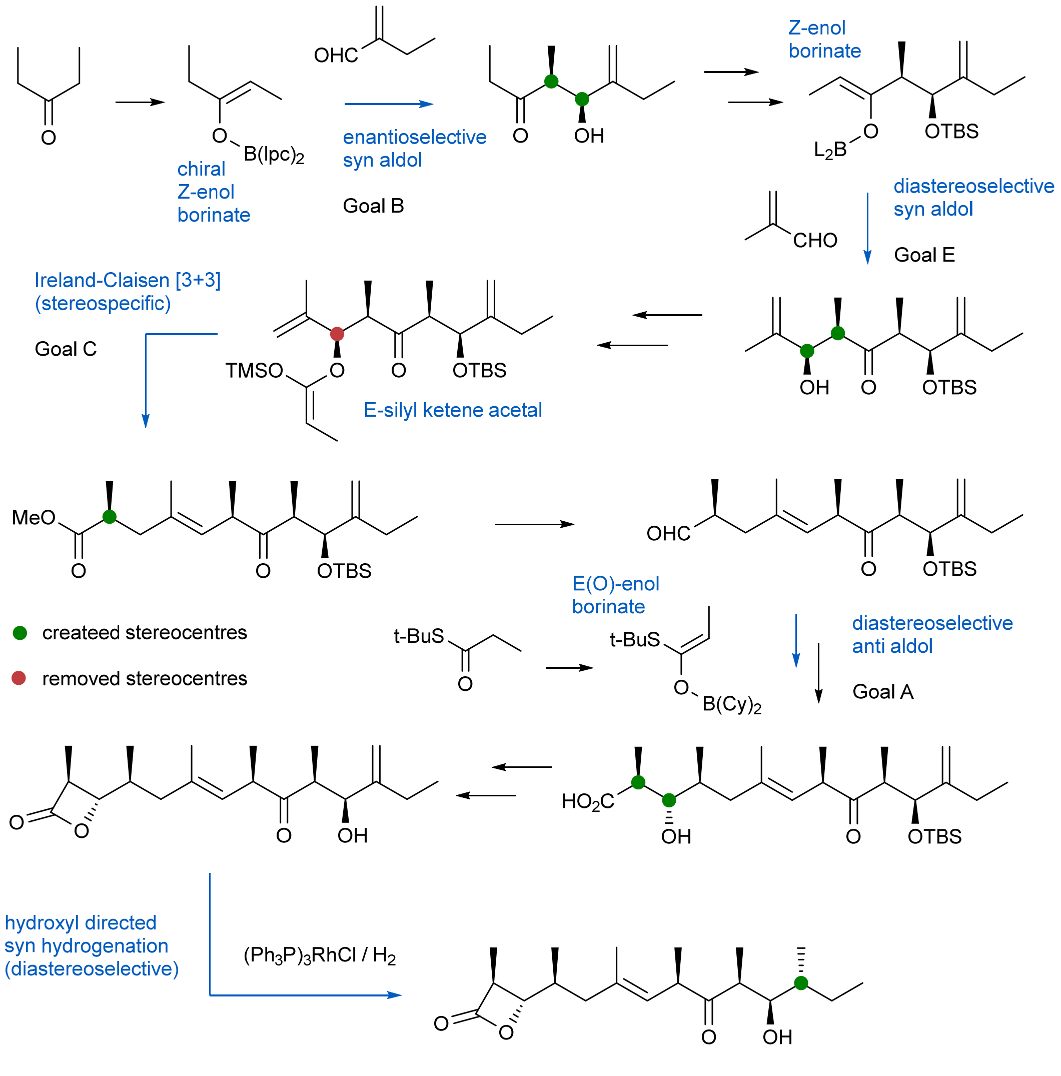


Figure 139 The key stereocontrolled steps in Paterson and Hulme's enantioselective synthesis of (-)-Ebelactone.

plan due to the lack of a suitable strategy pattern[d]. Synthetically this stereocentre was introduced in the closing step of the synthesis utilising a hydroxyl directed rhodium catalysed hydrogenation of the allylic alcohol precursor.[298] In the common sections of the theoretical and practical strategies, the synthesis was solved employing a sequence of two boron enolate mediated *syn* aldol reactions (goals B and E) followed by an Ireland-Claisen rearrangement (goal C) and then by a further boron enolate mediated *anti* aldol reaction (goal A) before the introduction of the final stereocentre by the hydroxyl directed hydrogenation step. The enantiopurity of Ebelactone was established *via* the first *syn* aldol reaction between diethyl ketone and 2-ethylacrolein with a reported stereoselectivity of ≥97% de and 86% ee.

This synthesis corresponds to the retrosynthetic goal satisfaction order A, C, E and B. Paterson and Hulme also investigated the alternative goal satisfaction sequence A, C, B and E but found that diastereoselective control in the second *syn* aldol step was good but not as effective as the preferred sequence.[359]

[d] This can be rectified by adding a new strategy pattern to cover the required transformation.

References

[1] Hendrickson, J. B.; Braun-Keller, E.; Toczko, G. A. A logic for synthesis design., *Tetrahedron*, **1981**, *37*, 359–370.

[2] Corey, E. J. General methods for the construction of complex molecules., *Pure Appl. Chem.*, **1967**, *14*, 19–38.

[3] Isler, O.; Ronco, A.; Guex, W.; Hindley, N. C.; Huber, W.; Dialer, K.; Kofler, M. Uber die Ester und Ather des synthetischen Vitamins A., *Helv. Chim. Acta*, **1949**, *32*, 489–505.

[4] Woodward, R. B.; Sondheimer, F.; Taub, D. The Total Synthesis of Cortisone., *J. Am. Chem. Soc.*, **1951**, *73*, 4057.

[5] Gates, M.; Tschudi, G. The Synthesis of Morphine., *J. Am. Chem. Soc.*, **1956**, *78*, 1380–1393.

[6] Woodward, R. B.; Ayer, W. A.; Beaton, J. M.; Bickelhaupt, F.; Bonnett, R.; Buchschacher, P.; Closs, G. L.; Dutler, H.; Hannah, J.; Hauck, F. P.; Itô, S.; Langemann, A.; Le Goff, E.; Leimgruber, W.; Lwowski, W.; Sauer, J.; Valenta, Z.; Volz, H. The Total Synthesis of Chlorophyll., *J. Am. Chem. Soc.*, **1960**, *82*, 3800–3802.

[7] Corey, E. J. The Logic of Chemical Synthesis: Multistep Synthesis of Complex Carbogenic Molecules (Nobel Lecture)., *Angew. Chem. Int. Ed. Engl.*, **1991**, *30*, 455–465.

[8] Corey, E. J.; Ohno, M.; Vatakencherry, P. A.; Mitra, R. B. Total Synthesis of d,l-Longifolene., *J. Am. Chem. Soc.*, **1961**, *83*, 1251–1253.

[9] Wyatt, P.; Warren, S. *Organic Synthesis: Strategy and Control*; Wiley, **2007**.

[10] Warren, S.; Wyatt, P. *Organic Synthesis: The Disconnection Approach*; 2nd Ed.; Wiley, **2008**.

[11] Corey, E.; Cheng, X.-M. *The logic of chemical synthesis*; John Wiley: New York, **1989**.

[12] Corey, E. J.; Wipke, W. T. Computer-Assisted Design of Complex Organic Syntheses., *Science*, **1969**, *166*, 178–192.

[13] Corey, E. J.; Wipke, W. T.; Cramer, R. D.; Howe, W. J. Techniques for perception by a computer of synthetically significant structural features in complex molecules., *J. Am. Chem. Soc.*, **1972**, *94*, 431–439.

[14] Ugi, I.; Bauer, J.; Blomberger, C.; Brandt, J.; Dietz, A.; Fontain, E.; Gruber, B.; v. Scholley-Pfab, A.; Senff, A.; Stein, N. Models, concepts, theories, and formal languages in chemistry and their use as a basis for computer assistance in chemistry., *J. Chem. Inf. Comput. Sci.*, **1994**, *34*, 3–16.

[15] Gasteiger, J.; Ihlenfeldt, W. D.; Röse, P.; Wanke, R. Computer-assisted reaction prediction and synthesis design., *Anal. Chim. Acta*, **1990**, *235*, 65–75.

[16] Hendrickson, J. B.; Parks, C. A. A program for the FORWARD generation of synthetic routes., *J. Chem. Inf. Comput. Sci.*, **1992**, *32*, 209–215.

[17] Todd, M. H. Computer-aided organic synthesis., *Chem. Soc. Rev.*, **2005**, *34*, 247.

[18] Salatin, T. D.; Jorgensen, W. L. Computer-assisted mechanistic evaluation of organic reactions. 1. Overview., *J. Org. Chem.*, **1980**, *45*, 2043–2051.

[19] Hanessian, S.; Franco, J.; Gagnon, G.; Laramee, D.; Larouche, B. Computer-assisted analysis and perception of stereochemical features in organic molecules using the CHIRON program., *J. Chem. Inf. Comput. Sci.*, **1990**, *30*, 413–425.

[20] Hanessian, S.; Botta, M.; Larouche, B.; Boyaroglu, A. Computer-assisted perception of similarity using the Chiron program: a powerful tool for the analysis and prediction of biogenetic patterns., *J. Chem. Inf. Comput. Sci.*, **1992**, *32*, 718–722.

[21] Wipke, W. T.; Rogers, D. Artificial intelligence in organic synthesis. SST: starting material selection strategies. An application of superstructure search., *J. Chem. Inf. Comput. Sci.*, **1984**, *24*, 71–81.

[22] Socorro, I. M.; Taylor, K.; Goodman, J. M. ROBIA: A Reaction Prediction Program., *Org. Lett.*, **2005**, *7*, 3541–3544.

[23] Socorro, I. M.; Goodman, J. M. The ROBIA Program for Predicting Organic Reactivity., *J. Chem. Inf. Model.*, **2006**, *46*, 606–614.

[24] Chen, J. H.; Baldi, P. No Electron Left Behind: A Rule-Based Expert System To Predict Chemical Reactions and Reaction Mechanisms., *J. Chem. Inf. Model.*, **2009**, *49*, 2034–2043.

[25] Ihlenfeldt, W.-D.; Gasteiger, J. Computer-Assisted Planning of Organic Syntheses: The Second Generation of Programs., *Angew. Chem. Int. Ed. Engl.*, **1996**, *34*, 2613–2633.

[26] Zefirov, N. S.; Gordeeva, E. V. Computer-assisted Synthesis., *Russ. Chem. Rev.*, **1987**, *56*, 1002 – 1012.

[27] Ugi, I.; Bauer, J.; Bley, K.; Dengler, A.; Dietz, A.; Fontain, E.; Gruber, B.; Herges, R.; Knauer, M.; Reitsam, K.; Stein, N. Computer-Assisted Solution of Chemical Problems—The Historical Development and the Present State of the Art of a New Discipline of Chemistry., *Angew. Chem. Int. Ed. Engl.*, **1993**, *32*, 201–227.

[28] Corey, E.; Long, A.; Rubenstein, S. Computer-assisted analysis in organic synthesis., *Science*, **1985**, *228*, 408–418.

[29] Wipke, W. T.; Braun, H.; Smith, G.; Choplin, F.; Sieber, W. SECS-Simulation and Evaluation of Chemical Synthesis - Strategy and Planning. In *Computer-Assisted Organic Synthesis*; ACS Symposium Series 61; American Chemical Society: Washington, DC, **1977**; Vol. 61, pp. 97–127.

[30] Hendrickson, J. B. Organic Synthesis in the Age of Computers., *Angew. Chem. Int. Ed. Engl.*, **1990**, *29*, 1286–1295.

[31] Hendrickson, J. B. The SYNGEN approach to synthesis design., *Anal. Chim. Acta*, **1990**, *235*, 103–113.

[32] Hendrickson, J. B.; Toczko, A. G. Systematic synthesis design: the SYNGEN program., *Pure Appl. Chem.*, **1989**, *61*, 589–592.

[33] Hendrickson, J. B.; Toczko, A. G. Synthesis design logic and the SYNGEN (synthesis generation) program., *Pure Appl. Chem.*, **1988**, *60*, 1563–1572.

[34] Gelernter, H. L.; Sanders, A. F.; Larsen, D. L.; Agarwal, K. K.; Boivie, R. H.; Spritzer, G. A.; Searleman, J. E. Empirical Explorations of SYNCHEM., *Science*, **1977**, *197*, 1041–1049.

[35] Hendrickson, J. B.; Miller, T. M. Reaction classification and retrieval. A linkage between synthesis generation and reaction databases., *J. Am. Chem. Soc.*, **1991**, *113*, 902–910.

[36] Gelernter, H.; Rose, J. R.; Chen, C. Building and refining a knowledge base for synthetic organic chemistry via the methodology of inductive and deductive machine learning., *J. Chem. Inf. Comput. Sci.*, **1990**, *30*, 492–504.

[37] Blurock, E. S. Computer-aided synthesis design at RISC-Linz: automatic extraction and use of reaction classes., *J. Chem. Inf. Comput. Sci.*, **1990**, *30*, 505–510.

[38] Wang, K.; Wang, L.; Yuan, Q.; Luo, S.; Yao, J.; Yuan, S.; Zheng, C.; Brandt, J. Construction of a generic reaction knowledge base by reaction data mining., *J. Mol. Graph. Model.*, **2001**, *19*, 427–433.

[39] Satoh, K.; Funatsu, K. A Novel Approach to Retrosynthetic Analysis Using Knowledge Bases Derived from Reaction Databases., *J. Chem. Inf. Comput. Sci.*, **1999**, *39*, 316–325.

[40] Law, J.; Zsoldos, Z.; Simon, A.; Reid, D.; Liu, Y.; Khew, S. Y.; Johnson, A. P.; Major, S.; Wade, R. A.; Ando, H. Y. Route Designer: A Retrosynthetic Analysis Tool Utilizing Automated Retrosynthetic Rule Generation., *J. Chem. Inf. Model.*, **2009**, *49*, 593–602.

[41] Barnard, J. M. Substructure searching methods: Old and new., *J. Chem. Inf. Comput. Sci.*, **1993**, *33*, 532–538.

[42] Myatt, G. J. *Computer Aided Estimation of Synthetic Accessibility*; Ph. D. Thesis.; The University of Leeds, **1994**.

[43] Corey, E. J.; Howe, W. J.; Orf, H. W.; Pensak, D. A.; Petersson, G. General methods of synthetic analysis. Strategic bond disconnections for bridged polycyclic structures., *J. Am. Chem. Soc.*, **1975**, *97*, 6116–6124.

[44] Corey, E. J.; Jorgensen, W. L. Computer-assisted synthetic analysis. Synthetic strategies based on appendages and the use of reconnective transforms., *J. Am. Chem. Soc.*, **1976**, *98*, 189–203.

[45] Corey, E. J.; Petersson, G. A. Algorithm for machine perception of synthetically significant rings in complex cyclic organic structures., *J. Am. Chem. Soc.*, **1972**, *94*, 460–465.

[46] Paton, K. An algorithm for finding a fundamental set of cycles of a graph., *Commun ACM*, **1969**, *12*, 514–518.

[47] Roos-Kozel, B. L.; Jorgensen, W. L. Computer-assisted mechanistic evaluation of organic reactions. 2. Perception of rings, aromaticity, and tautomers., *J. Chem. Inf. Comput. Sci.*, **1981**, *21*, 101–111.

[48] Plotkin, M. Mathematical Basis of Ring-Finding Algorithms in CIDS., *J. Chem. Doc.*, **1971**, *11*, 60–63.

[49] Vismara, P. Union of all the minimum cycle bases of a graph., *Electron J Comb*, **1997**, *4*, 73–87.

[50] Balducci, R.; Pearlman, R. S. Efficient exact solution of the ring perception problem., *J. Chem. Inf. Comput. Sci.*, **1994**, *34*, 822–831.

[51] Sysło, M. M. An Efficient Cycle Vector Space Algorithm for Listing All Cycles of a Planar Graph., *SIAM J. Comput.*, **1981**, *10*, 797.

[52] Downs, G. M.; Gillet, V. J.; Holliday, J. D.; Lynch, M. F. Theoretical aspects of ring perception and development of the extended set of smallest rings concept., *J. Chem. Inf. Comput. Sci.*, **1989**, *29*, 187–206.

[53] Berger, F.; Flamm, C.; Gleiss, P. M.; Leydold, J.; Stadler, P. F. Counterexamples in Chemical Ring Perception., *J. Chem. Inf. Comput. Sci.*, **2004**, *44*, 323–331.

[54] Downs, G. M.; Gillet, V. J.; Holliday, J. D.; Lynch, M. F. Review of ring perception algorithms for chemical graphs., *J. Chem. Inf. Comput. Sci.*, **1989**, *29*, 172–187.

[55] Hopkinson, G. A. *Computer-Assisted Organic Synthesis Design*; Ph. D. Thesis.; The University of Leeds, **1985**.

[56] Hoyle, P. L. M. *Computer Assisted Design of Organic Synthesis*; Ph. D. Thesis.; The University of Leeds.

[57] Dewar, M. J. S.; Gleicher, G. J. Ground States of Conjugated Molecules. II. Allowance for Molecular Geometry., *J. Am. Chem. Soc.*, **1965**, *87*, 685–692.

[58] Tumber, H. S. *Computer Assisted Synthesis Design - Selected Heterocyclic Systems*; Ph. D. Thesis.; The University of Leeds, **1985**.

[59] Seghal, S. K. *Computer Assisted Design of Organic Synthesis*; Ph. D. Thesis.; The University of Leeds, **1990**.

[60] Hanessian, S. *Total Synthesis of Natural Products: The "Chiron" Approach*; Pergamon Press: Elmsford. NY, **1983**.

[61] Gasteiger, J. *Handbook of Chemoinformatics*; Wiley-VCH, **2008**; Vol. 1.

[62] Wipke, W. T.; Dyott, T. M. Simulation and evaluation of chemical synthesis. Computer representation and manipulation of stereochemistry., *J. Am. Chem. Soc.*, **1974**, *96*, 4825–4834.

[63] Corey, E. J.; Long, A. K. Computer-assisted synthetic analysis. Performance of long-range strategies for stereoselective olefin synthesis., *J. Org. Chem.*, **1978**, *43*, 2208–2216.

[64] Blackwood, J. E.; Blower, P. E.; Layten, S. W.; Lillie, D. H.; Lipkus, A. H.; Peer, J. P.; Qian, C.; Staggenborg, L. M.; Watson, C. E. Chemical Abstracts Service Chemical Registry System. 13. Enhanced handling of stereochemistry., *J. Chem. Inf. Comput. Sci.*, **1991**, *31*, 204–212.

[65] Gunther, B.; Zugge, J. On the Recognition of Composed Systems of Stereocenters in Molecular Graph Theory by Wreath Products., *J. Comput. Chem. Jpn.*, **2007**, *6*, 235–244.

[66] Morgan, H. L. The Generation of a Unique Machine Description for Chemical Structures-A Technique Developed at Chemical Abstracts Service., *J. Chem. Doc.*, **1965**, *5*, 107–113.

[67] Freeland, R. G.; Funk, S. A.; O'Korn, L. J.; Wilson, G. A. The Chemical Abstracts Service Chemical Registry System. II. Augmented Connectivity Molecular Formula., *J. Chem. Inf. Comput. Sci.*, **1979**, *19*, 94–98.

[68] Wipke, W. T.; Dyott, T. M. Stereochemically unique naming algorithm., *J. Am. Chem. Soc.*, **1974**, *96*, 4834–4842.

[69] Wipke, W. T.; Krishnan, S.; Ouchi, G. I. Hash Functions for Rapid Storage and Retrieval of Chemical Structures., *J. Chem. Inf. Comput. Sci.*, **1978**, *18*, 32–37.

[70] Figueras, J. Morgan revisited., *J. Chem. Inf. Comput. Sci.*, **1993**, *33*, 717–718.

[71] Hodgson, R.; Nelson, A. A two-directional synthesis of the C58-C71 fragment of palytoxin., *Org. Biomol. Chem.*, **2004**, *2*, 373.

[72] Poss, C. S.; Schreiber, S. L. Two-directional chain synthesis and terminus differentiation., *Acc. Chem. Res.*, **1994**, *27*, 9–17.

[73] Magnuson, S. R. Two-directional synthesis and its use in natural product synthesis., *Tetrahedron*, **1995**, *51*, 2167–2213.

[74] Anstiss, M.; Holland, J. M.; Nelson, A.; Titchmarsh, J. R. Beyond Breaking the MirrorPlane: The Desymmetrisation of Centro-symmetric Molecules as an Efficient Strategy for Asymmetric Synthesis., *Synlett*, **2003**, 1213–1220.

[75] Garcia-Urdiales, E.; Alfonso, I.; Gotor, V. Enantioselective Enzymatic Desymmetrizations in Organic Synthesis., *Chem. Rev.*, **2005**, *105*, 313–354.

[76] Hoffmann, R. W. meso Compounds: Stepchildren or Favored Children of Stereoselective Synthesis?, *Angew. Chem. Int. Ed.*, **2003**, *42*, 1096–1109.

[77] Agarwal, K. K. An Algorithm for Computing the Automorphism Group of Organic Structures with Stereochemistry and a Measure of its Efficiency., *J. Chem. Inf. Comput. Sci.*, **1998**, *38*, 402–404.

[78] Shelley, C. A.; Munk, M. E. An Approach to the Assignment of Canonical Connection Tables and Topological Symmetry Perception., *J. Chem. Inf. Comput. Sci.*, **1979**, *19*, 247–250.

[79] Jochum, C.; Gasteiger, J. Canonical Numbering and Constitutional Symmetry., *J. Chem. Inf. Comput. Sci.*, **1977**, *17*, 113–117.

[80] Fan, B. T.; Barbu, A.; Panaye, A.; Doucet, J.-P. Detection of Constitutionally Equivalent Sites from a Connection Table., *J. Chem. Inf. Comput. Sci.*, **1996**, *36*, 654–659.

[81] Fan, B. T.; Panaye, A.; Doucet, J.-P. Comment on “Isomorphism, Automorphism Partitioning, and Canonical Labeling Can Be Solved in Polynomial-Time for Molecular Graphs.”, *J. Chem. Inf. Comput. Sci.*, **1999**, *39*, 630–631.

[82] Carhart, R. E. Erroneous Claims Concerning the Perception of Topological Symmetry., *J. Chem. Inf. Comput. Sci.*, **1978**, *18*, 108–110.

[83] Jochum, C.; Gasteiger, J. On the Misinterpretation of Our Algorithm for the Perception of Constitutional Symmetry., *J. Chem. Inf. Comput. Sci.*, **1979**, *19*, 49–50.

[84] Carhart, R. Letter to the Editor. Perception of Topological Symmetry., *J. Chem. Inf. Comput. Sci.*, **1979**, *19*, 56.

[85] Ruecker, G.; Ruecker, C. Computer perception of constitutional (topological) symmetry: TOPSYM, a fast algorithm for partitioning atoms and pairwise relations among atoms into equivalence classes., *J. Chem. Inf. Comput. Sci.*, **1990**, *30*, 187–191.

[86] Figueras, J. Automorphism and equivalence classes., *J. Chem. Inf. Comput. Sci.*, **1992**, *32*, 153–157.

[87] Röse, P.; Gasteiger, J. Automated derivation of reaction rules for the EROS 6.0 system for reaction prediction., *Anal. Chim. Acta*, **1990**, *235*, 163–168.

[88] Wipke, W. T.; Gund, P. Simulation and evaluation of chemical synthesis. Congestion: a conformation-dependent function of steric environment at a reaction center. Application with torsional terms to stereoselectivity of nucleophilic additions to ketones., *J. Am. Chem. Soc.*, **1976**, *98*, 8107–8118.

[89] Corey, E. J.; Cramer, R. D.; Howe, W. J. Computer-assisted synthetic analysis for complex molecules. Methods and procedures for machine generation of synthetic intermediates., *J. Am. Chem. Soc.*, **1972**, *94*, 440–459.

[90] Orf, H. W. *Computer-Assisted Synthetic Analysis*; Ph. D. Thesis.; Harvard University: Cambridge, Massachusetts, **1976**.

[91] Kappos, J. *Computer-Assisted Synthetic Analysis: Tactical Combinations of Transforms and a Generalised Procedure for Subgoal Transform Selection*; Ph. D. Thesis.; Harvard University: Cambridge, Massachusetts, **1991**.

[92] Long, A. K.; Kappos, J. C. Computer-assisted synthetic analysis. Performance of tactical combinations of transforms., *J. Chem. Inf. Comput. Sci.*, **1994**, *34*, 915–921.

[93] Fontain, E.; Reitsam, K. The generation of reaction networks with RAIN. 1. The reaction generator., *J. Chem. Inf. Comput. Sci.*, **1991**, *31*, 96–101.

[94] Ugi, I.; Dengler, A. The Algebraic and Graph Theoretical Completion of Truncated Reaction Equations., *J. Math. Chem.*, **1992**, *9*, 1–10.

[95] Gasteiger, J.; Jochum, C. EROS - a computer program for generating sequences of reactions., *Top. Curr. Chem.*, **1978**, *74*, 93–126.

[96] Hendrickson, J. B. Systematic synthesis design. IV. Numerical codification of construction reactions., *J. Am. Chem. Soc.*, **1975**, *97*, 5784–5800.

[97] Ernst, G.; Newell, A. *GPS: A case Study in generality and Problem Solving*; Academic Press: N. Y., **1969**.

[98] Wipke, W. T.; Ouchi, G. I.; Krishnan, S. Simulation and evaluation of chemical synthesis--SECS: An application of artificial intelligence techniques., *Artif. Intell.*, **1978**, *11*, 173–193.

[99] Corey, E. J.; Wipke, W. T.; Cramer, R. D.; Howe, W. J. Computer-assisted synthetic analysis. Facile man-machine communication of chemical structure by interactive computer graphics., *J. Am. Chem. Soc.*, **1972**, *94*, 421–430.

[100] Corey, E. J.; Jorgensen, W. L. Computer-assisted synthetic analysis. Generation of synthetic sequences involving sequential functional group interchanges., *J. Am. Chem. Soc.*, **1976**, *98*, 203–209.

[101] Gnas, Y.; Glorius, F. Chiral Auxiliaries - Principles and Recent Applications., *Synthesis*, **2006**, *2006*, 1899–1930.

[102] Paquette, L. *Chiral reagents for asymmetric synthesis*; Wiley: Chichester, **2003**.

[103] Ojima, I. *Catalytic asymmetric synthesis*; 2nd ed.; Wiley-VCH: New York, **2000**.

[104] Corey, E. J.; Howe, W. J.; Pensak, D. A. Computer-assisted synthetic analysis. Methods for machine generation of synthetic intermediates involving multistep look-ahead., *J. Am. Chem. Soc.*, **1974**, *96*, 7724–7737.

[105] Ott, M. A.; Noordik, J. H. Long-Range Strategies in the LHASA Program: The Quinone Diels−Alder Transform., *J. Chem. Inf. Comput. Sci.*, **1997**, *37*, 98–108.

[106] Corey, E. J.; Long, A. K.; Mulzer, J.; Orf, H. W.; Johnson, A. P.; Hewett, A. P. W. Computer-Assisted Synthetic Analysis. Long-Range Search Procedures for Antithetic Simplification of Complex Targets by Application of the Halolactonization Transform., *J. Chem. Inf. Comput. Sci.*, **1980**, *20*, 221–230.

[107] Corey, E. J.; Johnson, A. P.; Long, A. K. Computer-assisted synthetic analysis. Techniques for efficient long-range retrosynthetic searches applied to the Robinson annulation process., *J. Org. Chem.*, **1980**, *45*, 2051–2057.

[108] Corey, E. J.; Long, A. K. Computer-assisted synthetic analysis. Performance of long-range strategies for stereoselective olefin synthesis., *J. Org. Chem.*, **1978**, *43*, 2208–2216.

[109] Hendrickson, J. B. Systematic synthesis design. 6. Yield analysis and convergency., *J. Am. Chem. Soc.*, **1977**, *99*, 5439–5450.

[110] Johnson, A. P.; Marshall, C.; Judson, P. N. Starting material oriented retrosynthetic analysis in the LHASA program. 1. General description., *J. Chem. Inf. Comput. Sci.*, **1992**, *32*, 411–417.

[111] Johnson, A. P.; Marshall, C. Starting material oriented retrosynthetic analysis in the LHASA program. 2. Mapping the SM and target structures., *J. Chem. Inf. Comput. Sci.*, **1992**, *32*, 418–425.

[112] Johnson, A. P.; Marshall, C. Starting material oriented retrosynthetic analysis in the LHASA program. 3. Heuristic estimation of synthetic proximity., *J. Chem. Inf. Comput. Sci.*, **1992**, *32*, 426–429.

[113] Hart, P. E.; Nilsson, N. J.; Raphael, B. A Formal basis for the Heuristic Determination of Minimum Cost Paths., *IEEE Trans. Syst. Sci. Cybern.*, **1968**, *4*.

[114] Nilsson, N. J. *Principles of Artificial Intelligence*; Tioga Publishing Company: Palo Alto.

[115] Judea, P. *Heuristics: Intelligent Search Strategies for Computer Problem Solving*; Addison-Wesley, **1984**.

[116] Russell, S. J.; Norvig, P. *Artificial Intelligence: A Modern Approach*; Prentice Hall, **2002**.

[117] Corey, E. J.; Cheng, X.-M. *The Logic of Chemical Synthesis*; Wiley-Interscience: New York, **1989**.

[118] Coelho, F.; Deprés, J.-P.; Brocksom, T. J.; Greene, A. E. Direct approach to the bakkanes: A synthesis of (±)-homogynolide-B., *Tetrahedron Lett.*, **1989**, *30*, 565–566.

[119] Srikrishna, A.; Nagaraju, S.; Venkateswarlu, S.; Hiremath, U. S.; Reddy, T. J.; Venugopalan, P. A formal total synthesis of (±)-homogynolide-B., *J. Chem. Soc. [Perkin 1]*, **1999**, 2069–2076.

[120] Hendrickson, J. B.; Braun-Keller, E. Systematic synthesis design. 8. Generation of reaction sequences., *J. Comput. Chem.*, **1980**, *1*, 323–333.

[121] Lee, T. V. Expert systems in synthesis planning: A user’s view of the LHASA program., *Chemom. Intell. Lab. Syst.*, **1987**, *2*, 259–272.

[122] Gund, P.; Grabowski, E. J. J.; Hoff, D. R.; Smith, G. M.; Andose, J. D.; Rhodes, J. B.; Wipke, W. T. Computer-Assisted Synthetic Analysis at Merck., *J. Chem. Inf. Comput. Sci.*, **1980**, *20*, 88–93.

[123] Judson, P. N. *Knowledge-Based Expert Systems in Chemistry: Not Counting on Computers*; RSC Theoretical and Computational Chemistry Series; RSC Publishing, **2009**.

[124] Choplin, F.; Marc, R.; Kaufmann, G.; Wipke, W. T. Computer Design of Synthesis in Phosphorus Chemistry: Automatic Treatment of Stereochemistry., *J. Chem. Inf. Comput. Sci.*, **1978**, *18*, 110–118.

[125] Choplin, F. Computer graphics determination and display of stereoisomers in coordination compounds., *J. Organomet. Chem.*, **1978**, *152*, 101–109.

[126] Fernández, I.; Khiar, N. Recent Developments in the Synthesis and Utilization of Chiral Sulfoxides., *Chem. Rev.*, **2003**, *103*, 3651–3706.

[127] Morton, D.; Stockman, R. A. Chiral non-racemic sulfinimines: versatile reagents for asymmetric synthesis., *Tetrahedron*, **2006**, *62*, 8869–8905.

[128] Pyne, S. G. Asymmetric conjugate addition of organometallic reagents to chiral vinyl sulfoximines., *J. Org. Chem.*, **1986**, *51*, 81–87.

[129] Pyne, S. G.; Dong, Z.; Skelton, B. W.; White, A. H. Cyclopropanation Reactions of Enones with Lithiated Sulfoximines: Application to the Asymmetric Synthesis of Chiral Cyclopropanes., *J. Org. Chem.*, **1997**, *62*, 2337–2343.

[130] Hoffmann-Röder, A.; Krause, N. Synthesis and Properties of Allenic Natural Products and Pharmaceuticals., *Angew. Chem. Int. Ed.*, **2004**, *43*, 1196–1216.

[131] Walkup, R. D.; Kim, S. W. Syntheses of the Lower Portions of the Pamamycins from .gamma.-(Silyloxy)allenes Using Stereoselective Cyclization, Reduction, and Aldehyde Addition Methodologies., *J. Org. Chem.*, **1994**, *59*, 3433–3441.

[132] Hiroi, K.; Hiratsuka, Y.; Watanabe, K.; Abe, I.; Kato, F.; Hiroi, M. A novel direct catalytic asymmetric synthesis of cyclic indole derivatives by intramolecular carbopalladation of allenes and subsequent intramolecular amination., *Tetrahedron Asymmetry*, **2002**, *13*, 1351–1353.

[133] Chavarot, M.; Ménage, S.; Hamelin, O.; Charnay, F.; Pécaut, J.; Fontecave, M. “Chiral-at-Metal” Octahedral Ruthenium(II) Complexes with Achiral Ligands: A New Type of Enantioselective Catalyst., *Inorg. Chem.*, **2003**, *42*, 4810–4816.

[134] Rzepa, H. S.; Cass, M. E. In Search of the Bailar and Rây−Dutt Twist Mechanisms That Racemize Chiral Trischelates: A Computational Study of ScIII, TiIV, CoIII, ZnII, GaIII, and GeIV Complexes of a Ligand Analogue of Acetylacetonate., *Inorg. Chem.*, **2007**, *46*, 8024–8031.

[135] Togni, A. Planar-Chiral Ferrocenes: Synthetic Methods and Applications., *Angew. Chem. Int. Ed. Engl.*, **1996**, *35*, 1475–1477.

[136] Gómez Arrayás, R.; Adrio, J.; Carretero, J. C. Recent Applications of Chiral Ferrocene Ligands in Asymmetric Catalysis., *Angew. Chem. Int. Ed.*, **2006**, *45*, 7674–7715.

[137] Wasserman, E. The Preparation of Interlocking Rings: A Catenane., *J. Am. Chem. Soc.*, **1960**, *82*, 4433–4434.

[138] Herges, R. Topology in Chemistry: Designing Möbius Molecules†., *Chem. Rev.*, **2006**, *106*, 4820–4842.

[139] Lukin, O.; Vögtle, F. Knotting and Threading of Molecules: Chemistry and Chirality of Molecular Knots and Their Assemblies., *Angew. Chem. Int. Ed.*, **2005**, *44*, 1456–1477.

[140] Perret-Aebi, L.-E.; von Zelewsky, A.; Dietrich-Buchecker, C.; Sauvage, J.-P. Stereoselective Synthesis of a Topologically Chiral Molecule: The Trefoil Knot., *Angew. Chem. Int. Ed.*, **2004**, *43*, 4482–4485.

[141] Cahn, R. S.; Ingold, C.; Prelog, V. Specification of Molecular Chirality., *Angew. Chem. Int. Ed. Engl.*, **1966**, *5*, 385–415.

[142] Cahn, R. S.; Ingold, C. K.; Prelog, V. Errata: Specification of Molecular Chirality., *Angew. Chem. Int. Ed. Engl.*, **1966**, *5*, 511–511.

[143] Petrarca, A. E.; Lynch, M. F.; Rush, J. E. A Method for Generating Unique Computer Structural Representations of Stereoisomers., *J. Chem. Doc.*, **1967**, *7*, 154–165.

[144] Weininger, D. SMILES, a chemical language and information system. 1. Introduction to methodology and encoding rules., *J. Chem. Inf. Comput. Sci.*, **1988**, *28*, 31–36.

[145] Brecher, J. Graphical representation standards for chemical structure diagrams (IUPAC Recommendations 2008)., *Pure Appl. Chem.*, **2008**, *80*.

[146] Lin, S.-K. A proposal for the representation of the stereochemistry of quadrivalent centers., *Chirality*, **1992**, *4*, 274–278.

[147] Mata, P.; Lobo, A. M.; Marshall, C.; Johnson, A. P. Implementation of the Cahn-Ingold-Prelog System for Stereochemical Perception in the LHASA Program., *J. Chem. Inf. Comput. Sci.*, **1994**, *34*, 491–504.

[148] Dodziuk, H.; Mirowicz, M. A proposal for a modification of the Cahn, Ingold and Prelog classification of chirality., *Tetrahedron Asymmetry*, **1990**, *1*, 171–186.

[149] Perdih, M.; Razinger, M. Stereochemistry and sequence rules a proposal for modification of Cahn-Ingold-Prelog system., *Tetrahedron Asymmetry*, **1994**, *5*, 835–861.

[150] Petrarca, A. E.; Rush, J. E. Methods for Computer Generation of Unique Configurational Descriptors for Stereoisomeric Square-Planar and Octahedral Complexes., *J. Chem. Doc.*, **1969**, *9*, 32–37.

[151] Allen, F. H.; Barnard, J. M.; Cook, A. P. F.; Hall, S. R. The Molecular Information File (MIF): Core Specifications of a New Standard Format for Chemical Data., *J. Chem. Inf. Comput. Sci.*, **1995**, *35*, 412–427.

[152] Peishoff, C. E.; Jorgensen, W. L. Computer-assisted mechanistic evaluation of organic reactions. 10. Stereochemistry., *J. Org. Chem.*, **1985**, *50*, 3174–3184.

[153] Gasteiger, J. A Representation of π Systems for Efficient Computer Manipulation., *J. Chem. Inf. Comput. Sci.*, **1979**, *19*, 111–115.

[154] Warr, W. *Chemical structures 2 : the international language of chemistry : proceedings of the second international conference, Leeuwenhorst Congress Center, Noordwijkerhout, the Netherlands, 3rd June to 7th*; Springer-Verlag: Berlin ;;New York, **1993**.

[155] Dietz, A. Yet Another Representation of Molecular Structure., *J. Chem. Inf. Comput. Sci.*, **1995**, *35*, 787–802.

[156] Rohde, B. *GM-Search: A System for Stereochemical Substructure Search. Inaugural Dissertation*; Ph. D. Thesis.; University of Zurich: Zurich, **1988**.

[157] Chen, L.; Nourse, J. G.; Christie, B. D.; Leland, B. A.; Grier, D. L. Over 20 Years of Reaction Access Systems from MDL: A Novel Reaction Substructure Search Algorithm., *J Chem Inf Comput Sci*, **2002**, *42*, 1296–1310.

[158] Moock, T. E.; Nourse, J. G.; Grier, D.; Hounshell, W. D. The Implementation of AAM and Related Reaction Features in the Reaction Access System (REACCS). In *Chemical Structures, The International Language of Chemistry*; Springer-Verlag: Noordwijkerhout, The Netherlands, **1988**; pp. 303 – 313.

[159] Donated for research purposes by Accelrys, Inc. 10188 Telesis Court, Suite 100, San Diego, CA 92121, USA.

[160] Supplied by FIZ CHEMIE, Franklinstr. 11, D-10587 Berlin.

[161] Lynch, M. F.; Willett, P. The Automatic Detection of Chemical Reaction Sites., *J. Chem. Inf. Comput. Sci.*, **1978**, *18*, 154–159.

[162] McGregor, J. J.; Willett, P. Use of a maximum common subgraph algorithm in the automatic identification of ostensible bond changes occurring in chemical reactions., *J. Chem. Inf. Comput. Sci.*, **1981**, *21*, 137–140.

[163] Jochum, C.; Gasteiger, J.; Ugi, I. The Principle of Minimum Chemical Distance (PMCD)., *Angew. Chem. Int. Ed.*, **1980**, *19*, 495–505.

[164] Vladutz, G. Do We Still Need a Classification orf Reactions. In *Modern Approaches to Chemical Reaction Searching*; Gower: University of York, **1985**; pp. 202 – 220.

[165] Fujita, S. Description of organic reactions based on imaginary transition structures. 1. Introduction of new concepts., *J. Chem. Inf. Comput. Sci.*, **1986**, *26*, 205–212.

[166] Vladutz, G.; Gould, S. R. Joint Compound/Reaction Storage and Retrieval and Possibilities of a Hyperstructure-based Solution. In *Chemical Structures: The International Language of Chemistry*; Springer-Verlag: Noordwijkerhout, The Netherlands, **1988**; pp. 371 – 384.

[167] Fujita, S. Canonical numbering and coding of imaginary transition structures. A novel approach to the linear coding of individual organic reactions., *J. Chem. Inf. Comput. Sci.*, **1988**, *28*, 128–137.

[168] Randić, M. On Unique Numbering of Atoms and Unique Codes for Molecular Graphs., *J Chem Inf Comput Sci*, **1975**, *15*, 105–108.

[169] Randić, M. On the recognition of identical graphs representing molecular topology., *J. Chem. Phys.*, **1974**, *60*, 3920.

[170] Moreau, G. Topological Code for Molecular Structures: A Modified Morgan Algorithm., *Nouv J Chim*, **1980**, *4*, 17–22.

[171] Ihlenfeldt, W. D.; Gasteiger, J. Hash codes for the identification and classification of molecular structure elements., *J. Comput. Chem.*, **1994**, *15*, 793–813.

[172] Willett, P. An algorithm for chemical superstructure searching., *J. Chem. Inf. Comput. Sci.*, **1985**, *25*, 114–116.

[173] Willett, P. A Screen Set Generation Algorithm., *J. Chem. Inf. Comput. Sci.*, **1979**, *19*, 159–162.

[174] Willett, P. Searching Techniques for Databases of Two- and Three-Dimensional Chemical Structures., *J. Med. Chem.*, **2005**, *48*, 4183–4199.

[175] Cringean, J. K.; Lynch, M. F. Subgraphs of reduced chemical graphs as screens for substructure searching of specific chemical structures., *J. Inf. Sci.*, **1989**, *15*, 211–222.

[176] Sussenguth, E. H. A Graph-Theoretic Algorithm for Matching Chemical Structures., *J. Chem. Doc.*, **1965**, *5*, 36–43.

[177] Ullmann, J. R. An Algorithm for Subgraph Isomorphism., *J ACM*, **1976**, *23*, 31–42.

[178] Corneil, D. G.; Gotlieb, C. C. An Efficient Algorithm for Graph Isomorphism., *J. ACM*, **1970**, *17*.

[179] Wipke, W. T.; Rogers, D. Rapid subgraph search using parallelism., *J. Chem. Inf. Comput. Sci.*, **1984**, *24*, 255–262.

[180] Ullmann, J. R. Bit-vector algorithms for binary constraint satisfaction and subgraph isomorphism., *ACM J. Exp. Algorithmics*, **2011**, *15*, 1 – 64.

[181] Von Scholley, A. A relaxation algorithm for generic chemical structure screening., *J. Chem. Inf. Comput. Sci.*, **1984**, *24*, 235–241.

[182] Figueras, J. Substructure Search by Set Reduction., *J. Chem. Doc.*, **1972**, *12*, 237–244.

[183] http://accelrys.com/products/informatics/decision-support/isentris.html; Accessed October 2011.

[184] http://www.cambridgesoft.com/software/chembiofinder/; Accessed October 2011.

[185] Lipkus, A. H.; Blower, P. E. Relative Configurations in Stereochemical Substructure Search. In *Chemical Structures 2: The International Language of Chemistry*; Noordwijkerhout, The Netherlands, **1990**.

[186] McGregor, J. J. Relational consistency algorithms and their application in finding subgraph and graph isomorphisms., *Inf. Sci.*, **1979**, *19*, 229–250.

[187] Crawford, B.; Castro, C.; Monfroy, E. Solving Sudoku with Constraint Programming. In *Cutting-Edge Research Topics on Multiple Criteria Decision Making*; Communications in Computer and Information Science; **2009**; Vol. 35, pp. 345 – 348.

[188] Lewis, R. Metaheuristics can solve sudoku puzzles., *J. Heuristics*, **2007**, *13*, 387–401.

[189] Naanaa, W. A domain decomposition algorithm for constraint satisfaction., *J Exp Algor*, **2009**, *13*, 13–23.

[190] Mackworth, A. K. Consistency in networks of relations., *Artif. Intell.*, **1977**, *8*, 99–118.

[191] Mohr, R.; Henderson, T. C. Arc and path consistency revisited., *Artif. Intell.*, **1986**, *28*, 225–233.

[192] Fowler, G.; Haralick, R.; Gray, F. G.; Feustel, C.; Grinstead, C. Efficient graph automorphism by vertex partitioning., *Artif. Intell.*, **1983**, *21*, 245–269.

[193] McKay, B. D. Practical Graph Isomorphism., *Congr. Numerantium*, **1981**, *30*, 45–87.

[194] Wallace, R. J.; Freuder, E. C. Ordering heuristics for arc consistency algorithms. In *In AI/GI/VI '92*; **1992**; pp. 163 – 169.

[195] Miguel, I.; Shen, Q. Hard, Flexible and Dynamic Constraint Satisfaction., *Knowl. Eng. Rev.*, **1999**, *14*, 199–220.

[196] Freuder, E. C.; Wallace, R. J. Partial constraint satisfaction., *Artif. Intell.*, **1992**, *58*, 21–70.

[197] McKay, B. D. Isomorph Free Exhaustive Generation., *J Algorithms*, **1998**, *26*, 306–324.

[198] Ivanov, J.; Schuurmann, G. Simple Algorithms for Determining the Molecular Symmetry., *J. Chem. Inf. Comput. Sci.*, **1999**, *39*, 728–737.

[199] Shelley, C. A.; Munk, M. E. Signal number prediction in carbon-13 nuclear magnetic resonance spectrometry., *Anal. Chem.*, **1978**, *50*, 1522–1527.

[200] Christie, B. D.; Munk, M. E. The role of two-dimensional nuclear magnetic resonance spectroscopy in computer-enhanced structure elucidation., *J. Am. Chem. Soc.*, **1991**, *113*, 3750–3757.

[201] Liu, X.; Balasubramanian, K.; Munk, M. E. Computational techniques for vertex partitioning of graphs., *J. Chem. Inf. Comput. Sci.*, **1990**, *30*, 263–269.

[202] Razinger, M.; Balasubramanian, K.; Munk, M. E. Graph automorphism perception algorithms in computer-enhanced structure elucidation., *J. Chem. Inf. Comput. Sci.*, **1993**, *33*, 197–201.

[203] Druffel, L. E.; Schmidt, D. C.; Wang, D.-L. A Generator Set for Representing All Automorphisms of a Graph., *SIAM J. Appl. Math.*, **1978**, *34*, 593–596.

[204] McKay, B. D. NAUTY User's Guide (Version 2.4) **2009**.

[205] Faulon, J.-L. Isomorphism, Automorphism Partitioning, and Canonical Labelling Can Be Solved in Polynomial-Time for Molecular Graphs., *J. Chem. Inf. Comput. Sci.*, **1998**, *38*, 432–444.

[206] Shelley, C. A.; Munk, M. E. Computer Perception of Topological Symmetry., *J. Chem. Inf. Comput. Sci.*, **1977**, *17*, 110–113.

[207] Ouyang, Z.; Yuan, S.; Brandt, J.; Zheng, C. An Effective Topological Symmetry Perception and Unique Numbering Algorithm., *J. Chem. Inf. Comput. Sci.*, **1999**, *39*, 299–303.

[208] Ruecker, G.; Ruecker, C. On using the adjacency matrix power method for perception of symmetry and for isomorphism testing of highly intricate graphs., *J. Chem. Inf. Comput. Sci.*, **1991**, *31*, 123–126.

[209] Liu, X.; Klein, D. J. The graph isomorphism problem., *J. Comput. Chem.*, **1991**, *12*, 1243–1251.

[210] Read, R. C.; Corneil, D. G. The graph isomorphism disease., *J. Graph Theory*, **1977**, *1*, 339–363.

[211] Ruecker, G.; Ruecker, C. Isocodal and isospectral points, edges, and pairs in graphs and how to cope with them in computerized symmetry recognition., *J. Chem. Inf. Comput. Sci.*, **1991**, *31*, 422–427.

[212] Knop, J. V.; Müller, W. R.; Szymanski, K.; Trinajstić, N.; Kleiner, A. F.; Randić, M. On irreducible endospectral graphs., *J. Math. Phys.*, **1986**, *27*, 2601.

[213] Hong, H.; Xin, X. ESSESA: An Expert System for Structure Elucidation from Spectra. 4. Canonical Representation of Structures., *J. Chem. Inf. Comput. Sci.*, **1994**, *34*, 730–734.

[214] Hu, C.-Y.; Xu, L. Algorithm for computer perception of topological symmetry., *Anal. Chim. Acta*, **1994**, *295*, 127–134.

[215] McKay, B. D. Computing Automorphism and Canonical Labelling of Graphs. In *Combinatorial Mathematics*; Lecture Notes in Mathematics; Springer-Verlag: Berlin, **1978**; pp. 223–232.

[216] Hartke, S.; Radcliffe, A. J. McKay's Canonical Graph Labeling Algorithm. In *Communicating Mathematics*; Contemporary Mathematics; AMS, **2009**; Vol. 479, pp. 99–111.

[217] Miyazaki, T. The Complexity of McKay's Canonical Labelling Algorithm. In *Groups and Computation II: Workshop on Groups and Computation*; DIMACS Series in Discrete Mathematics and Computer Science; American Mathematical Society, **1995**; Vol. 28.

[218] Darga, P. T.; Sakallah, K. A.; Markov, I. L. Faster Symmetry Discovery using Sparsity of Symmetries. In *Proceedings of the 45th annual Design Automation Conference*; ACM: Anaheim, California, **2008**; pp. 149–154.

[219] Junttila, T.; Kaski, P. Engineering an Efficient Canonical Labelling Tool for Large and Sparse Graphs. In *Proceedings of the Ninth Workshop on Algorithm Engineering and Experiments and the Fourth Workshop on Analytic Algorithms and Combinatorics*; SIAM, **2007**; pp. 135–149.

[220] Tener, G.; Deo, N. Efficient isomorphism of Miyazaki graphs. In *39th Southeastern International Conference on Combinatorics*; Boca, **2008**.

[221] McKay, B. D. Backtrack Programming and Isomorph Rejection on Ordered Subsets., *Ars Comb.*, **1978**, *5*, 65–99.

[222] Brown, C.; Finkelstein, L.; Purdom, P. Backtrack searching in the presence of symmetry. In *Applied Algebra, Algebraic Algorithms and Error-Correcting Codes*; Mora, T., Ed.; Lecture Notes in Computer Science; Springer Berlin / Heidelberg, **1989**; Vol. 357, pp. 99–110.

[223] Crawford, J.; Ginsberg, M.; Luks, E.; Roy, A. Symmetry-Breaking Predicates for search problems. In *Principles of Knowledge Representation and Reasoning*; Morgan Kaufman, **1996**; Vol. 429, pp. 148–159.

[224] Cohen, D.; Jeavons, P.; Jefferson, C.; Petrie, K.; Smith, B. Symmetry definitions for constraint satisfaction problems., *Constraints*, **2006**, *11*, 115–137.

[225] Gent, I. P.; Kelsey, T.; Linton, S.; Roney-Dougal, C. Symmetry and Consistency. In *Proceedings of the 11th International Conference on Principles and Practice of Constraint Programming*; Springer, **2005**; pp. 271–285.

[226] Gent, I. P.; Harvey, W.; Kelsey, T. Groups and Constraints: Symmetry Breaking during Search. In *Proceedings of CP-02, LNCS 2470*; Springer, **2002**; pp. 415–430.

[227] Gent, I. P.; Harvey, W.; Kelsey, T.; Linton, S. Generic SBDD using computational group theory. In *Proceedings of CP'03*; Springer, **2003**; pp. 333–347.

[228] Uchino, M. Algorithms for unique and unambiguous coding and symmetry perception of molecular structure diagrams. 5. Unique coding by the method of orbit graphs., *J. Chem. Inf. Comput. Sci.*, **1982**, *22*, 201–206.

[229] Mihalić, Z.; Veljan, D.; Amić, D.; Nikolić, S.; Plavšić, D.; Trinajstić, N. The distance matrix in chemistry., *J. Math. Chem.*, **1992**, *11*, 223–258.

[230] Bersohn, M. A fast algorithm for calculation of the distance matrix of a molecule., *J. Comput. Chem.*, **1983**, *4*, 110–113.

[231] Müller, W. R.; Szymanski, K.; Knop, J. V.; Trinajstić, N. An algorithm for construction of the molecular distance matrix., *J. Comput. Chem.*, **1987**, *8*, 170–173.

[232] Ayral Kaloustian, S.; Kaloustian, M. K. Determining homotopic, enantiotopic, and diastereotopic faces in organic molecules., *J. Chem. Educ.*, **1975**, *52*, 56.

[233] Long, A. K.; Kappos, J. C.; Rubenstein, S. D.; Walker, G. E. Computer-assisted synthetic analysis. A generalized procedure for subgoal transform selection based on a two-dimensional pattern language., *J. Chem. Inf. Comput. Sci.*, **1994**, *34*, 922–933.

[234] Corey, E. J.; Orf, H. W.; Pensak, D. A. Computer-assisted synthetic analysis. The identification and protection of interfering functionality in machine-generated synthetic intermediates., *J. Am. Chem. Soc.*, **1976**, *98*, 210–221.

[235] Hoyle, P. L. M. *Computer Assisted Design of Organic Synthesis*; Ph. D. Thesis.; The University of Leeds, **1986**.

[236] Esack, A.; Bersohn, M. A program for rapid and automatic functional group recognition., *J. Chem. Soc. [Perkin 1]*, **1974**, 2463.

[237] Myatt, G. J. *Computer Aided Estimation of Synthetic Accessibility*; Ph. D. Thesis.; The University of Leeds, **1994**.

[238] Hopkinson, G. A. *Computer-Assisted Organic Synthesis Design*; Ph. D. Thesis.; The University of Leeds, **1985**.

[239] Bohlen, J.; Parlow, A.; Welske, C.; Gasteiger, J. Cheminform — An Integrated Production Process for the Building of a Reaction Database and the Publishing of a Printed Abstracts Service. In *Software Development in Chemistry 5*; Springer Berlin Heidelberg, **1990**; pp. 37–43.

[240] Bertelsen, S.; Jørgensen, K. A. Organocatalysis—after the gold rush., *Chem. Soc. Rev.*, **2009**, *38*, 2178.

[241] Dalko, P. I.; Moisan, L. In the Golden Age of Organocatalysis., *Angew. Chem. Int. Ed.*, **2004**, *43*, 5138–5175.

[242] Jamieson, S. Likert scales: how to (ab)use them., *Med. Educ.*, **2004**, *38*, 1217–1218.

[243] Zwillinger, D.; Kokoska, S. *CRC Standard Probability and Statistics Tables and Formulae*; CRC Press, **2000**.

[244] Furniss, B. S.; Hannaford, A. J.; Smith, P. W. G.; Tatchell, A. R. *Vogel's Textbook of Practical Organic Chemistry (5th ed.)*; Longman: Harlow, **1989**.

[245] Masamune, S.; Choy, W.; Petersen, J. S.; Sita, L. R. Double Asymmetric Synthesis and a New Strategy for Stereochemical Control in Organic Synthesis., *Angew. Chem. Int. Ed. Engl.*, **1985**, *24*, 1–30.

[246] Xia, Q.-H.; Ge, H.-Q.; Ye, C.-P.; Liu, Z.-M.; Su, K.-X. Advances in Homogeneous and Heterogeneous Catalytic Asymmetric Epoxidation., *Chem Rev*, **2005**, *105*, 1603–1662.

[247] Katsuki, T.; Sharpless, K. B. The first practical method for asymmetric epoxidation., *J Am Chem Soc*, **1980**, *102*, 5974–5976.

[248] Curci, R.; Fiorentino, M.; Serio, M. R. Asymmetric epoxidation of unfunctionalized alkenes by dioxirane intermediates generated from potassium peroxomonosulphate and chiral ketones., *J. Chem. Soc. Chem. Commun.*, **1984**, 155–156.

[249] Tu, Y.; Wang, Z.-X.; Shi, Y. An Efficient Asymmetric Epoxidation Method for trans-Olefins Mediated by a Fructose-Derived Ketone., *J. Am. Chem. Soc.*, **1996**, *118*, 9806–9807.

[250] Zhang, W.; Loebach, J. L.; Wilson, S. R.; Jacobsen, E. N. Enantioselective epoxidation of unfunctionalized olefins catalyzed by salen manganese complexes., *J. Am. Chem. Soc.*, **1990**, *112*, 2801–2803.

[251] Sharpless, K. B.; Woodard, S. S. On the mechanism of titanium-tartrate catalyzed asymmetric epoxidation., *Pure Appl. Chem.*, **1983**, *55*.

[252] Woodard, S. S.; Finn, M. G.; Sharpless, K. B. Mechanism of asymmetric epoxidation. 1. Kinetics., *J. Am. Chem. Soc.*, **1991**, *113*, 106–113.

[253] Finn, M. G.; Sharpless, K. B. Mechanism of asymmetric epoxidation. 2. Catalyst structure., *J. Am. Chem. Soc.*, **1991**, *113*, 113–126.

[254] McGarrigle, E. M.; Gilheany, D. G. Chromium− and Manganese−salen Promoted Epoxidation of Alkenes., *Chem. Rev.*, **2005**, *105*, 1563–1602.

[255] Kürti, L.; Blewett, M. M.; Corey, E. J. Origin of Enantioselectivity in the Jacobsen Epoxidation of Olefins., *Org. Lett.*, **2009**, *11*, 4592–4595.

[256] Wong, O. A.; Shi, Y. Organocatalytic Oxidation. Asymmetric Epoxidation of Olefins Catalyzed by Chiral Ketones and Iminium Salts., *Chem. Rev.*, **2008**, *108*, 3958–3987.

[257] Singleton, D. A.; Wang, Z. Isotope Effects and the Nature of Enantioselectivity in the Shi Epoxidation. The Importance of Asynchronicity., *J. Am. Chem. Soc.*, **2005**, *127*, 6679–6685.

[258] Lattanzi, A. Advances in Asymmetric Epoxidation of α,β -Unsaturated Carbonyl Compounds:The Organocatalytic Approach., *Curr. Org. Synth.*, **2008**, *5*, 117–133.

[259] Hashimoto, T.; Maruoka, K. Recent Development and Application of Chiral Phase-Transfer Catalysts., *Chem. Rev.*, **2007**, *107*, 5656–5682.

[260] Marigo, M.; Franzén, J.; Poulsen, T. B.; Zhuang, W.; Jørgensen, K. A. Asymmetric Organocatalytic Epoxidation of α,β-Unsaturated Aldehydes with Hydrogen Peroxide., *J. Am. Chem. Soc.*, **2005**, *127*, 6964–6965.

[261] Arai, S.; Tsuge, H.; Oku, M.; Miura, M.; Shioiri, T. Catalytic asymmetric epoxidation of enones under phase-transfer catalyzed conditions., *Tetrahedron*, **2002**, *58*, 1623–1630.

[262] Jacques, O.; Richards, S. J.; Jackson, R. F. W. Catalytic asymmetric epoxidation of aliphatic enones using tartrate-derived magnesium alkoxides., *Chem. Commun.*, **2001**, 2712–2713.

[263] Daikai, K.; Hayano, T.; Kino, R.; Furuno, H.; Kagawa, T.; Inanaga, J. Asymmetric catalysis with self-organized chiral lanthanum complexes: Practical and highly enantioselective epoxidation of α,β-unsaturated ketones., *Chirality*, **2003**, *15*, 83–88.

[264] Lee, N. H.; Jacobsen, E. N. Enantioselective epoxidation of conjugated dienes and enynes. Trans-epoxides from cis-olefins., *Tetrahedron Lett.*, **1991**, *32*, 6533–6536.

[265] Geary, L. M.; Hultin, P. G. The state of the art in asymmetric induction: the aldol reaction as a case study., *Tetrahedron Asymmetry*, **2009**, *20*, 131–173.

[266] Guillena, G.; Nájera, C.; Ramón, D. J. Enantioselective direct aldol reaction: the blossoming of modern organocatalysis., *Tetrahedron Asymmetry*, **2007**, *18*, 2249–2293.

[267] Zimmerman, H. E.; Traxler, M. D. The Stereochemistry of the Ivanov and Reformatsky Reactions. I., *J. Am. Chem. Soc.*, **1957**, *79*, 1920–1923.

[268] Evans, D. A.; Downey, C. W.; Shaw, J. T.; Tedrow, J. S. Magnesium Halide-Catalyzed Anti-Aldol Reactions of Chiral N-Acylthiazolidinethiones., *Org. Lett.*, **2002**, *4*, 1127–1130.

[269] Arya, P.; Qin, H. Advances in Asymmetric Enolate Methodology., *Tetrahedron*, **2000**, *56*, 917–947.

[270] Li, J.; Menche, D. Direct Methods for Stereoselective Polypropionate Synthesis: A Survey., *Synthesis*, **2009**, *2009*, 2293–2315.

[271] Nelson, S. G. Catalyzed enantioselective aldol additions of latent enolate equivalents., *Tetrahedron Asymmetry*, **1998**, *9*, 357–389.

[272] Evans, D. A.; Nelson, J. V.; Vogel, E.; Taber, T. R. Stereoselective aldol condensations via boron enolates., *J Am Chem Soc*, **1981**, *103*, 3099–3111.

[273] Tang, Z.; Jiang, F.; Yu, L.-T.; Cui, X.; Gong, L.-Z.; Mi, A.-Q.; Jiang, Y.-Z.; Wu, Y.-D. Novel Small Organic Molecules for a Highly Enantioselective Direct Aldol Reaction., *J. Am. Chem. Soc.*, **2003**, *125*, 5262–5263.

[274] Mase, N.; Nakai, Y.; Ohara, N.; Yoda, H.; Takabe, K.; Tanaka, F.; Barbas, C. F. Organocatalytic Direct Asymmetric Aldol Reactions in Water., *J. Am. Chem. Soc.*, **2006**, *128*, 734–735.

[275] Evans, D. A.; Bartroli, J.; Shih, T. L. Enantioselective aldol condensations. 2. Erythro-selective chiral aldol condensations via boron enolates., *J. Am. Chem. Soc.*, **1981**, *103*, 2127–2129.

[276] Walker, M. A.; Heathcock, C. H. Acyclic stereoselection. 54. Extending the scope of the Evans asymmetric aldol reaction: preparation of anti and “non-Evans” syn aldols., *J. Org. Chem.*, **1991**, *56*, 5747–5750.

[277] Crimmins, M. T.; King, B. W.; Tabet, E. A. Asymmetric Aldol Additions with Titanium Enolates of Acyloxazolidinethiones: Dependence of Selectivity on Amine Base and Lewis Acid Stoichiometry., *J. Am. Chem. Soc.*, **1997**, *119*, 7883–7884.

[278] Crimmins, M. T.; King, B. W.; Tabet, E. A.; Chaudhary, K. Asymmetric Aldol Additions: Use of Titanium Tetrachloride and (−)-Sparteine for the Soft Enolization of N-Acyl Oxazolidinones, Oxazolidinethiones, and Thiazolidinethiones., *J. Org. Chem.*, **2001**, *66*, 894–902.

[279] Pihko, P. M.; Erkkilä, A. Enantioselective synthesis of prelactone B using a proline-catalyzed crossed-aldol reaction., *Tetrahedron Lett.*, **2003**, *44*, 7607–7609.

[280] Loh, T.-P.; Feng, L.-C.; Yang, H.-Y.; Yang, J.-Y. l-Proline in an ionic liquid as an efficient and reusable catalyst for direct asymmetric aldol reactions., *Tetrahedron Lett.*, **2002**, *43*, 8741–8743.

[281] Kotrusz, P.; Kmentová, I.; Gotov, B.; Toma, Š.; Solčániová, E. Proline-catalysed asymmetric aldol reaction in the room temperature ionic liquid [bmim]PF6., *Chem. Commun.*, **2002**, 2510–2511.

[282] Chandrasekhar, S.; Narsihmulu, C.; Reddy, N. R.; Sultana, S. S. Asymmetric aldol reactions in poly(ethylene glycol) catalyzed by l-proline., *Tetrahedron Lett.*, **2004**, *45*, 4581–4582.

[283] Matsuo, J.; Murakami, M. The Mukaiyama Aldol Reaction: 40 Years of Continuous Development., *Angew. Chem. Int. Ed.*, **2013**, *52*, 9109–9118.

[284] Job, A.; Janeck, C. F.; Bettray, W.; Peters, R.; Enders, D. The SAMP-/RAMP-hydrazone methodology in asymmetric synthesis., *Tetrahedron*, **2002**, *58*, 2253–2329.

[285] Evans, D. A.; Tedrow, J. S.; Shaw, J. T.; Downey, C. W. Diastereoselective Magnesium Halide-Catalyzed anti-Aldol Reactions of Chiral N-Acyloxazolidinones., *J. Am. Chem. Soc.*, **2002**, *124*, 392–393.

[286] Enders, D.; Breuer, K.; Runsink, J.; Teles, J. H. The First Asymmetric Intramolecular Stetter Reaction. Preliminary Communication., *Helv. Chim. Acta*, **1996**, *79*, 1899–1902.

[287] Breslow, R. On the Mechanism of Thiamine Action. IV.1 Evidence from Studies on Model Systems., *J. Am. Chem. Soc.*, **1958**, *80*, 3719–3726.

[288] Kerr, M. S.; Read de Alaniz, J.; Rovis, T. An Efficient Synthesis of Achiral and Chiral 1,2,4-Triazolium Salts: Bench Stable Precursors for N-Heterocyclic Carbenes., *J. Org. Chem.*, **2005**, *70*, 5725–5728.

[289] De Alaniz, J.; Rovis, T. The Catalytic Asymmetric Intramolecular Stetter Reaction., *Synlett*, **2009**, *2009*, 1189–1207.

[290] Matsumoto, Y.; Tomioka, K. C2 Symmetric chiral N-heterocyclic carbene catalyst for asymmetric intramolecular Stetter reaction., *Tetrahedron Lett.*, **2006**, *47*, 5843–5846.

[291] Read de Alaniz, J.; Rovis, T. A Highly Enantio- and Diastereoselective Catalytic Intramolecular Stetter Reaction., *J. Am. Chem. Soc.*, **2005**, *127*, 6284–6289.

[292] De Alaniz, J. R.; Kerr, M. S.; Moore, J. L.; Rovis, T. Scope of the Asymmetric Intramolecular Stetter Reaction Catalyzed by Chiral Nucleophilic Triazolinylidene Carbenes., *J. Org. Chem.*, **2008**, *73*, 2033–2040.

[293] Um, J. M.; DiRocco, D. A.; Noey, E. L.; Rovis, T.; Houk, K. N. Quantum Mechanical Investigation of the Effect of Catalyst Fluorination in the Intermolecular Asymmetric Stetter Reaction., *J. Am. Chem. Soc.*, **2011**, *133*, 11249–11254.

[294] Taura, Y.; Tanaka, M.; Funakoshi, K.; Sakai, K. Asymmetric cyclization reactions by Rh(I) with chiral ligands., *Tetrahedron Lett.*, **1989**, *30*, 6349–6352.

[295] Taura, Y.; Tanaka, M.; Wu, X.-M.; Funakoshi, K.; Sakai, K. Asymmetric cyclization reactions. Cyclization of substituted 4-pentenals into cyclopentanone derivatives by rhodium(I) with chiral ligands., *Tetrahedron*, **1991**, *47*, 4879–4888.

[296] Evans, D. A.; Dart, M. J.; Duffy, J. L.; Yang, M. G.; Livingston, A. B. Diastereoselective Aldol and Allylstannane Addition Reactions. The Merged Stereochemical Impact of .alpha. and .beta. Aldehyde Substituents., *J. Am. Chem. Soc.*, **1995**, *117*, 6619–6620.

[297] Mitchell, H. J.; Nelson, A.; Warren, S. Strategies for the stereoselective synthesis of molecules with remote stereogenic centres across a double bond of fixed configuration., *J. Chem. Soc. [Perkin 1]*, **1999**, 1899–1914.

[298] Hoveyda, A. H.; Evans, D. A.; Fu, G. C. Substrate-directable chemical reactions., *Chem. Rev.*, **1993**, *93*, 1307–1370.

[299] Chérest, M.; Felkin, H.; Prudent, N. Torsional strain involving partial bonds. The stereochemistry of the lithium aluminium hydride reduction of some simple open-chain ketones., *Tetrahedron Lett.*, **1968**, *9*, 2199–2204.

[300] Chérest, M.; Felkin, H. Torsional strain involving partial bonds. The steric course of the reaction between allyl magnesium bromide and 4-t-butyl-cyclohexanone., *Tetrahedron Lett.*, **1968**, *9*, 2205–2208.

[301] Anh, N. T.; Eisentein, O. Theoretical Interpretation of 1-2 Asymmetric Induction - Importance of Anti-periplanarity., *Nouv. J. Chim.*, **1977**, *1*, 61–70.

[302] Cram, D. J.; Wilson, D. R. Studies in Stereochemistry. XXXII. Models for 1,2-Asymmetric Induction., *J. Am. Chem. Soc.*, **1963**, *85*, 1245–1249.

[303] Reetz, M. T. Chelation or Non-Chelation Control in Addition Reactions of Chiral α- and β-Alkoxy Carbonyl Compounds., *Angew. Chem. Int. Ed. Engl.*, **1984**, *23*, 556–569.

[304] Houk, K. N.; Paddon-Row, M. N.; Rondan, N. G.; Wu, Y.-D.; Brown, F. K.; Spellmeyer, D. C.; Metz, J. T.; Li, Y.; Loncharich, R. J. Theory and Modeling of Stereoselective Organic Reactions., *Science*, **1986**, *231*, 1108–1117.

[305] Washington, I.; Houk, K. N. CH···O Hydrogen Bonding Influences π-Facial Stereoselective Epoxidations., *Angew. Chem. Int. Ed.*, **2001**, *40*, 4485–4488.

[306] Itoh, T.; Jitsukawa, K.; Kaneda, K.; Teranishi, S. Vanadium-catalyzed epoxidation of cyclic allylic alcohols. Stereoselectivity and stereocontrol mechanism., *J. Am. Chem. Soc.*, **1979**, *101*, 159–169.

[307] Rossiter, B. E.; Verhoeven, T. R.; Sharpless, K. B. Stereoselective epoxidation of acyclic allylic alcohols. A correction of our previous work., *Tetrahedron Lett.*, **1979**, *20*, 4733–4736.

[308] Sharpless, K. B.; Verhoeven, T. R. Metal-catalyzed, highly selective oxygenations of olefins and acetylenes with tert-butyl hydroperoxide. Practical considerations and mechanisms., *Aldrichimica Acta*, **1979**, *12*, 52–63.

[309] Cram, D. J.; Elhafez, F. A. A. Studies in Stereochemistry. X. The Rule of "Steric Control of Asymmetric Induction" in the Syntheses of Acyclic Systems., *J. Am. Chem. Soc.*, **1952**, *74*, 5828–5835.

[310] Cornforth, J. W.; Cornforth, R. H.; Mathew, K. K. 24. A general stereoselective synthesis of olefins., *J. Chem. Soc. Resumed*, **1959**, 112–127.

[311] Anh, N. T.; Eisenstein, O. Induction asymetrique 1–2: comparaison ab initio des modeles de cram, de cornforth, de karabatsos et de felkin., *Tetrahedron Lett.*, **1976**, *17*, 155–158.

[312] Burgi, H. B.; Dunitz, J. D.; Shefter, E. Geometrical reaction coordinates. II. Nucleophilic addition to a carbonyl group., *J. Am. Chem. Soc.*, **1973**, *95*, 5065–5067.

[313] Burgi, H. B.; Dunitz, J. D.; Lehn, J. M.; Wipff, G. Stereochemistry of reaction paths at carbonyl centres., *Tetrahedron*, **1974**, *30*, 1563–1572.

[314] Cee, V. J.; Cramer, C. J.; Evans, D. A. Theoretical Investigation of Enolborane Addition to α-Heteroatom-Substituted Aldehydes. Relevance of the Cornforth and Polar Felkin−Anh Models for Asymmetric Induction., *J. Am. Chem. Soc.*, **2006**, *128*, 2920–2930.

[315] Evans, D. A.; Siska, S. J.; Cee, V. J. Resurrecting the Cornforth Model for Carbonyl Addition: Studies on the Origin of 1,2-Asymmetric Induction in Enolate Additions to Heteroatom-Substituted Aldehydes., *Angew. Chem. Int. Ed.*, **2003**, *42*, 1761–1765.

[316] Evans, D. A.; Duffy, J. L.; Dart, M. J. 1,3-Asymmetric induction in the aldol addition reactions of methyl ketone enolates and enolsilanes to β-substituted aldehydes. A model for chirality transfer., *Tetrahedron Lett.*, **1994**, *35*, 8537–8540.

[317] Bahmanyar, S.; Houk, K. N.; Martin, H. J.; List, B. Quantum Mechanical Predictions of the Stereoselectivities of Proline-Catalyzed Asymmetric Intermolecular Aldol Reactions., *J. Am. Chem. Soc.*, **2003**, *125*, 2475–2479.

[318] Armstrong, A.; Boto, R. A.; Dingwall, P.; Contreras-García, J.; Harvey, M. J.; Mason, N. J.; Rzepa, H. S. The Houk–List transition states for organocatalytic mechanisms revisited., *Chem. Sci.*, **2014**, *5*, 2057–2071.

[319] Gennari, C.; Moresca, D.; Vulpetti, A.; Pain, G. Reagent control in the aldol addition reaction of chiral boron enolates with chiral aldehydes. Total synthesis of (3S,4S)-Statine., *Tetrahedron*, **1997**, *53*, 5593–5608.

[320] Stanton, G. R.; Johnson, C. N.; Walsh, P. J. Overriding Felkin Control: A General Method for Highly Diastereoselective Chelation-Controlled Additions to α-Silyloxy Aldehydes., *J. Am. Chem. Soc.*, **2010**, *132*, 4399–4408.

[321] Stanton, G. R.; Koz, G.; Walsh, P. J. Highly Diastereoselective Chelation-Controlled Additions to α-Silyloxy Ketones., *J. Am. Chem. Soc.*, **2011**, *133*, 7969–7976.

[322] Ghosh, A. K.; Kim, J.-H. Stereoselective Chloroacetate Aldol Reactions: Syntheses of Acetate Aldol Equivalents and Darzens Glycidic Esters., *Org. Lett.*, **2004**, *6*, 2725–2728.

[323] Majewski, M.; Nowak, P. Aldol Addition of Lithium and Boron Enolates of 1,3-Dioxan-5-ones to Aldehydes. A New Entry into Monosaccharide Derivatives., *J. Org. Chem.*, **2000**, *65*, 5152–5160.

[324] Enders, D.; Grondal, C. Direct Organocatalytic De Novo Synthesis of Carbohydrates., *Angew. Chem. Int. Ed.*, **2005**, *44*, 1210–1212.

[325] Jolley, C. C.; Douglas, T. Bipartite Network Analysis of (bio)CHEMICAL Reaction Systems., *Biophys. J.*, **2011**, *100*, 166a.

[326] Schaus, S. E.; Jacobsen, E. N. Asymmetric Ring Opening of Meso Epoxides with TMSCN Catalyzed by (pybox)lanthanide Complexes., *Org. Lett.*, **2000**, *2*, 1001–1004.

[327] Wang, Z.; Law, W. K.; Sun, J. Chiral Phosphoric Acid Catalyzed Enantioselective Desymmetrization of meso-Epoxides by Thiols., *Org. Lett.*, **2013**, *15*, 5964–5966.

[328] Ananthan, B.; Chang, W.-C.; Lin, J.-S.; Li, P.-H.; Yan, T.-H. A C2-Symmetric Chiral Pool-Based Flexible Strategy: Synthesis of (+)- and (−)-Shikimic Acids, (+)- and (−)-4-epi-Shikimic Acids, and (+)- and (−)-Pinitol., *J. Org. Chem.*, **2014**, *79*, 2898–2905.

[329] Chang, Y.-K.; Lo, H.-J.; Yan, T.-H. A Flexible Strategy Based on a C2-Symmetric Pool of Chiral Substrates: Concise Synthesis of (+)-Valienamine, Key Intermediate of (+)-Pancratistatin, and Conduramines A-1 and E., *Org. Lett.*, **2009**, *11*, 4278–4281.

[330] Carpenter, R. D.; Fettinger, J. C.; Lam, K. S.; Kurth, M. J. Asymmetric Catalysis: Resin-Bound Hydroxyprolylthreonine Derivatives in Enamine-Mediated Reactions., *Angew. Chem. Int. Ed.*, **2008**, *47*, 6407–6410.

[331] Denmark, S. E.; Stavenger, R. A. The Chemistry of Trichlorosilyl Enolates. Aldol Addition Reactions of Methyl Ketones., *J. Am. Chem. Soc.*, **2000**, *122*, 8837–8847.

[332] Itsuno, S.; Arima, S.; Haraguchi, N. Asymmetric Mukaiyama aldol reaction of silyl enol ethers with aldehydes using a polymer-supported chiral Lewis acid catalyst., *Tetrahedron*, **2005**, *61*, 12074–12080.

[333] Ollevier, T.; Plancq, B. Highly enantioselective Mukaiyama aldol reaction in aqueous conditions using a chiral iron(II) bipyridine catalyst., *Chem. Commun.*, **2012**, *48*, 2289–2291.

[334] Lafantaisie, M.; Mirabaud, A.; Plancq, B.; Ollevier, T. Iron(II)-Derived Lewis Acid/Surfactant Combined Catalysis for the Enantioselective Mukaiyama Aldol Reaction in Pure Water., *ChemCatChem*, **2014**, *6*, 2244–2247.

[335] Corey, E. J.; Long, A. K.; Greene, T. W.; Miller, J. W. Computer-assisted synthetic analysis. Selection of protective groups for multistep organic syntheses., *J. Org. Chem.*, **1985**, *50*, 1920–1927.

[336] Li, X.; Tanasova, M.; Vasileiou, C.; Borhan, B. Fluorinated Porphyrin Tweezer: A Powerful Reporter of Absolute Configuration for erythro and threo Diols, Amino Alcohols, and Diamines., *J. Am. Chem. Soc.*, **2008**, *130*, 1885–1893.

[337] Torii, S.; Liu, P.; Bhuvaneswari, N.; Amatore, C.; Jutand, A. Chemical and Electrochemical Asymmetric Dihydroxylation of Olefins in I2−K2CO3−K2OsO2(OH)4 and I2−K3PO4/K2HPO4−K2OsO2(OH)4 Systems with Sharpless' Ligand., *J. Org. Chem.*, **1996**, *61*, 3055–3060.

[338] Motorina, I.; Crudden, C. M. Asymmetric Dihydroxylation of Olefins Using Cinchona Alkaloids on Highly Ordered Inorganic Supports., *Org. Lett.*, **2001**, *3*, 2325–2328.

[339] Choudary, B. M.; Chowdari, N. S.; Kantam, M. L.; Raghavan, K. V. Catalytic Asymmetric Dihydroxylation of Olefins with New Catalysts: The First Example of Heterogenization of OsO42- by Ion-Exchange Technique., *J. Am. Chem. Soc.*, **2001**, *123*, 9220–9221.

[340] Kihumbu, D.; Stillger, T.; Hummel, W.; Liese, A. Enzymatic synthesis of all stereoisomers of 1-phenylpropane-1,2-diol., *Tetrahedron Asymmetry*, **2002**, *13*, 1069–1072.

[341] Wang, L.; Sharpless, K. B. Catalytic asymmetric dihydroxylation of cis-disubstituted olefins., *J. Am. Chem. Soc.*, **1992**, *114*, 7568–7570.

[342] Li, Y.-M.; Tang, Y.-Q.; Hui, X.-P.; Huang, L.-N.; Xu, P.-F. Synthesis of new β-hydroxy amide ligands and their Ti(IV) complex-catalyzed enantioselective alkynylation of aliphatic and vinyl aldehydes., *Tetrahedron*, **2009**, *65*, 3611–3614.

[343] Ramachandran, P. V.; Teodorovic, A. V.; Rangaishenvi, M. V.; Brown, H. C. Chiral synthesis via organoboranes. 34. Selective reductions. 47. Asymmetric reduction of hindered .alpha.,.beta.-acetylenic ketones with B-chlorodiisopinocampheylborane to propargylic alcohols of very high enantiomeric excess. Improved workup procedure for the isolation of product alcohols., *J. Org. Chem.*, **1992**, *57*, 2379–2386.

[344] Hirose, T.; Sugawara, K.; Kodama, K. Switching of Enantioselectivity in the Catalytic Addition of Diethylzinc to Aldehydes by Regioisomeric Chiral 1,3-Amino Sulfonamide Ligands., *J. Org. Chem.*, **2011**, *76*, 5413–5428.

[345] Xu, Z.; Mao, J.; Zhang, Y. Highly enantioselective alkynylation of aldehydes catalyzed by a new oxazolidine–titanium complex., *Org. Biomol. Chem.*, **2008**, *6*, 1288–1292.

[346] Boobalan, R.; Chen, C.; Lee, G.-H. Camphor-based Schiff base ligand SBAIB: an enantioselective catalyst for addition of phenylacetylene to aldehydes., *Org. Biomol. Chem.*, **2012**, *10*, 1625–1638.

[347] Hsieh, S.-H.; Gau, H.-M. Asymmetric Alkynyl Additions to Aldehydes Catalyzed by Tunable -Oxovanadium(V) Complexes of Schiff Bases of β-Amino Alcohols., *Synlett*, **2006**, *2006*, 1871–1874.

[348] Wu, K.-H.; Gau, H.-M. Remarkably Efficient Enantioselective Titanium(IV)−(R)-H8-BINOLate Catalyst for Arylations to Aldehydes by Triaryl(tetrahydrofuran)aluminum Reagents., *J. Am. Chem. Soc.*, **2006**, *128*, 14808–14809.

[349] Harper, K. C.; Sigman, M. S. Using Physical Organic Parameters To Correlate Asymmetric Catalyst Performance., *J. Org. Chem.*, **2013**, *78*, 2813–2818.

[350] Harper, K. C.; Bess, E. N.; Sigman, M. S. Multidimensional steric parameters in the analysis of asymmetric catalytic reactions., *Nat. Chem.*, **2012**, *4*, 366–374.

[351] Winstein, S.; Holness, N. J. Neighboring Carbon and Hydrogen. XIX. t-Butylcyclohexyl Derivatives. Quantitative Conformational Analysis., *J. Am. Chem. Soc.*, **1955**, *77*, 5562–5578.

[352] Taft, R. W. Polar and Steric Substituent Constants for Aliphatic and o-Benzoate Groups from Rates of Esterification and Hydrolysis of Esters1., *J. Am. Chem. Soc.*, **1952**, *74*, 3120–3128.

[353] Taft, R. W. Linear Steric Energy Relationships., *J. Am. Chem. Soc.*, **1953**, *75*, 4538–4539.

[354] Charton, M. Steric effects. I. Esterification and acid-catalyzed hydrolysis of esters., *J. Am. Chem. Soc.*, **1975**, *97*, 1552–1556.

[355] Charton, M. Steric effects. 7. Additional V constants., *J. Org. Chem.*, **1976**, *41*, 2217–2220.

[356] Verloop, A.; Hoogenstraaten, W.; Tipker, J. Development and Application of New Steric Substituent Parameters in Drug Design. In *Drug Design*; Medicinal Chemistry: A Series of Monographs; Academic Press, **1976**; Vol. 7, pp. 165–207.

[357] Corey, R. B.; Pauling, L. Molecular Models of Amino Acids, Peptides, and Proteins., *Rev. Sci. Instrum.*, **1953**, *24*, 621–627.

[358] Corey, E. J.; Bakshi, R. K.; Shibata, S. Highly enantioselective borane reduction of ketones catalyzed by chiral oxazaborolidines. Mechanism and synthetic implications., *J. Am. Chem. Soc.*, **1987**, *109*, 5551–5553.

[359] Paterson, I.; Hulme, A. N. Total Synthesis of (-)-Ebelactone A and B., *J. Org. Chem.*, **1995**, *60*, 3288–3300.

[360] Mandal, A. K. Stereocontrolled Total Synthesis of (−)-Ebelactone A., *Org. Lett.*, **2002**, *4*, 2043–2045.

List of Abbreviations

ASAP	As Soon As Possible (algorithm)
BE matrix	Bond-electron matrix
BFGI	Bond orientated Functional Group Interconversion
BINOL	1,1’-Bi-2-napthol
CAS	Chemical Abstracts Service
CE	Cornforth-Evans (model)
CIP	Cahn-Ingold-Prelog
CIRX	ChemInform Reaction Library (database)
CRC	Characteristic Reaction Core
CSP	Constraint Satisfaction Problem
de	Diastereomeric excess
DET	Diethyl tartrate
dr	Diastereomeric ratio
DSD	Double Stereo Differentiation
EBNF	Extended Backus-Naur Form
EDG	Electron-donating Group
ee	Enantiomeric excess
EF	Electrofuge
er	Enantiomeric ratio
ESSR	Extended Set of Smallest Rings
EWG	Electron-withdrawing Group
FA	Felkin-Anh (model)
FG	Functional Group
FGA	Functional Group Addition
FGI	Functional Group Interconversion
FGR	Functional Group Removal
HOMO	Highest Occupied Molecular Orbital
IQR	Interquartile Range
IUPAC	International Union of Pure and Applied Chemistry

JSON	Javascript Object Notation
LUMO	Lowest Unoccupied Molecular Orbital
MEA	Means-Ends Analysis
MIF	Molecule Interchange Format
MOS	Methods of Organic Synthesis (database)
NF	Nucleofuge
NHC	*N*-Heterocyclic Carbene
NP-complete	Nondeterministic Polynomial
oct	Octahedral
PE	Principal Eigenvector
PFA	Polar Felkin-Anh (model)
R matrix	Reaction matrix
RAM	Random Access Memory
RC	Reactive Conformer
RD	Regular Disconnection
REFLIB	Reference Library (database)
RSWB	Retrosynthesis Workbench (application)
SBC	Smallest Binary Code (algorithm)
SBDD	Symmetry Breaking by Dominance Detection (algorithm)
SBDS	Symmetry Breaking During Search (algorithm)
SCRC	Stereochemical Characteristic Reaction Core
SEMA	Stereochemically Extended Morgan Algorithm
S-goal	Structure goal
SMARTS	SMILES Arbitrary Target Specification
SMD	Standard Molecular Data Format
SMILES	Simplified Molecular Input Line Entry Specification
spl	Square planar
SS	Structure Search
SSS	Substructure Search
SSSR	Smallest Set of Smallest Rings
TBHP	tert-Butyl hydroperoxide
tbp	Trigonal bipyramidal

TC	Tactical Combination
T-goal	Target goal
XML	Extensible Markup Language

Appendix A

Use of Set Theory

Terminology and Definitions

Algorithms presented in this thesis use symbols drawn from both set and graph theory. Table 71 lists the important set theory symbols and their interpretations. Table 72 lists additional Boolean logic symbols used in conditional tests.

∀	for all	∈	element of
⊂	subset of	∉	not an element of
⊃	superset of	∋	contains as member
⊄	not subset of	∌	does not contain as member
⊅	not superset of	∩	set intersection
⊆	subset of or equal to	∪	set union
⊇	superset of or equal to	−	set difference
⊈	neither a subset of nor equal to	⊗	set symmetric difference
⊉	neither a superset of nor equal to	∅	empty set
|S|	count of members in set S	{a, b}	set containing members a and b

Table 71 Set theory symbols

∧	logical and	≠	not equal to
∨	logical or	=	equal to

Table 72 Logic symbols used in algorithm descriptions

Appendix B
Use of Graph Theory

Terminology and Definitions

A molecule is represented as an undirected *graph* **G (V, E)** were **V** is a set of *vertices* representing atoms and **E** is a set of *edges* representing bonds. In an *undirected graph*, edges are viewed as an unordered pair of vertices such that **E ⊆ V x V**. The exception is that self-loops or multiple edges are not allowed i.e. atoms can't be bonded to themselves nor can multiple independent bonds exist between the same pair of atoms. A *subgraph* **G' (V', E')** of graph **G (V, E)** is a graph that satisfies the condition **E' ⊆ E** and **V' ⊆ V**. The number of edges that are incident with a vertex is known as its *degree*. A *path* **P** is a subgraph in **G** which is a sequence of edges **(e_1, e_2, … e_{q-1})** were **e_i = {v_i, v_{i+1}}** for **i = 1 … q-1** and the vertices **v_i** for **i = 1 … q** are all different except that **v_1** may occasionally be the same as **v_q**. When we do find a path where **v_1 = v_q**, we have a *cycle* and thus every vertex in the cycle path has an even degree (Figure 140).

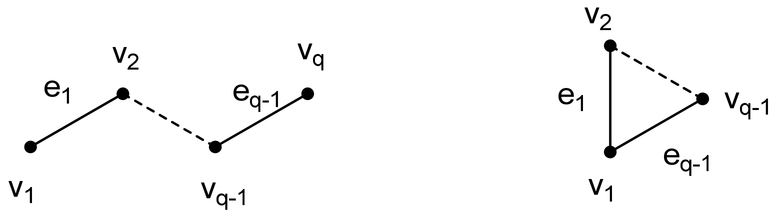


Figure 140 Paths and cycles.

A graph **G** is *connected* if, for any pair of vertices **w, x** in the set of vertices **V**, there is a path such that **w = v_1** and **x = v_q**. The *connected components* of a graph are the maximal connected subgraphs and the count of these is represented as **c (G)**. For most molecules **c (G)** is 1, whereas for binary salts **c (G)** is 2 *etc*. The *cyclomatic number* **μ(G)** of a graph (the number of basis cycles) is given by the Cauchy equation:

$$\mu(G) = |E| - |V| + c\,(G)$$