The Application of Multiobjective Optimisation to Protein-Ligand Docking

Sally Mardikian

Doctor of Philosophy

Department of Information Studies

University of Sheffield

September 2007

Acknowledgements

I would like to thank my supervisors Dr. Val Gillet at the University of Sheffield, and Dr. Dave Westhead and Dr. Richard Jackson at the University of Leeds for all their support and guidance throughout the research period of this work. I would like to thank Dave and Val for all their very insightful comments and for their immense support during the writing-up period. This research was supported by the Medical Research Council to whom I am grateful.

I would like to thank different members of the chemoinformatics group at Sheffield, and the bioinformatics group at Leeds, for all their support, and for being wonderful company. From Sheffield I thank Kris Birchall and Eleanor Gardiner, and a special thank you to Simon Cottrell. From Leeds I would like to thank James Bradford, Nicola Gold, Sean Killen, Binbin Liu, Monika Rella, Archana Sharma-Oates and Liz Webb. I am especially grateful to Nicholas Burgoyne and Peter Oledzki for all our great discussions.

Special thanks to my parents, and brother and sister, for their continuous support and encouragements. Finally I would like to thank my husband Nicholas Casswell, whose love, patience and support gave me strength in the most difficult of times. I dedicate this work to him.

Abstract

Despite the intense efforts that have been devoted to the development of scoring functions for protein-ligand docking, they are still limited in their ability to identify the correct binding pose of a ligand within a protein binding site. A deeper understanding of the intricacies of scoring functions is therefore essential in order to develop these effectively. The aim of the work described in this thesis is to analyse the individual interaction energy types which form the individual components of a force field-based scoring function.

To do this, a protein-ligand docking algorithm that is based on multiobjective optimisation has been developed. Multiobjective optimisation allows for the optimisation of several objectives simultaneously and this has been applied to the individual interaction energy types of the GRID scoring function. Traditionally these interaction energy types are summed together and the total energy is used to guide the search. By using individual energy types during optimisation, their roles can be better understood. The interaction energy types that have been used here are the electrostatic and hydrogen bond interactions combined, and van der Waals interactions.

The algorithm is first tested on two datasets containing twenty complexes. The results show that the different interaction energy types have varying influences when it comes to successfully docking certain complexes, and that it is important to find the right balance of interaction energy types so as to find correct solutions. Of the twenty complexes, the algorithm found correct solutions for fifteen.

To improve the performance of the algorithm, a few enhancements were introduced. This includes a simplex minimisation process with a Lamarckian element. The algorithm was retested on the twenty complexes, and the newer version was found to outperform the original version, finding correct solutions for seventeen of the twenty complexes.

To extensively study the capabilities of the algorithm, it was tested on varied datasets, including the FlexX dataset. The algorithm's performance was also compared to a single-objective docking tool, Q-fit. The comparison between the multiobjective and single-objective methodologies revealed that single-objective methods can sometimes fail at finding correct docked solutions because they are unable to correctly balance the interaction energy types comprising a scoring function. The study also showed that a multiobjective optimisation method can reveal the reasons why a given docking algorithm may fail at finding a correct solution.

Finally, the algorithm was extended to incorporate desolvation energy as a third objective. Though these results are preliminary, they revealed some interesting relationships between the different objectives.

Contents

List of Figures

List of Tables

Abbreviations

GA – genetic algorithm

MOEA – multiobjective evolutionary algorithm

NSGA-II – elitist non-dominated sorting genetic algorithm

SBDD – structure-based drug design

SGA – standard genetic algorithm

vdw – van der Waals

1 Introduction

The path of developing a small molecule into a drug for the market is long and arduous, but one that every pharmaceutical company must take so as to ensure that they survive and flourish in the healthcare industry. High attrition rates, mounting costs of drug development (estimated at over US$1 billion per drug) and regulatory hurdles are all factors which have driven the need to develop methods that increase the efficiency and speed of finding viable leads that can be developed into drugs. Computational methods in particular have seen an increase in popularity within the industry. These range from chemoinformatics techniques such as similarity searching, to virtual screening, a computational technique used to identify ligands from a library of compounds based on various criteria, such as activity against a specified protein target. Virtual screening is one of the tools used in structure-based drug design (SBDD), a concept which uses three dimensional (3-D) structural information of a target protein to find and design drugs. For example, protein-ligand docking is a SBDD tool often used in virtual screening, which tries to predict the structure of a small molecule-protein complex from the molecules' atom coordinates only.

Several of these methods apply search techniques, many of which are borrowed from engineering applications, to try to predict the behaviour and characteristics of small molecules computationally. The computational search methods vary, depending on the type of problem and the method's aims. One such method is multiobjective optimisation, which has been used in several drug discovery/chemoinformatics applications, including library selection (Gillet *et al.*, 2002), and pharmacophore generation (Cottrell *et al.*, 2004). The theory behind this method and a discussion of its different implementations is presented in Chapter 2. Chapter 2 also introduces genetic algorithms (GAs), which have been used to perform multiobjective optimisation. GAs are heuristic search algorithms based on Darwinian concepts of evolution and natural selection.

Using computational methods, such as protein-ligand docking, to predict the behaviour of molecules requires a good understanding of biological notions that drive

these molecules' interactions. Chapter 3 discusses molecular recognition concepts and the various energy components that are considered when modelling biological systems. The use of these concepts in protein-ligand docking is discussed, along with the different types of algorithms that have successfully been applied to the problem (Kitchen *et al.*, 2004).

Dicussion of the experimental work carried out in the thesis begins in Chapter 4, which explains the development of a single-objective GA that performs protein-ligand docking. This algorithm is tested on a small dataset, the results of which are also discussed in this chapter.

An area of weakness of current protein-ligand docking methods is their ability to correctly score different docked solutions which leads to deficiencies in virtual screening experiments which attempt to rank a large number of ligands. As the title of the thesis indicates, the aim of this work is to apply multiobjective optimisation to protein-ligand docking in order to gain insights that will lead to improved scoring functions. Chapter 5 discusses how the single-objective, protein-ligand docking algorithm is modified into a multiobjective algorithm. To understand the capabilities of the algorithm, it is tested on two datasets, the results of which are presented in Chapter 6.

An algorithm can be modified in several ways in order to improve its performance, as well as provide different levels of control that can be adjusted depending on the problem. Chapter 7 focuses on the algorithmic side of the research, and describes major modifications carried out on the multiobjective algorithm. To understand the effects of these modifications, the algorithm's performance is compared to that of the unmodified version described in Chapter 6.

In Chapter 8 the biological attributes of the algorithm are explored on a larger scale by testing it on different datasets, the results of which provide a deeper understanding of the algorithm and its benefits. In Chapter 9 the algorithm is once again modified to include a novel component, which, though not fully explored, provides a basis for future work. The thesis is concluded with Chapter 10, which summarises and

discusses the results obtained from this research, and provides a discussion on the future prospects of this work.

2 Multiobjective optimisation: Theories

Optimisation is the process which attempts to find the global solution or solutions, and which describes extreme values of one or more objectives. A problem which involves only one objective function will require finding a single extreme solution, in a process termed single-objective optimisation.

When more than one objective function affects a problem, then the task of finding one or more optimal solutions is known as multiobjective optimisation. Multiobjective optimisation elucidates real-life problems more realistically since these will naturally involve multiple, conflicting, objectives. Therefore finding the extreme solution for one objective will not be sufficient in these cases, since the other objectives need to be considered. In scenarios focusing on one or more objectives, a number of solutions will exist, which may have conflicting trade-offs, or compromises among the objectives. An extreme solution in one objective (one that is optimal in that objective) will require a compromise in the other objective. Given the existence of different solutions at the optimal extremes of the objectives, there will also exist solutions in between the extremes, with different compromises of the different objectives (Figure 2.1).

Multiobjective optimisation problems occur in many everyday decision-making situations. For example, when buying a car, and if, hypothetically, cost is the only criterion to base the decision by, then only the least expensive cars would ever be bought. However, the reality is that other factors which conflict with cost also affect this decision, such as the comfort level of the car. The cheapest cars will have the lowest comfort levels, whereas the most expensive cars will be the most comfortable. Rich individuals to whom comfort is important will therefore opt for the most expensive cars. In between the two extremes there is a whole array of cars with different trade-offs of cost versus comfort.

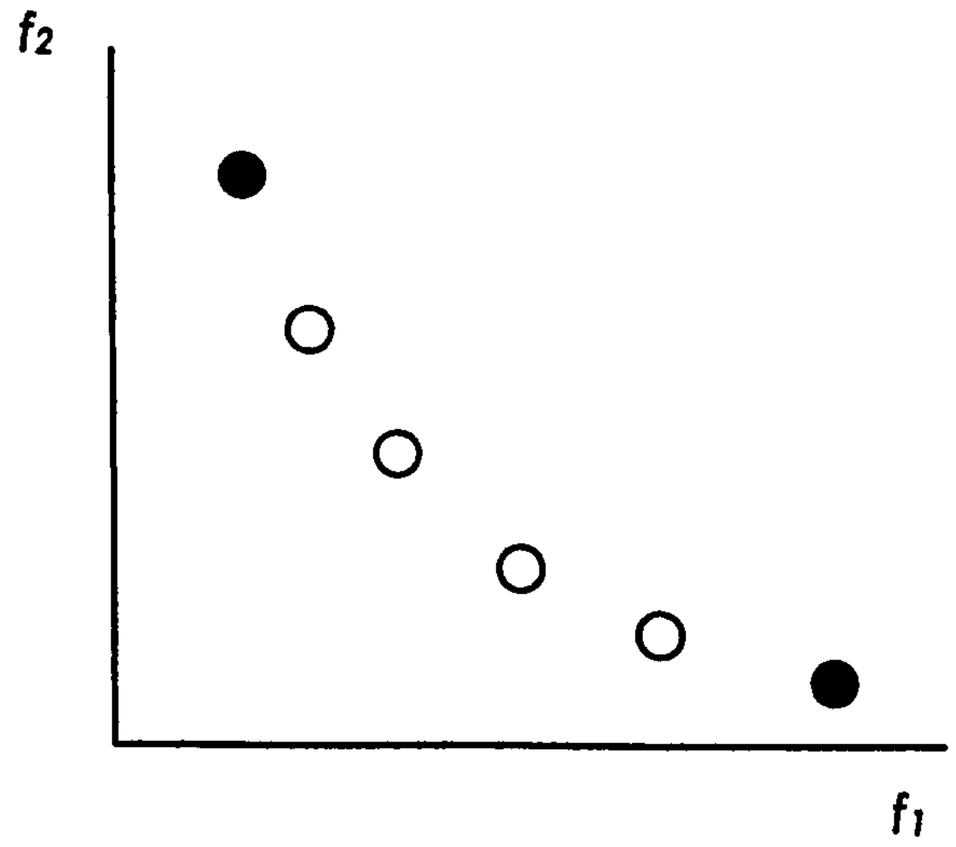


Figure 2-1 In a multiobjective optimisation problem, the (conflicting) objectives f_1 and f_2 are minimised. The solid circle solutions are optimal in terms of one objective, but at the expense of the other. The solutions in open circles are not optimal in either of the objectives but show different compromises in terms of the objectives.

Traditionally multiobjective optimisation problems have been approached by transforming the problem into a single-objective optimisation one by using user-defined parameters (classical methods). The weighted sum technique is one such example, and is described in section 2.3.1. The main problem with such an approach is deciding on the ideal parameters which correctly describe the relationships between the objectives. Also, it is difficult to define these parameters when the objectives themselves are non-commensurate. A true multiobjective optimisation technique will remove any bias towards a particular objective during optimisation and will consider all objectives to be equally important. The process is not completely objective though because, since most optimisation problems require a single solution, then the user, or decision maker, will have to select one of the trade-off solutions. At this point it is hoped that higher level information, information that is qualitative and driven by experience, will assist in making that decision. Therefore the task of a multiobjective optimisation technique is to find optimal solutions which have a good variety in compromised objectives, giving the decision maker a good basis for judging which of the optimal solutions solve the particular problem.

2.1 Evolutionary approach to multiobjective optimisation

The classical approach to multiobjective optimisation combines all objectives into a composite function that is then optimised single objectively. Classical search methods, which operate on a point-by-point approach, modifying a single solution per iteration until the best solution is found are therefore very well suited for single-objective optimisation problems, where only one optimal solution is needed. However, to perform true multiobjective optimisation, multiple solutions must exist during the optimisation process, all of which are optimised simultaneously in order to result in a set of multiple optimised solutions.

Evolutionary methods are therefore ideally suited for multiobjective optimisation. These maintain a population of solutions at all times during optimisation, and therefore the end of a run always contains a final population of solutions. Evolutionary methods are routinely implemented on single-objective optimisation problems, where it is hoped that all population members will converge to a single optimal solution. These methods can be adapted though, so that different population members converge towards multiple optimum points. The details of multiobjective optimisation using evolutionary algorithms, along with evolutionary computation concepts, are discussed in the following sections.

2.2 Evolutionary computation

Evolutionary computation methods are a type of optimisation algorithms which mimic biological evolution (Holland, 1992). They work on the principle that by combining different parts of good but non-optimal solutions together, then the global minimum will be reached. Evolutionary computational techniques consist of three components; a data structure representing decision variables, a fitness function for evaluating the quality of the solutions and a strategy for moving from one generation to the next.

Different types of evolutionary computation methods exist, such as genetic programming, evolutionary strategies and genetic algorithms (Foster, 2001). This thesis focuses largely on genetic algorithms (GAs), therefore this section will focus solely on these evolutionary computation methods.

2.2.1 Genetic algorithms

Genetic algorithms attempt to find a set of decision variables, represented by a vector x, that minimises the value of a fitness function f(x). The process mimics evolutionary Darwinian theory in many ways. For example, the decision variables representing solutions of the population are known as chromosomes. The other similarities to biological processes are in the methods of optimisation as described below.

The genetic operators are the methods applied on the chromosomes to effectively explore the search space and find optimal solutions. Two of these are crossover and mutation. Crossover combines the genetic information (the decision variables), of two "parents" in the population to result in two "offspring" which are different, and hopefully "better", than the parents. The mutation process mutates a part of the information stored in the chromosome at random. These are discussed in detail below.

The chromosomes onto which the genetic operators are applied are selected randomly, though the selection is weighted so that fitter members are more likely to be selected. This process is analogous to the biological theory that fitter members of a population are more likely to survive and reproduce, passing on their genetic traits to the next generation. Least fit chromosomes are removed from the population; in biological scenarios these unfit individuals are least likely to survive and perish before they can reproduce.

2.2.1.1 Chromosome structure

The chromosome, a member of a GA population, encodes information in decision variables into elements known as genes, which are the smallest data units that can be manipulated independently. There are various ways of representing genes, including as floating point numbers or integers which directly represent the problem (real coding). Binary numbers can also be used to represent decision variable information; each binary digit is regarded as a gene. The type of gene encoding selected affects the "resolution" of the data structure, and will affect how the genetic operators are applied to them. The genes of a member of the GA population represent that member's genotype, whereas the phenotype is a tangible characteristic displayed by that member, and that is defined by its genotype.

2.2.1.2 Genetic operators: mutation

The mutation operator makes small changes in a gene of a chromosome. Binary representation allows one of two values in the gene therefore during mutation, the binary digit of a randomly selected gene is simply switched to the other one. In real number encoding of chromosomes, the real number floating point or integers can theoretically take any value. These are usually limited by the gene step-size, which is the maximum value by which a gene can change. This type of mutation is known as random. Creep mutation changes a randomly selected gene by a fixed, pre-determined step-size.

2.2.1.3 Genetic operators: crossover

Crossover, the swapping of information, or genes, between two chromosomes, usually occurs by swapping contiguous sequences of genes between two parent chromosomes. The crossover process begins by selecting a random point, the crossover point, on the two chromosomes. The genes before the crossover point from one parent, are combined with the genes after the crossover point of the second

parent. Figure 2.2 illustrates this process. This is known as single-point crossover because a single point is selected on the chromosome. Two point crossovers divide the chromosome into three sections, and the central portions are swapped between the two parents. A third type of crossover is known as uniform crossover, where corresponding genes from either parent are selected to form the offspring chromosomes. This is equivalent to having 5 crossover points for the chromosomes depicted in Figure 2.2.

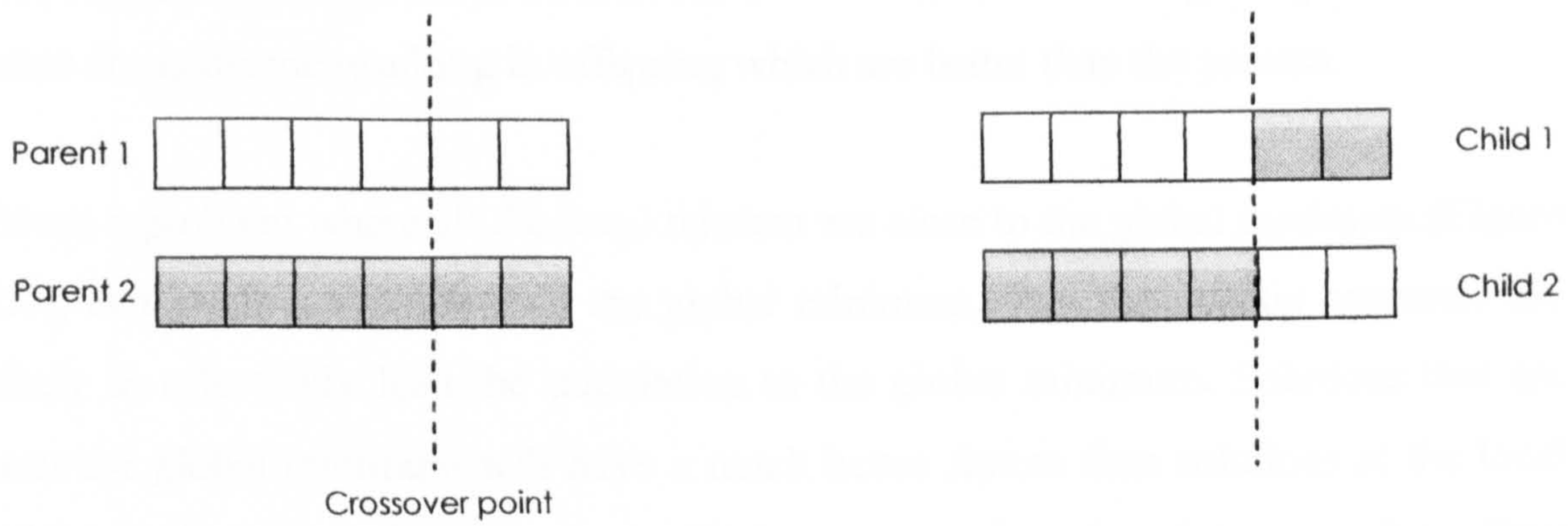


Figure 2-2 Single-point crossover between two chromosomes

2.2.1.4 Selection and replacement strategy theories

The starting population of a GA will consist of randomly generated solutions that are represented by chromosomes. The aim of the GA is to improve the fitness of the entire population over a run of the algorithm. This is performed by ensuring that the genetic operators only act on the fitter chromosomes to produce improved offspring. To keep the population size constant, as new offspring are produced, these must replace other chromosomes in the population; this is done by removing less fit members of the population.

Given two similar chromosomes, it is likely that both will have similar fitness values. Applying the mutation operator on a reasonably fit but non-optimal chromosome will very slightly change the fitness of the chromosome, ideally towards a better fitness. The gradual improvements of solutions in a population will result in the population gradually evolving towards a better fitness.

Two chromosomes which have similar fitness values will not necessarily have similar chromosomes. Different sets of genes may be contributing towards the fitness of the chromosomes. By combining the two chromosomes through crossover, it is hoped that the good genes from one chromosome will combine with the good genes from the other chromosome resulting in offspring which are better than the parents.

Given a problem where all the local minima are close to the global minimum (Figure 2.3), and are much worse than the global minimum, then the genetic operators are likely to effectively lead the population to the global minimum. Solutions that are near the global minimum will have a much better fitness than solutions at the local minima. These are therefore more likely to get selected and to reproduce fitter offspring that will eventually lead to global convergence.

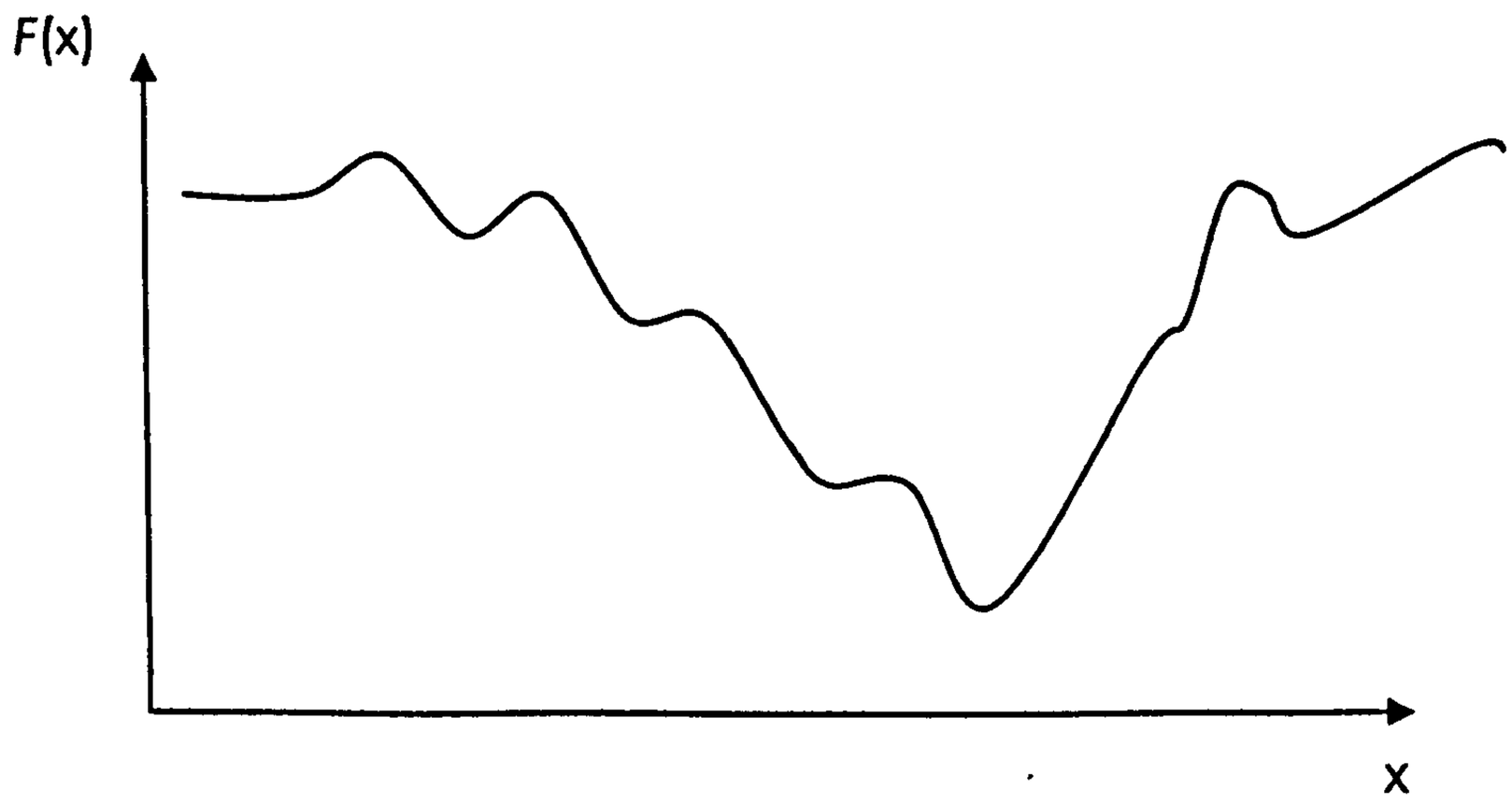


Figure 2-3 A function F(x) that is being minimised. Local optima are all close to the global minimum and have worse fitness. In this scenario it is likely that the genetic operators will easily guide the search to the global minimum.

However, in the scenario shown in Figure 2.4, where there are several local minima with fitness close to the global minimum that are separated by large barriers in the fitness landscape, converging to the true global minimum is more problematic. This is because solutions may exist at any of the local minima basins, and if these groups of solutions are continuously selected for the genetic operators to act on, then the population can very easily converge into one of the local minima. This process is known as genetic drift (Goldberg, 1989).

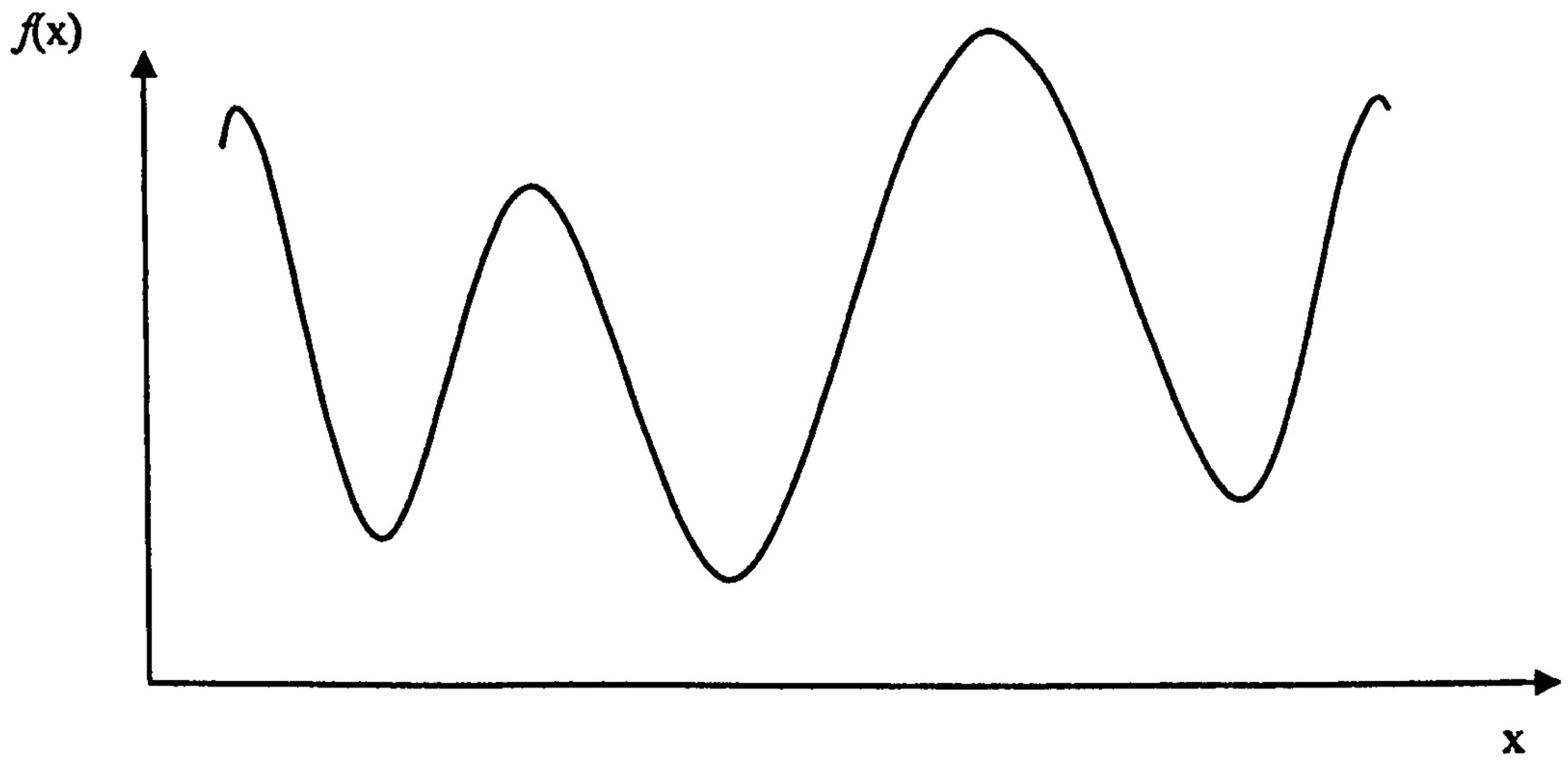


Figure 2-4 A function f(x) that is being minimised. Here the local minima are only slightly worse than the global minimum, are separated by large fitness barriers and the population may converge into any of the local minima.

One tactic employed for avoiding genetic drift is to maintain diversity in the population at all times, so that the population always contains solutions near as many optima as possible. Niching is one process that can be used to maintain diversity and is described in section 2.4.4. The different types of selection and replacement strategies are also discussed in sections 2.4.1 and 2.4.2.

2.3 Multiobjective optimisation techniques

As was discussed earlier, multiobjective optimisation has traditionally been performed by transforming multiple objectives into a single one that is then optimised

as a single objective. One of the more popular of these methods is the weighted sum approach.

2.3.1 Weighted sum approach

The weighted sum approach scalarises a set of objectives into a single objective by multiplying each objective with a user-defined weight to produce a composite function. The first issue that arises here is deciding which weights to select. A guide to setting weights is to decide the relative importance of each objective and setting a weight accordingly. A second issue which must be considered is scaling of the objectives. If the objectives have different orders of magnitude, then it is useful to normalise these values before forming the composite function. Also, it is worth noting that forming a composite function by summing physical quantities of different dimensions does not have a direct physical meaning, i.e. the physical quantities are non-commensurate. Figure 2.5 illustrates how a composite function can find optimal solutions in multiobjective space.

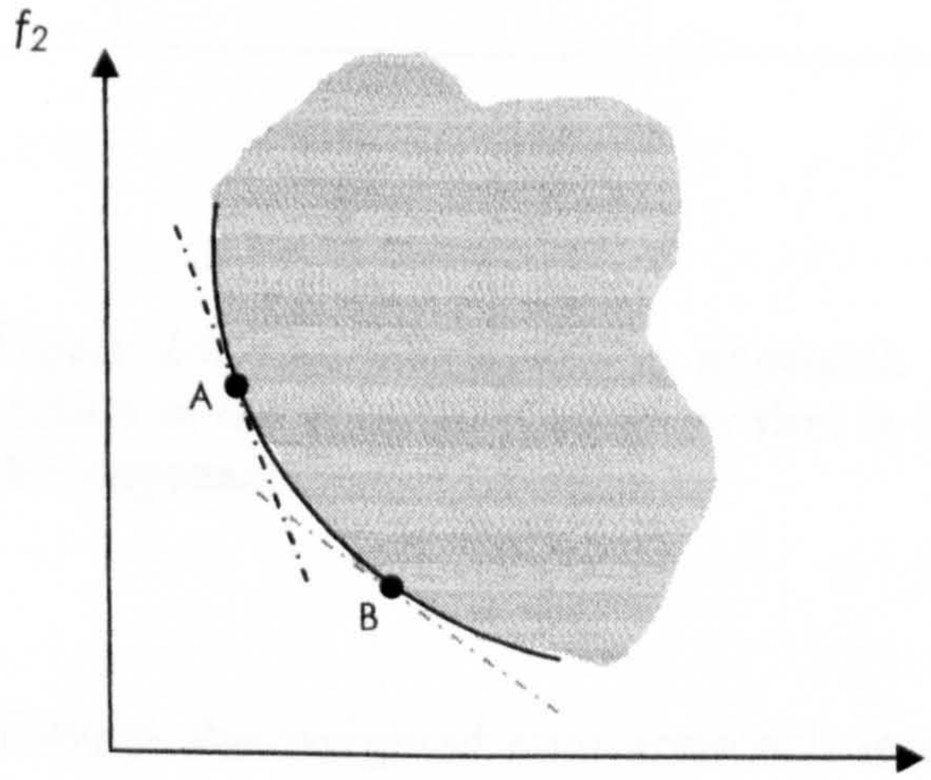


Figure 2-5 A weighted sum approach to finding a single optimal solution in objective space. The gradient of the dotted lines (representing the composite function) is dependent on the weights selected; both points A and B at the tangents of the accessible area of the objective space (coloured grey) have been found by using different sets of weights in the composite function.

Changing the weights of the objectives will result in changes in the gradient of the dotted line, therefore the optimisation process will find a different optimal solution. Theoretically it is possible to find all optimal solutions by changing the weights of the objectives systematically. This however, will only apply if the shape of the objective space region is convex. If the shape of the objective region is non-convex, the weighted sum approach will not find all optimal solutions (Figure 2.6). As the figure shows, certain weights will only find certain solutions, but any solutions that lie on the red portion of the objective space boundary will not be found because it is not possible to find a tangent at that point due to the shape of the curve.

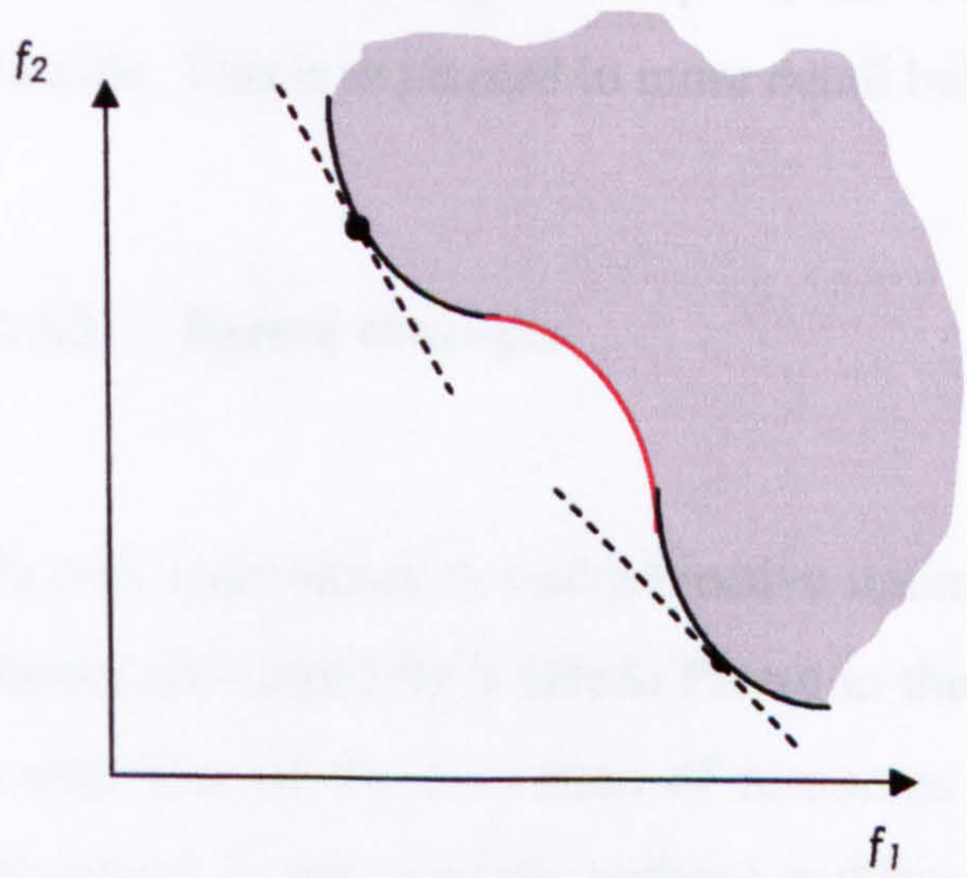


Figure 2-6 The weighted sum approach applied to a non-convex problem. The section of the objective boundary that is coloured in red will not be accessible by this approach.

Though the weighted sum approach is intuitive and theoretically straightforward to implement, optimal solutions will be missed if the shape of the objective space boundary is non-convex. It is also not possible to tell the shape of the objective space boundary *a priori*, which will make it difficult to decide on whether this approach is most suitable for a given problem. Additionally the designation of weights, as mentioned earlier, can be problematic and introduces a subjective element to the optimisation process, which will undoubtedly bias the search for optimal solutions.

Given the assumtion that many objectives control a given situation and are all equally important indicates why multiobjective optimisation aims to find a set of solutions which all display a variety of different objective compromises, or trade-offs. This in turn explains why it is that GAs are employed for this type of search. Since a set of solutions is the output expected, then the population-based nature of GAs is ideal because it allows for the existence of several solutions simultaneously (unlike other search methods, such as simulated annealing and tabu searches). This optimal solution set is said to lie on the Pareto front, named after the Italian economist Vilfredo Pareto, who popularised the notion of multiobjective optimisation. These solutions are collectively known as the Pareto solutions. They are all considered to be equally important- and it is left up to the decision maker to decide which one of the solutions is the most desirable. All the solutions in the population are ranked based on where they lie in objective space, and how they relate to each other in a Pareto fashion. This is explained in more detail below.

2.3.2 Pareto concepts

Recent approaches to multiobjective optimisation have been based on the economic theory developed by Vilfredo Pareto in the 19th and early 20th centuries. The theories centre around the allocation of resources in a society, and how the reallocation of resources is not possible without making at least one individual worse off. This is known as the Pareto optimality or efficiency.

Pareto-optimal distributions can vary. For examples resources can be evenly distributed across all individuals or all resources could be distributed to one individual. The latter would still be a Pareto-optimal distribution if that one individual became worse off by redistributing the resources more equally.

The application of this theory has been adapted to scientific optimisation problems. In multiobjective optimisation of real-life problems with conflicting objectives, no solution can exist which is optimal in all objectives. With a Pareto-optimal solution it is not possible to make an improvement in one objective without worsening the

solution's other objective. There can therefore be many different Pareto-optimal solutions, with different compromises, or balances, of the objectives, to a given problem.

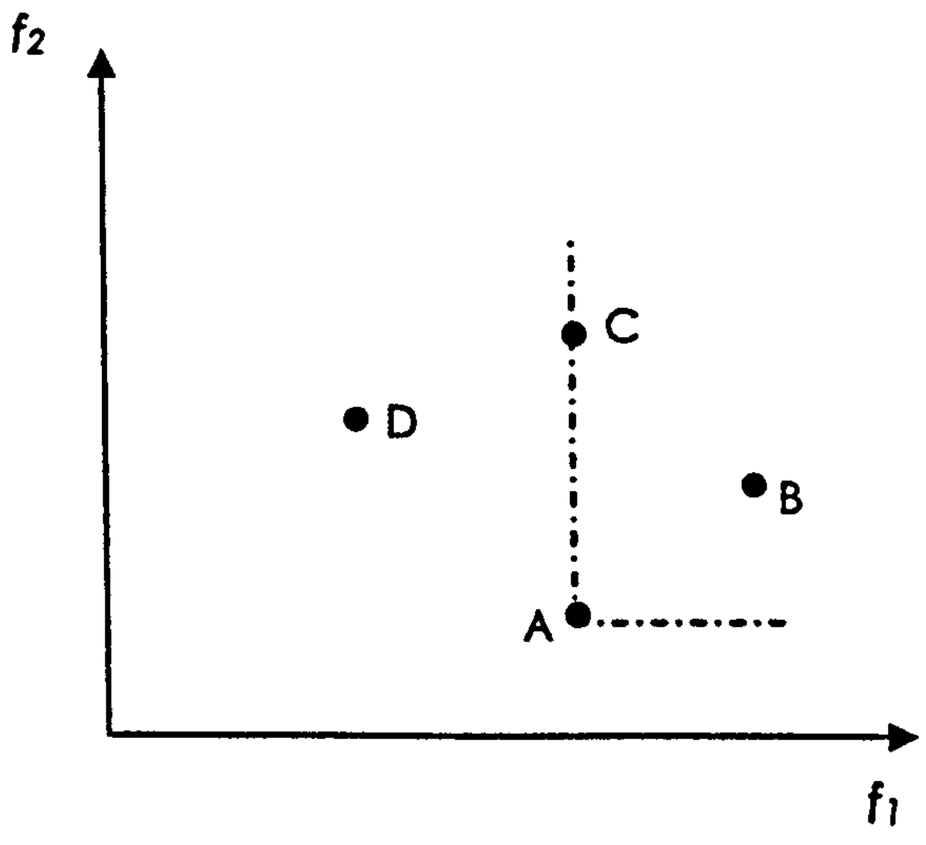


Figure 2-7 Illustration of Pareto dominance. All solutions within shaded area are dominated by A.

2.3.2.1 Pareto dominance

Most multiobjective optimisation algorithms use the concept of dominance to compare solutions to each other. Given two solutions, Y and Z, Y will dominate Z if it is better than Z in at least one objective, and worse than Z in no objectives. Figure 2.7 illustrates this concept. Solution A is better than solution B for both f_1 and f_2. Therefore A dominates B. although C is equal to A in terms of f_1, it is worse with respect to f_2. Therefore A dominates C. Looking at the position of solution D, it can be seen that A is better than D in terms of f_2, but D is better in f_1. This relationship means that neither solution dominates the other. The elegance of this process avoids the need to quantitatively compare two objectives of a solution and to arbitrarily assign values that manipulate the importance of each objective.

2.3.2.2 Pareto optimal set

As the previous section showed, both solutions A and D do not dominate each other, and are not dominated by any other solution shown on the plot. Therefore both these solutions are showing different, but equally valid compromises between the two objectives. These solutions are said to be non-dominated. In a population of solutions, the non-dominated solutions are considered to be equally as valid, due to the different balance in objectives which they display. Also for every other solution in the population there will exist at least one non-dominated solution that is better than itself in all objectives. Therefore all the non-dominated solutions can be considered equally optimal, and are known as the Pareto-optimal, or the Pareto solutions.

2.3.2.3 Pareto frontier

The points in objective space representing the Pareto solutions are said to lie on the Pareto front. If the objective scores are continuous quantities, then theoretically the Pareto front will consist of an infinite number of Pareto solutions on a continuous surface, although some discontinuous parts may exist as well. In practical terms, when dealing with Pareto solutions in a GA population, these cannot form a continuous surface. As Figure 2.8 shows, it can be assumed that the solutions on the Pareto front will dominate any solution that lies with the step-like area in grey. It should not be possible to interpolate between Pareto solutions to form a smooth curve, since no solutions were found in between those points, though they may still exist. In practice, if the gap between two Pareto solutions is small, then forming a continuous surface to link them is reasonable, such as between A and B. However if there is a large gap between the solutions, such as B and C, then it should not be assumed that these are joined by a continuous surface.

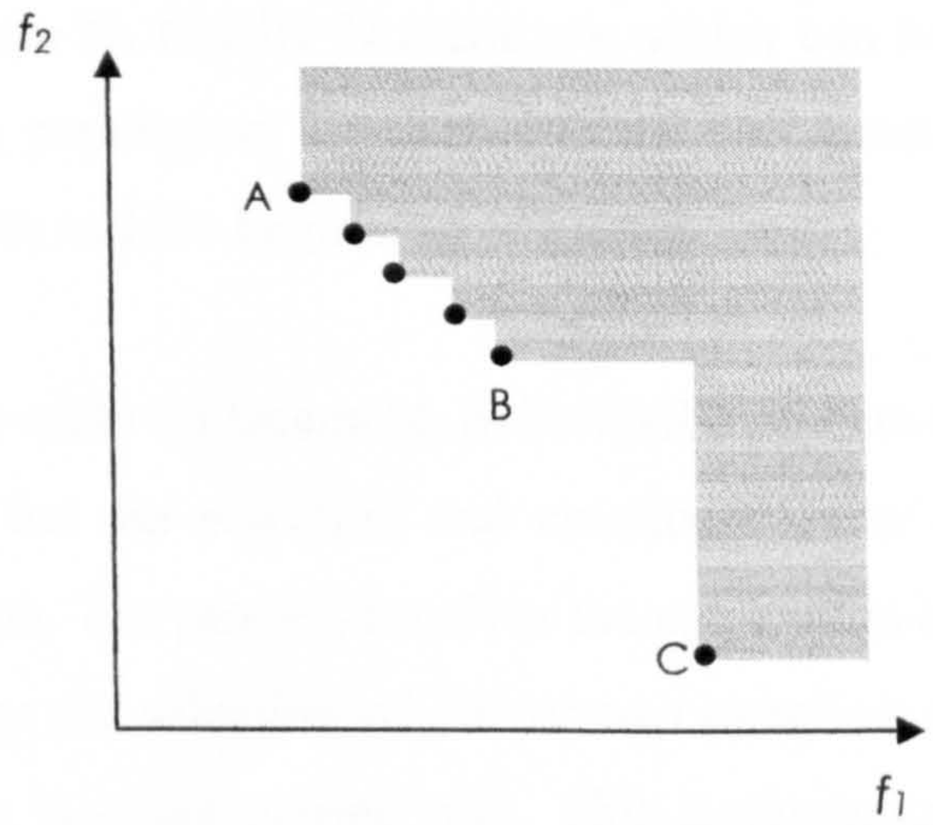


Figure 2-8 The attainment surface for a finite set of Pareto solutions.

2.4 Comparison of GAs and MOEAs

The basic outline of the operation of GAs is described in the preceding sections. However, the details of the implementation vary considerably. Some of these variations apply equally to single-objective and multiobjective GAs, whilst others are relevant only to multiobjective methods. These are discussed below.

2.4.1 Replacement strategies

The replacement strategy describes the way in which a population is maintained and evolved over the period of an algorithmic run. There are two types of replacement strategy, generational and steady-state. The more popular strategy is generational, which involves the potential replacement of an entire population with a new one at each iteration, or generation. During a generation, from a (parent) population of size N, members are selected to fill a mating pool. The process of selection is weighted so that fitter members are more likely to be selected (these can be selected more than once) and the least fit members are less likely to be selected (Deb, 2001). This is

followed by a random selection of members from within the gene pool for crossover and mutation (sections 2.2.1.2 and 2.2.1.3), to result in a new, offspring population, also of size N. Finally N members, which can be selected from both the parent and offspring population, are passed to the new generation, which signals the end of one generation and the beginning of the next one.

In steady-state replacement, individual members of the population gradually undergo change, but the selection and creation of new members are not structured into a generation. The process involves the application of one genetic operator at a time, and begins by the selection of one or two members (depending on whether crossover or mutation is being carried out). This again occurs with a stronger bias towards the fitter members. Selection is followed by the application of the genetic operator, and the replacement of the resulting member(s) with one or two less fit members in the current population.

An issue that may arise with a replacement strategy, the generational replacement strategy in particular, is that non-dominated solutions generated during a run are discarded during the selection process and better solutions are not found and retained in the final population. To ensure that the current, fittest solutions are always retained many algorithms employ elitism which ensures that the current fittest, non-dominated solutions are never discarded, whether through maintaining an archive of these solutions, or through keeping these in the main population, presuming of course, that the number of non-dominated solutions is smaller than the size of the single population. The NSGA-II and PAES described below are examples of algorithms with highly elitist strategies.

2.4.2 Selection methods

Selection is integral to evolutionary algorithms in directing a population towards an optimal solution, the theory of which was discussed in section 2.2.1.4. All selection procedures are ultimately characterised by a stronger bias towards fitter individuals so that the probabilities of their selection is increased.

Proportionate selection is an implementation of a procedure which gives each member of a population a weight related to its fitness; fitter members will have stronger weights, and vice versa. Members with higher weights will therefore a stronger probability of being selected than weaker, less fit members. Roulette wheel selection is a type of proportionate selection (Goldberg, 1989) and, as its name suggests, is analogous to a roulette wheel, where every member of the population is represented by a segment on the wheel, the size of the segment being proportional to the weight, or fitness, of the member. With a "spin" of the wheel, a member is selected; those with higher weights will have a higher chance of being chosen since the size of their segment is larger than weaker individuals.

In tournament selection, the alternative to proportionate selection, tournaments are performed between pairs of members from the population. Two members are selected randomly, the fitter of these "wins" and is added to the mating pool.

2.4.3 Ranking methodologies

With single objective optimisation, the ranking of the population is straightforward: minimising functions require that the lowest scoring chromosome is given the highest rank and vice versa. In multiobjective optimisation, population ranking is more complicated since several objectives must be considered. It is important that the value of each objective for a given chromosome is taken into account separately (summing the objectives defeats the purpose of multiobjective optimisation), and that all non-dominated solutions are considered as jointly optimal. One technique that is implemented is that which was proposed by Fonseca and Fleming (1998a) and which is best described using a graphical representation shown in Figure 2.9.

Figure 2.9 shows solutions to a problem in objective space. The two objectives being minimised are f_1 and f_2, on the x and y axes respectively. The Pareto solutions are represented as closed circles on the Pareto front - these are also considered to be the non-dominated solutions. A non-dominated solution is a solution where an improvement in one objective results in the deterioration of another objective. All

non-dominated solutions are considered to be superior and are given the top rank 0 (as labelled). The rest of the solutions are dominated solutions (open circles) and are ranked according to the number of times they are dominated. For example the solution ranked 1 is ranked so because there is one other solution that is better than itself in *both* objectives. A simple way of obtaining this information is to draw perpendicular lines to the axes from a solution and counting the number of solutions within the rectangle (shown for one of the solutions ranked 0 and the solution ranked 4).

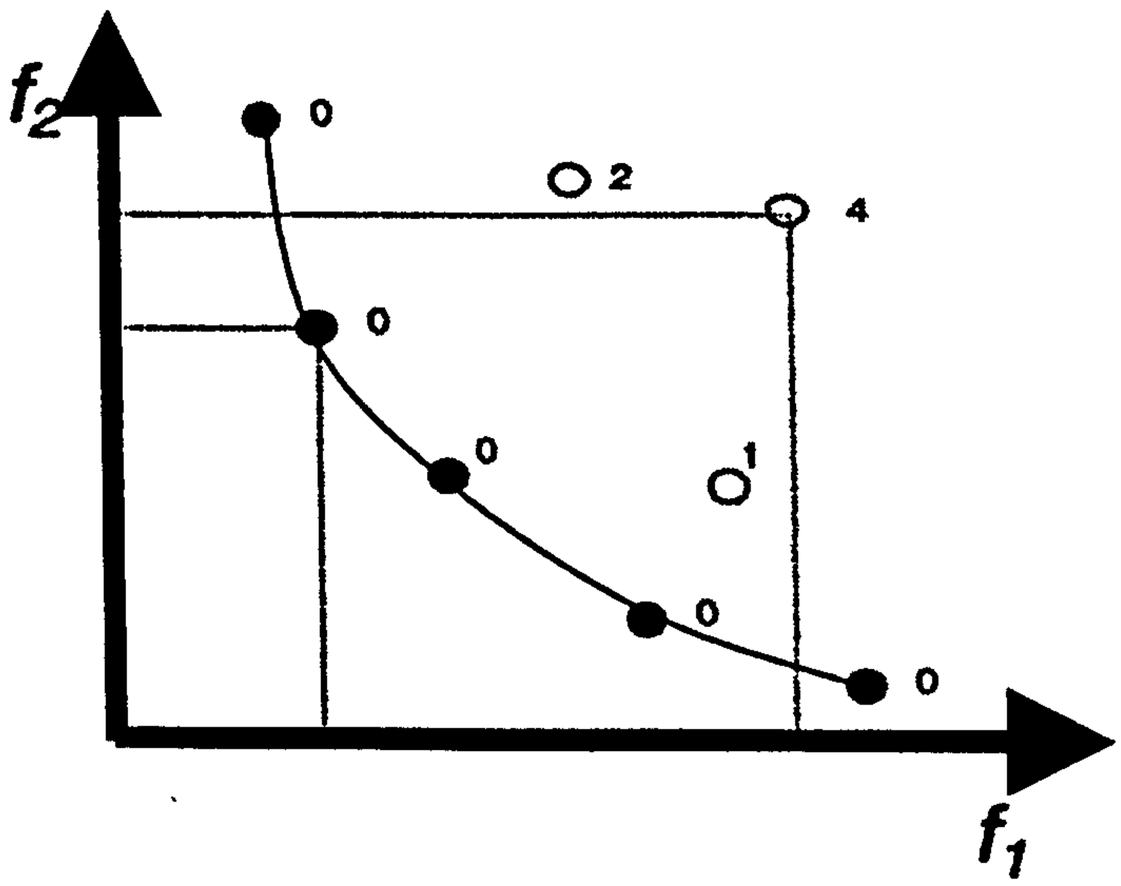


Figure 2-9 Pareto ranking: Solutions in objective space, where the two objectives, f_1 and f_2 are being minimised. Solid circles represent the Pareto or non-dominated solutions. The open circles are the dominated solutions and are ranked as indicated. The number of times a solution is dominated can be determined by drawing perpendicular lines to the axes and counting the number of solutions that fall within the rectangle (from Gillet *et al.*, 2002).

An alternative ranking scheme is to first give all the non-dominated solutions a rank of 0. These are then temporarily ignored and the non-dominating solutions in the remainder of the population are determined and given a rank 1. These are then also ignored, non-dominating solutions are determined from the remainder of the population and these are given a rank of 2. The process continues until all the members of the population have been given a rank.

2.4.4 Niching

Preserving diversity is essential during multiobjective optimisation. Diversity is also necessary in single-objective optimisation, since a population consisting of chromosomes with diverse genes is more likely to find the optimal solution than a population consisting of similar, mediocre, chromosomes. Because solving multiobjective problems generates a set of optimal solutions, then ideally these also must be as diverse as possible, and widely spread out on the Pareto front. Diversity along a front ensures that a good representation of varying compromises by the solutions has been achieved. Niching is the technique that is most widely implemented to preserve the diversity of a population and is modelled on the behaviour of species in biological macrosystems. Individuals within a species competing for a finite amount of resources in an environmental niche will have a reduced fitness, whereas those that occupy less dense niches will have more resources available, which in turn will increase their fitness. In an analogous manner a given member of a population will exist in a niche, specified by a niche radius (Figure 2.10). The selection probability of the chromosome is reduced to a level that is correlated to the number of chromosomes which fall within its niche. In this way chromosomes in sparse niches have a higher likelihood of being selected. This mode of niching which influences selection probabilities of chromosomes is known as sharing. The alternative is known as crowding, and this acts by restricting the number of individuals that may exist in a specific niche. This can be achieved by, for example, not accepting a new solution if too many other solutions exist in the same niche or, if using a steady-state replacement strategy, then a solution from the same niche is removed to make space for a new solution.

To quantify niching, the "distance" between solutions must be measured in some way. In multiobjective optimisation, this can be done in objective, or in decision space. Objective space niching measures the Euclidian distance between solutions in objective space, using the magnitude of the solutions' individual objectives. The definition of the decision space depends on the problem, and is usually a characteristic that is exhibited by the solution.

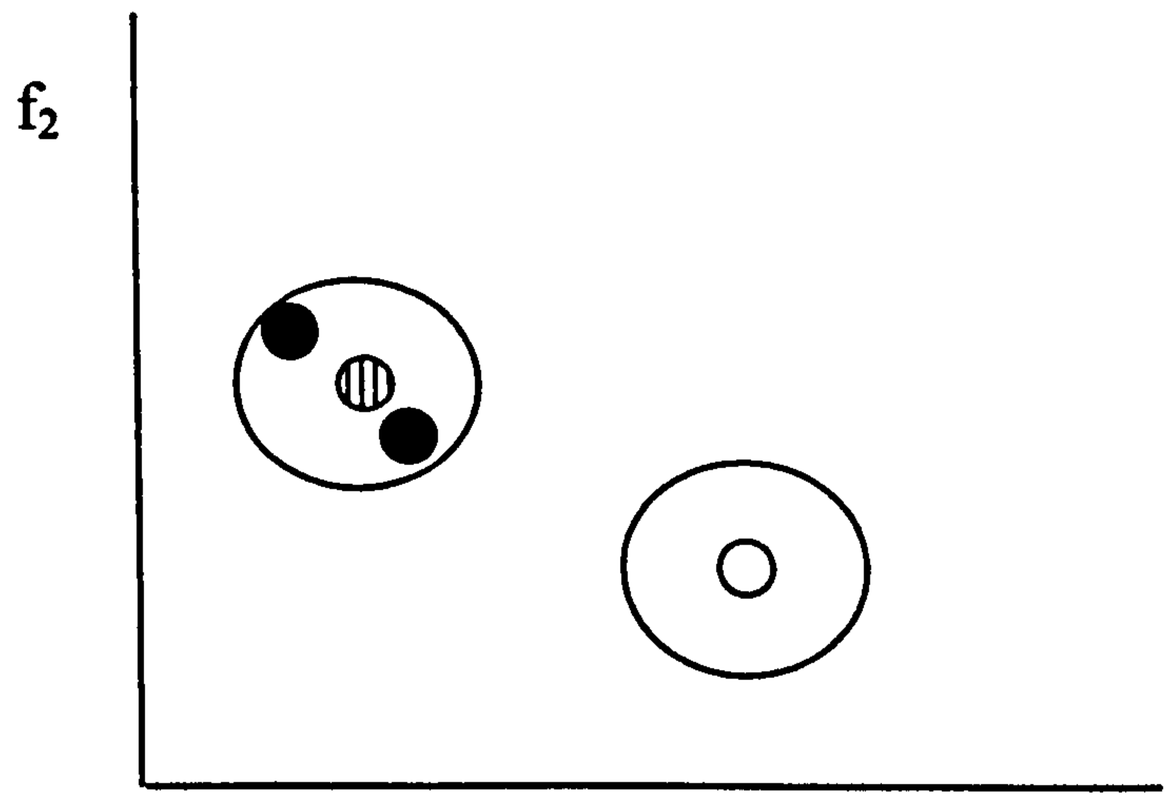


Figure 2-10 Niching in objective space. The niche radii of two solutions: the first shown as a striped circle and the second as an open circle. The striped circle solution will have a reduced niche fitness because two other solutions lie within its niche radius whereas the open circle solution will have the maximum fitness since no other solution occupies its niche. Therefore the open circle solution has a higher probability of being selected relative to the striped solution. This allows for diversity along the Pareto rank and ensures that potentially interesting outliers are not lost.

2.5 MOEA implementations

The previous section has shown the different aspects of multiobjective optimisation and described the features which control the process. Several different algorithms, which implement these features in various ways have, as a result, been developed. An outline of some of these is given below.

2.5.1 Vector Evaluated Genetic Algorithm

The Vector Evaluated Genetic Algorithm (VEGA) was published in 1984 by Schaffer and is possibly the first GA which aims to find non-dominated solutions. The VEGA is a simple extension of a single-objective GA, and operates by ranking certain members of the population by a given single objective and independent of the other

objectives. Given that M objectives are to be handled, the GA population is, at every generation, split into M subpopulations randomly. Each subpopulation is given a fitness value according to one of the objectives; therefore each subpopulation will be evaluated based on one of the objectives. The selection operator is then used to fill a subpopulation mating pool, using only members from a single subpopulation. In this way, members which are particularly good in one objective are likely to be represented more than once within the mating pool. The mating pools from all sub-populations are combined, and the genetic operators, crossover and mutation, take place. Since the focus of this method is per objective, then members who are particularly good at a single objective will be favoured. Schaffer proposed that by allowing crossover between different members of the subpopulations, solutions may be created which are good at more than one objective. It was noted however, that crossover between solutions which excel at different objectives did not succeed at finding diverse solutions with good compromises of the objectives, and eventually the VEGA converges to a few solutions that are good at only one of the objectives. The main advantage of a VEGA is in the simplicity in the implementation. It is also suitable for problems where best solutions per objective are desired.

2.5.2 Multiobjective genetic algorithms

A multiobjective genetic algorithm (MOGA), developed by Fonseca and Fleming (1998a) was the first algorithm that was designed to explicitly emphasise non-dominated solutions and simultaneously maintain diversity along a front. The MOGA used the Pareto ranking scheme described in section 2.3.2, where the rank of a member is determined by the number of members it is dominated by. The algorithm is generational and uses a fitness-sharing technique for its niching. Niching takes place in objective space between solutions of the same rank.

The MOGA also uses the concept of "goals" as well as regular objectives in its selection of members. A goal is an objective for which a member needs to achieve a value over a certain threshold regardless of what it achieves in other objectives.

2.5.3 Pareto Archived Evolutionary Strategy

The Pareto Archived Evolutionary Strategy (PAES), by Knowles and Corne (2002), uses a steady-state, rather than a generational strategy, and was originally designed for solving real-world telecommunications network design problems. It differs from other multiobjective algorithms in that the crossover operator is not applied and only the mutation operator is used to produce offspring. This algorithm therefore uses a local search strategy, rather than a global search, to find non-dominated solutions. A PAES, as its name may imply, also keeps an archive of the best solutions found during the run, which has a limit in terms of the numbers of solutions it can hold.

During the search process, a random member, *p*, is selected from the population and mutated to produce an offspring *c*. If *c* dominates *p*, *c* is added to the archive of good solutions. If *p* dominates *c*, *c* is discarded and another mutated solution is created for processing.

In scenarios where both *p* and *c* are non-dominated relative to each other, then *c* is compared to members in the archive. If there are solutions in the archive which dominate *c*, then this is discarded and *p* is mutated again to produce another offspring. If *c* dominates any solutions in the archive, then these are deleted and *c* is added to the archive. If *c* is not dominated, and does not dominate any solution in the archive, then it is only accepted in the archive if a free slot is available. On the other hand if *c* occupies a less crowded region in the search space, then it is retained and replaces a randomly-selected member from a more crowded region.

This version of the PAES is known as (1+1). Other versions have been developed which vary in terms of the number of solutions generated by mutating the parent *p* λ times, to produce λ offspring, one of which is then selected to be added to the archive, in the same way as described above. This version is known as $(1 + \lambda)$. In the $(\mu + \lambda)$ version, there are μ current solutions and one of these is selected as a parent by tournament selection. λ offspring are then generated from this parent via mutation and one of these is selected, as described above, for addition to the archive.

2.5.4 Elitist Non-dominated Sorting Genetic Algorithm (NSGA-II)

An NSGA-II is a highly elitist method developed by Deb in 2000. An offspring population of size *N* is created from a parent population of size *N*, and these two are combined to form 2*N*, upon which a Pareto ranking process is carried out. This ensures that relatively good solutions are never lost from the population. After the population is ranked and sorted, solutions in 2*N* are taken, rank by rank, to fill the offspring population of size *N*, meaning that the lowest ranking *N* solutions are discarded. The NSGA-II uses a crowded tournament selection operator to select parents for crossover. This works in the same way as standard tournament selection, with the exception that if both members are in the same rank then a crowding distance measure is used for selecting one of the two members. The crowding distance is a fitness-sharing scheme with the advantage of not requiring a user-defined measure to define the size of a niche. Its estimation is based on the distance of a given solution to adjacent solutions per objective. The crowding distance measure may also be used in the step of reducing 2*N* down to *N*. This situation may arise when the final rank that is being added to the offspring population *N* contains more solutions than available slots. In this case the crowding distance measure is applied to select enough solutions from the least crowded areas to fill the final vacant slots of the offspring population.

2.6 Multiobjective Optimisation in chemoinformatics and bioinformatics

The optimisation methods of GAs have been applied to various problems in chemoinformatics and bioinformatics. Section 3.4.3.5 discusses the application of GAs to protein-ligand docking. Multiobjective optimisation has also been applied to several chemoinformatics and bioinformatics problems. These applications have been reviewed by Nicolaou *et al.*, (2007). One of the first chemoinformatics applications for Pareto based GAs was the GAMMA program for the superposition of flexible molecules (Handschuh *et al.*, 1998; Handschuh and Gasteiger, 2000). The program optimises two criteria, the number of matching atoms between two molecules – which has to be maximised – and the devitations of the coordinates of the superimposed atoms – which is minimised. The geometric fit between two conformations is further

improved by changing torsional angles. This is done using a GA and a directed tweak method.

Other recent multiobjective approaches include:

- The MoSELECT program (Gillet *et al.*, 2002; Wright *et al.*, 2003) applies a MOGA to combinatorial library design. In the original version, up to six objectives were optimised – the cost of producing the library, the diversity of the library, and the profiles of four physicochemical properties. The size of the libraries and the configuration (number of reactants needed to generate the products) were fixed throughout a given run. A later version of the program allowed the latter two properties to be varied during a run, and hence be treated as additional optimisation objectives.

- A method for generating multiple pharmacophore hypotheses using a MOGA (Cottrell *et al.*, 2004). The method looks for a representative ensemble of overlays that show different structure-activity hypotheses, in a concerted manner, and allowing full conformational flexibility. Three objectives are optimised simulatenously, the feature score which is a similarity score that measures the degree to which feature alignments are optimal, the van der Waals energy of individual conformers, and the volume integral of the overlay. A later version of the method allowed for partial matches to exist between ligands within a set (Cottrell *et al.*, 2006).

- Brown *et al.* (2004) applied a multiobjective approach to the generation of median molecules. The aim is to generate a series of molecules, known as *median molecules*, that are as similar as possible to each of a set of two or more input molecules. The number of objectives is equal to the number of input molecules, and the similarity of the median molecules to each of the input molecules forms an objective. Since the molecule that is most similar to each input molecule is that molecule itself, the objectives are in conflict.

- In bioinformatics, multiobjective optimisation has been applied to clustering (Handl and Knowles, 2005). Their MOCK algorithm optimises two objectives, what the

authors describe as "the compactness of clusters", and the "connectedness of data points". The algorithm has the capacity of selecting a good solution from the Pareto front, and automatically determines the optimal number of clusters in a given set.

2.7 Summary

In this chapter, the theory of multiobjective optimisation in relation to single-objective optimisation was described, as well as the principles of evolutionary algorithms. Fundamental Pareto techniques in multiobjective optimisation were discussed, and their application to evolutionary algorithms. Various applications of evolutionary algorithms in chemoinformatics and bioinformatics were also cited.

3 Docking and scoring

3.1 Molecular recognition

Most biological activity begins with a union of individual elements to form one entity. If one of the elements is smaller than the other, then it is usually referred to as the ligand, and the latter as the receptor. Thus in an enzyme-substrate complex, the substrate is the ligand, and the enzyme is the receptor. Other examples of receptors to which ligands bind are antibodies, DNA, and membrane-bound proteins. The action of drugs tends to be synonymous to these interactions; a drug will bind to a receptor protein to initiate a desired therapeutic effect. Molecules need to recognise each other in order to unite, or bind together, thus triggering the cascade of events that leads to a biological outcome.

One of the first theories that attempted to explain how molecules recognise each other, or molecular recognition, was proposed by Emil Fischer in 1894, using the "lock and key" hypothesis. This theory envisaged molecules as wooden puzzles that need to have a geometric match in order to interact with each other. This theory was modified in the late fifties by Daniel Koshland, who proposed the "induced fit" model; molecules induce changes in their conformations as they bind to each other; Koshland likened the process to a hand fitting a glove, the glove is the receptor, and the hand is the ligand (Koshland, 2004).

Molecular recognition has been studied through supramolecular chemistry, an interdisciplinary field which explores molecular interactions through specially designed artificial systems. The 1987 Nobel Prize for chemistry was awarded to Cram, Lehn and Pederson for their work in this field (Cram, 1988). Their work on "host-guest" complexes in particular, was cited as important. Host-guest systems describe the study of complexes composed of two molecules and held together by non-bonded interactions, and which are used in the detailed analyses of the binding properties of molecules. This type of information can be used to understand

biological processes and subsequently help in the selection and design of ligands which can manipulate these processes.

3.2 Energetics of protein-ligand interactions

The interactions between a protein and a ligand, or drug, takes place if the reaction is energetically favourable. Though covalent bonding between protein and drug can occur, these interactions tend to be non-bonded. The association of molecules and the affinity towards each other, is driven by the thermodynamics.

3.2.1 Affinity and dissociation constants

The process of non-covalent binding between ligand *A* and protein *B*, is described by the changes in enthalpy and entropy of a system. The system consists of free protein and free ligand molecules and solvent, and bound molecules *A* and *B* with solvent. The association relationship between the molecules is described as:

$$A + B \underset{k-1}{\overset{k+1}{\rightleftharpoons}} AB \qquad \text{Equation 3.1}$$

A represents the uncomplexed ligand and *B* represents the uncomplexed protein. *AB* is the complexed ligand and protein. k_{+1} is the association rate constant for the reaction going from left to right and k_{-1} is the dissociation rate constant going from right to left.

The binding affinity between two molecules can be expressed as the dissociation equilibrium constant (molar) at the thermodynamics equilibrium, K_d, shown by:

$$K_d = \frac{[A][B]}{[AB]} \qquad \text{Equation 3.2}$$

or by the association equilibrium constant K_a (in M^{-1}), given by:

$$K_a = \frac{[AB]}{[A][B]} \qquad \text{Equation 3.3}$$

These equilibrium constants can be determined experimentally by measuring the concentrations of *A*, *B*, and *AB*. The following equation relates the equilibrium constant to the change in Gibb's free energy of dissociation of *AB*

$$\Delta G = \Delta G^o - RT \ln (K_d) \qquad \text{Equation 3.4}$$

G is Gibb's free energy constant, ΔG is the change in free energy for the reaction, *T* is the absolute temperature and *R* is the gas constant. ΔG^o is the free energy change of the reaction under standard conditions. Standard conditions are denoted by a 1 M concentration of all reactants and products, T = 298 K and pressure is at 1 atm. At equilibrium $\Delta G = 0$, therefore:

$$\Delta G^o = RT \ln (K_d) \qquad \text{Equation 3.5}$$

ΔG^o is the binding free energy of an interaction. This is made up of two componenets, entropy and enthalpy, which are associated with ΔG^o through:

$$\Delta G^o = \Delta H^o - T\Delta S^o \qquad \text{Equation 3.6}$$

ΔH^o represents the change in enthalpy of a system and $T\Delta S^o$ is the change in entropy. If ΔH^o is positive (i.e. unfavourable energy value) then the interaction is described as being entropy-driven and $T\Delta S^o$ is positive (so $-T\Delta S^o$ is favourable). If the signs for both terms are negative (i.e. ΔH^o is favourable and $-T\Delta S^o$ is unfavourable) then the reaction is regarded as enthalpy-driven.

Changes in enthalpy of a system can be measured experimentally by isothermal calorimetry or ITC (isothermal titration calorimetry), which measures binding equilibrium directly by using sensitive calorimeters which can measure ΔH and K_d in a single experiment. Protein-ligand interactions can also be determined using IC_{50}

values. The IC_{50} measures the concentration of inhibitor required to reduce the binding of a ligand (or the rate of reaction) by half. The IC_{50} is not suitable for theoretical studies because its value depends on the amount of ligand available to the receptor, which makes comparisons between data obtained under different conditions unfeasible. On the other hand K_d values from different experiments can be compared, assuming that these were performed under equilibrium conditions (Ajay and Murcko, 1995).

3.2.2 Computational free energy calculations

Using computational methods to accurately calculate binding affinities/free energies are essential in biomolecular simulations. Energy functions are used to estimate the binding energy between a ligand and a protein. Though there are several methods which are used to estimate free energies, these vary in terms of the levels of theory they employ and also in terms of their speed. Currently these methods are not capable of estimating binding energies to the level of experimental methods (Gilson and Zhou, 2007), though they continue to have huge potential, particularly in the field of structure-based design for the discovery of drugs (section 3.4). These methods are likely to improve as our understanding of physical principles increase and with improvements in computer hardware and speed.

3.2.3 Free Energy Perturbation

Free energy perturbation (FEP) is a computational technique that can be used to calculate relative binding free energies (Bash *et al.*, 1987). Free energy changes are defined by the initial and final thermodynamic states, regardless of the path taken to get from one state to the other. By making small changes in a molecule's atom composition, and calculating the free energy between each change, it is possible to sum all the free energies from each of the changes to get the free energy between the starting and the final molecule. For example, given two molecules A and E, the difference in free energy between both can be calculated by taking into account

intermediate structures B, C and D. Therefore the free energy between each of the intermediate molecules can be summed up to give the difference in free energy between A and E (Figure 3.1).

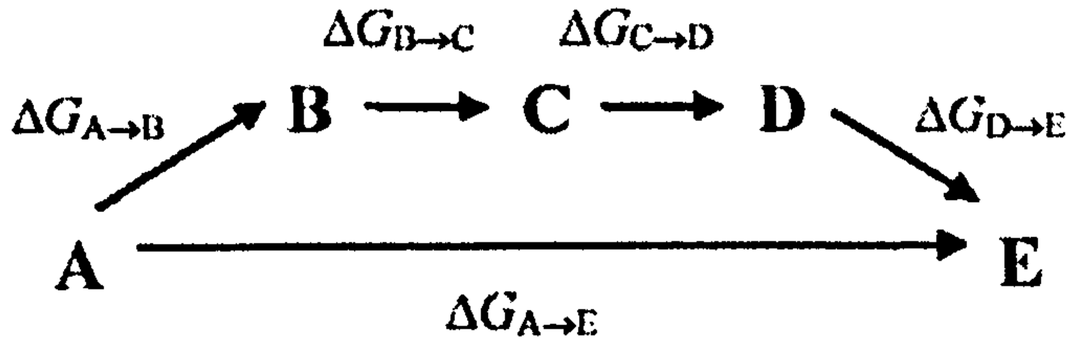


Figure 3-1 Free energy of change from moving between molecule *A* and *E* is represented by $\Delta G_{A\to E}$. This is calculated by calculating the changes in free energy between intermediates, B, C and D which is represented by $\Delta G_{A\to E} = \Delta G_{A\to B} + \Delta G_{B\to C} + \Delta G_{C\to D} + \Delta G_{D\to E}$.

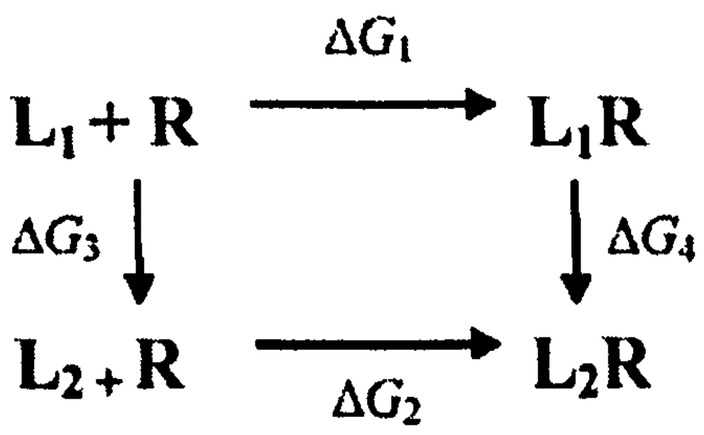


Figure 3-2 Thermodynamics cycle of inhibitor ligands L_1 and L_2, binding to receptor R.

These principles can be applied to thermodynamics cycles, which attempt to calculate the relative binding energy between two different complexes, L_1R and L_2R in Figure 3.2. The difference in free energies to form the two complexes can be written as ΔG_2 - ΔG_1, or as $\Delta\Delta G$. In principle it is possible to calculate $\Delta\Delta G$, ΔG_1 and ΔG_2 by simulating the process, but in practice this is difficult as it would entail the reorganisation of receptor, ligand and solvent, and would require large computational power for the extensive sampling of the phase space. It is easier to, instead, focus on ΔG_3 and ΔG_4. Since the value around a thermodynamics cycle is zero, then $\Delta G_2 - \Delta G_1 = \Delta G_4 - \Delta G_3$, then $\Delta\Delta G$ can be calculated from ΔG_3 and ΔG_4. Though physically ΔG_3

and ΔG_4 cannot be performed in the laboratory, they can be simulated computationally by mutating L_1 to L_2 in solution and also within the receptor. Taking this route rather than using ΔG_1 and ΔG_2 is much more reliable as this involves less reorganisation of the system.

3.2.3.1 Empirical factor models

Calculating the Gibbs free energy of binding can be very time-consuming and complex because the calculation requires long, all-atom molecular dynamics simulations in order to cover the full phase space and obtain reliable free energy estimates. It is possible to use methods which approximate and simplify these calculations. By breaking down the Gibbs free energy of binding into separate, tangible components that are calculated separately and then summed together, it is possible to infer a good estimate of the free energy of binding. Force fields employ what is known as a "master equation" that comprises all, or some, depending on the level of approximation needed, of the energetic factors involved in a ligand binding reaction. One representation of the master equation is as follows:

$$\Delta G_{bind} = \Delta G_{int} + \Delta G_{solv} + \Delta G_{conf} + \Delta G_{motion} \qquad \text{Equation 3.7}$$

ΔG_{bind} is the binding free energy, ΔG_{int} is the change in interaction energy, ΔG_{solv} is the change in solvation energy, ΔG_{conf} is the change in conformational energy, and ΔG_{motion} is the change in molecular motion.

Empirical factor models are much less computationally exhaustive than FEP, though it is important to be aware of their relatively simplistic nature compared with more exhaustive methods. Reasons which contribute to this are, for example, the disregard of certain terms, such as entropic terms, which are difficult to estimate. Also it is difficult to weigh each of the terms in the master equation to give a good approximation of the free energy of binding. The other point to be aware of is that often the individual energy components comprise large, approximate, numbers may be both favourable (i.e. negative energy values) and unfavourable (positive energy

values). By summing these approximate values together, a small total binding free energy is obtained which, because it was derived from large approximate numbers, may likely be prone to statistical and systematic errors (Kollman, 1993). In the next section the individual terms of the master equation shown in Equation 3.7 are discussed.

3.2.3.1.1 Change in interaction energy (ΔG_{int})

This term describes the non-bonded interactions between a ligand and protein, in particular the electrostatics and van der Waals (vdw) interactions. These terms form part of the enthalpic contributions to binding and are discussed further in section 3.2.3.2.4.

3.2.3.1.2 Change in solvation energy (ΔG_{solv})

Change in solvation energy is represented by ΔG_{solv}, and represents the contribution of water to the free energy of a system. Biological systems where the binding of a protein and a ligand take place occur in an aqueous environment. In their free, unbound states, the surfaces of a protein and ligand are surrounded by water. When the two molecules bind and form a complex, the complex is also surrounded by water, but the actual surface available for water- molecule contact (the solvent accessible area) has been reduced. Conversely the surfaces of molecules buried from the solvent (the solvent inaccessible area) have increased. The reduction in solvent accessible area is termed the desolvation energy and will contribute towards ΔG_{bind}. The approach of modelling a system's water will depend on the level of approximation necessary; explicit modelling of desolvation include solving the Poisson-Boltzmann equation and the Generalised Born method (Lee *et al.*, 2005). Implicit solvation calculations can be performed using the solvent accessible surface area (Dill *et al.*, 2005).

3.2.3.1.3 Change in conformational energy (ΔG_{conf})

The energetic contributions made to ΔG_{bind} through changes in conformation of the protein and ligand are represented by ΔG_{conf}. The degree of conformational changes made by the protein and ligand upon binding differs between complexes. For most complexes the protein side chains may change substantially whereas changes in the conformation main chain are minimal. There are proteins which do undergo large changes in conformation of the main chain upon binding, such as HIV protease. The enthalpic contribution to free energy from conformational changes can be modelled using a molecular mechanics force field. The entropic component of the protein and ligand ΔG_{conf} is more problematic to compute. A ligand, in its free state, adopts several conformations and, when bound to the protein, does so in only one conformation, resulting in loss of conformational entropy. Sampling all possible conformations of the ligand in its free state is impractical, so a method of circumventing this issue is to estimate the number of conformational bonds in the ligand, and so this term is usually estimated by counting the ligand's number of rotatable bonds. Changes in movement of the protein main chain and side chains also contribute to entropic losses. It is possible to estimate this by looking at the distribution of side chains from experimentally determined structures (Pickett and Sternberg, 1993).

3.2.3.1.4 Change in molecular motion (ΔG_{motion})

The change in energy due to motion of molecules is represented by ΔG_{motion}, and is represented by changes in rotation, translation and vibrational movements of the molecules. Upon binding, the six degrees of freedom, three translational and three rotational, are lost, thus lowering the entropy of the system and negatively contributing to the free binding energy.

3.2.3.2 Molecular mechanics force fields

Molecular mechanics force fields have been developed to perform calculations of energy on systems containing large numbers of atoms. They are ideal for these types of systems because they ignore the electrons in the system and model atoms based on their nuclear positions, thus substantially reducing computational time. With biomolecules force fields are described as being two-body additive, meaning the potential energy is estimated as a function between pairs of two atoms. A potential energy function calculates the potential energy of a system and is a function of the position of the atoms under study. Energy terms which can be applied to estimate the potential energy of a system are shown in the following equation:

$$E = E_{bond} + E_{angle} + E_{torsion} + E_{vdw} + E_{elec} \qquad \text{Equation 3.8}$$

E_{bond} represents the stretching of bonds, E_{angle} is the opening and closing of valance angles, $E_{torsion}$ represents torsional terms, and E_{vdw} and E_{elec} are the non-bonded vdw and electrostatic interactions respectively.

Transferability of a force field and the parameters which define the terms in Equation 3.8 is an important feature of force fields as it allows the same sets of parameters to model related molecules, and avoids having to generate a new set of parameters for every molecule. This is also important for making predictions; a force field that has been parameterised on a given molecule type can be applied to a related molecule and, theoretically, a reasonable estimation of the potential energy can be achieved. Force fields also utilise the concept of an atom type; every atom in the system under study is assigned an atom type, containing information about the atom, such as its number, hybridisation states, and local environment. For example, carbon atoms with sp^1, sp^2 and sp^3 hybridisation states will be assigned individual atom types.

3.2.3.2.1 Bond stretching

The energy of a bond between two atoms is said to be at its lowest when its length is at its "natural", equilibrium length. A bond is stretched when its length deviates from its equilibrium state. When two bonded atoms are brought close to each other, their electron clouds overlap, thus increasing the energy of the bond. Similarly if the bond is stretched beyond its equilibrium point its energy will increase. Hooke's law is usually applied to describe bond stretching.

$$v(l) = k/2\ (l - l_0)^2 \qquad \text{Equation 3.9}$$

$v(l)$ represents the potential energy, k is a stretching constant of the bond, l_0 is the reference bond length and $(l - l_0)$ is the change in bond length. The reference bond length is the value that a bond adopts when all other terms in the force field are set to zero.

3.2.3.2.2 Angle bending

Angle bending describes the bending of a valence angle. The valence angle describes the angle between the two bonds of three atoms that are bonded together consecutively. Similar to bond stretching Hooke's law is also applied to estimate the deviations of a given angle from a reference state:

$$v(l) = k/2\ (\theta - \theta_0)^2 \qquad \text{Equation 3.10}$$

$v(\theta)$ is the potential energy, k is a constant, θ_0 is the reference angle and $(\theta - \theta_0)$ is the displacement of the reference angle.

3.2.3.2.3 Torsional terms

Torsional terms account for rotations about a covalent bond, which tend to have a large impact on the conformation of a molecule. Torsional angle interactions differ from bond stretching and angle bending interactions in two ways. The internal rotational energy barriers of a torsion bond are quite low therefore changes in torsion angles can be large. Secondly change in the torsional potential is periodic through a 360° rotation. Therefore, the torsion angles between atoms determine the torsional potential and E_{tors} can adopt many different forms depending on the atoms forming it. Hence, torsional terms in force fields model a variety of different potentials. Torsional potentials are usually expressed as a cosine series expansion, such as:

$$v(\omega) = \sum_{n=0}^{N} \frac{V_n}{2}[1 + \cos(n\omega - \gamma)]$$

Equation 3.11

where, ω is the torsion angle, and *Vn* is often referred to as the 'barrier height,' giving the relative energy barriers to rotation. *n* is the multiplicity and gives the number of minimum points in the function as the bond is rotated through 360°. γ is the phase factor and defines where the torsion angles passes through its minimum value.

3.2.3.2.4 Non-bonded interactions

Non-bonded interactions in a force field usually comprise the electrostatic and vdw interactions. These forces are distance-dependent and are usually expressed between non-bonded atoms. The non-bonded interactions of 1,2 and 1,3 atom pairs (atoms separated by one and two covalent bonds respectively) are usually not considered; the bond stretching and angle bending terms usually suffice in these cases. 1,4 bonds or greater are usually considered because these affect the conformational energies of a molecule.

3.2.3.2.4.1 Electrostatic interactions

Because of differences of electronegativity of atoms, the electron cloud on a molecule tends to be unequally distributed. A couloumbic electrostatic potential is used to calculate these interactions, as shown by the equation:

$$E_{elec} = 332 \frac{q_i q_j}{r\varepsilon} \qquad \text{Equation 3.12}$$

i and *j* are the two atoms between which the interaction energy is being calculated, q_i and q_j are the atomic charges for *i* and *j* respectively, *r* is the distance between the two atoms and ε is the dielectric constant. 332 is a constant for expressing the value in units (kcal mol^{-1}). Estimation of the dielectric can be problematic, and different dielectric constants can vary the electrostatic interactions substantially. Examples of dielectric values implemented are $\varepsilon = 1$ for a vacuum and $\varepsilon = 80$ for bulk water. Dielectric constants are usually selected during parameter optimisation to fit empirical data.

3.2.3.2.4.2 Van der Waals (vdw) interactions

vdw interactions describe dispersion, repulsion and induction between atoms. They determine the shape and volume a given atom occupies. Dispersion forces are due to correlations between electrons in different atoms; they lower the overall energy and are therefore an attractive force. Repulsion is a positive force that is due to electron clouds overlapping. Induction is due to the distortion of the charge distribution of an atom by a neighbouring atom. The strength of the vdw term E_{vdw} is distance dependent. At small distances between atoms the repulsion force is unfavourable, making E_{vdw} unfavourable and as the interatomic distance decreases E_{vdw} approaches infinity. At larger interatomic distances the dispersion interactions first result in favourable, negative values for E_{vdw}; as the interatomic distance increases to infinity,

E_{vdw} value approaches 0. These distance-dependent changes in E_{vdw} values are most frequently described using the Lennard-Jones '12-6' function.

$$E_{vdw} = \varepsilon\left(\frac{A_{ij}}{r^{12}}\right) - \left(\frac{C_{ij}}{r^{6}}\right) \qquad \text{Equation 3.13}$$

r is the distance between a pair of atoms and ε is the vdw well depth. A is equal to $4\sigma^{12}$ and C is equal to $4\sigma^{6}$ where σ is the collision diameter (the separation between atoms for which the energy is equal to zero).

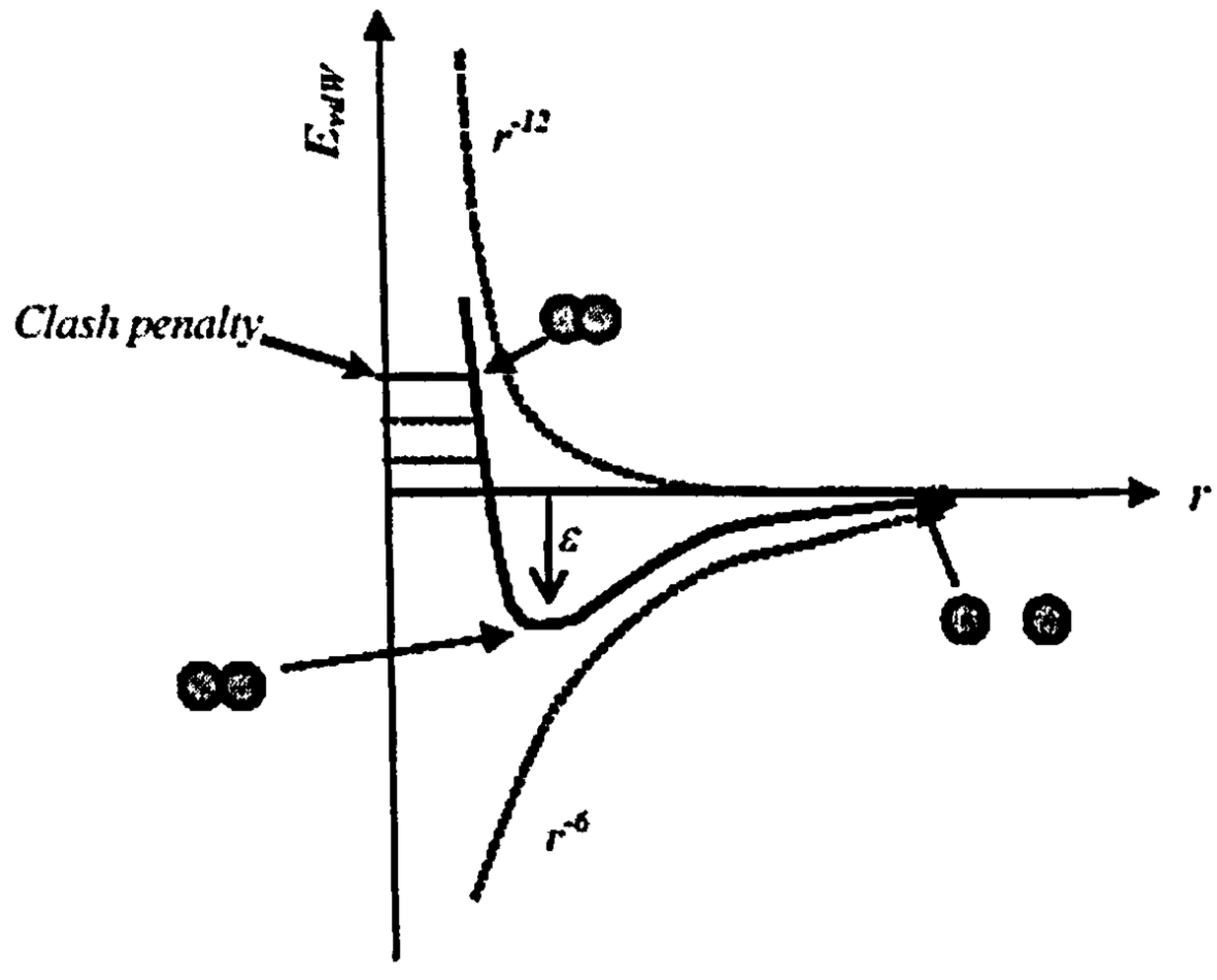


Figure 3-3 The Lennard-Jones potential consisting of a repulsive component (r^{-12}) and an attractive component (r^{-6}). The relative positions of the well depth ε and the clash penalty are also shown.

A diagrammatic representation of the Lennard-Jones potential is shown in Figure 3.3. The attractive component of E_{vdw} varies with r^{-6} and the repulsive component varies with r^{-12}. The clash penalty is a cut-off value that represents the magnitude of the maximum E_{vdw} value allowed between two atoms. The clash penalty can be decreased to dampen the effect of large, unfavourable vdw interactions.

3.3 Molecular modelling

Molecular modelling applies variations of the theoretical principles described above, combined with computational methods, to molecules in order to understand and predict their behaviour in biological systems. The interactions of molecules can be studied to gain insight into the fundamental principles which govern biological processes. Similarly molecular modelling methods can be applied to the rational design of therapeutic drugs that interact with proteins to treat different disease conditions.

Ligand-receptor interactions have traditionally been elucidated through laboratory experiments, which are often time-consuming and expensive. A drive towards molecular simulation has been triggered by the increase in protein structural information in the PDB (Berman *et al.*, 2000) and the decreasing costs of computational hardware. These methods complement laboratory experiments by providing fast and inexpensive investigations to guide further examination through laboratory experiments. An example of this is the application of protein-ligand docking in virtual screening of compound libraries for the identification of hits, a technique which is now widely implemented in the pharmaceutical industry (Jalaie and Shanmugasundaram, 2006).

3.4 Protein-ligand docking

In the very simplest of terms protein-ligand docking attempts to, computationally, predict the correct bound association of a protein and a ligand from their atomic coordinates only. Research over the past two decades has focused on finding the most successful methods to try to achieve this. The fundamental drive for developing such methods is for their use in structure-based drug design (SBDD) for drug discovery. The potential to predict which ligands effectively bind to a protein target is valuable in the process of finding lead compounds that can be developed into effective drugs. This, combined with the mounting costs and lengthy process of taking a drug to market, have caused computational methods of structure based drug design to receive

significant attention (Jorgensen, 2004; Klebe, 2006; Kitchen *et al.*, 2004) as a viable strategy to drug discovery.

Traditionally lead compounds have been discovered serendipitously, by modifying existing drugs or by isolating active ingredients in herbal remedies. However, with the increase in the elucidation of pharmacologically relevant protein structures from experimental (X-ray crystallography and NMR) and computational (homology modelling) methods, a structure-based approach to finding a drug which interacts with a protein to trigger a specific physiological response has become more feasible.

The application of SBDD methods to drug discovery ultimately requires two elements, the 3-D structure of a protein and the location of the appropriate ligand-binding site. As mentioned in the previous paragraph, X-ray crystallography and NMR methods are some of the experimental techniques which are used to obtain structural information, though it is important to be aware of these methods' limitations, in particular the uncertainties introduced during the derivation of an atomic model from the experimentally observed electron density data in X-ray crystallography (Davis, *et al.*, 2003, Acharya and Lloyd, 2005).

Computationally driven homology/comparative modelling can also be used to obtain suitable structures, particularly for structures which are difficult to elucidate experimentally, such as membrane-bound proteins. A notable example of such proteins are GPCRs (G protein-coupled receptors), which are perhaps the most important group of targets for therapeutics (Jacoby, *et al.*, 2006), with 50% of the most recently launched drugs targetting GPCRs (Klabunde and Hessler, 2002). Building a homology model of a protein is possible if there is good sequence homology between the target and another protein whose structure is known experimentally. Homology modelling therefore allows for the discovery of suitable structures of targets, such as GPCRs (Flower, 1999), upon which SBDD methods can be applied.

The binding site within the protein can be inferred by the co-crystallisation of the protein and ligand. Other methods exist which look at sequence or structural similarity between putative sites to known binding sites (Campbell *et al.*, 2003).

Other methods use *de novo* identification and mapping of small-molecules binding sites (Sotriffer and Klebe, 2002, Laurie and Jackson, 2005).

3.4.1 Methodologies of protein-ligand docking

The approach to protein-ligand docking followed by many algorithms divides the process into two components; a search procedure and a scoring function. The aim of the search procedure is to sample the search space efficiently in order to predict the correct pose of a ligand within the protein binding site. With rigid-body docking, six degrees of freedom apply, three for rotation and three for translation. The scoring function is what is used to assess the "quality" of the structures generated by the search procedure. This step is necessary to guide the search towards the correct ligand pose, and is essential in the ranking of generated poses, with good poses being in the top ranks.

The earliest docking algorithms considered both the protein and ligand as rigid bodies, with no consideration of the molecules' flexibility. Nowadays most docking tools allow for flexibility within a ligand. Full receptor flexibility, on the other hand, remains a challenge, due to the immense computational resources that would be needed in order to effectively sample all possible conformations and backbone rearrangements of the target protein. Attempts have been made, however, to include partial protein flexibility. This includes methods such as soft docking (Jiang and Kim, 1991) and partial side-chain flexibility (Leach, 1994, Jones *et al.*, 1995). Current protein-ligand docking algorithms are, in general, considered to be successful at predicting the binding pose of a ligand within a protein binding site (Leach *et al.*, 2006).

Protein-ligand docking methods have been reviewed in a number of publications (Abagyan and Totrov, 2001; Brooijmans and Kuntz, 2003; Campbell *et al.*, 2003; Halperin *et al.*, 2002; McConkey *et al.*, 2002; Shoichet *et al.*, 2002; Taylor *et al.*, 2002, Kitchen *et al.*, 2004). In the following sections a description of scoring functions is presented, followed by a discussion on the search techniques employed in

docking and their application in different algorithms. A review of studies that compare and evaluate different docking methods is also covered, in Section 3.4.4.

3.4.2 Scoring functions

Accurate scoring of ligand conformations is essential for discerning the best solution from a variety of different conformations a ligand adopts during a conformational search, and for differentiating binders from non-binders from a compound library. In addition to pose discrimination, ideally scoring functions should also be capable of ranking sets of ligands according to experimentally determined binding affinities. This assumption is made in virtual screening methods, by using the calculated binding affinities to rank-order a hit list obtained from docking a library of compounds. In practice, this is very difficult to achieve, mainly due to the challenge of accurately calculating the energy components that influence the binding energy.

A good scoring function also needs to be fast, since numerous conformations, generated by the search component, need to be assessed at one time. A huge effort has gone into developing effective scoring functions, which is apparent in the numerous reviews (Tame, 1999; Gohlke and Klebe, 2001; Gohlke and Klebe, 2002; Jansen and Martin, 2004) and performance comparative studies (Wang *et al.*, 2003; Ferrara *et al.*, 2004; Wang *et al.*, 2004) which have been published in the area. Despite this drive, scoring functions continue to have their limitations (Leach *et al.*, 2006), and to-date no scoring function has been developed which succeeds at discerning correct poses in all docking experiments. Attempts to understand scoring functions have included the employment of decoy structures to validate the capabilities of scoring functions (Wang *et al.*, 2003; Perola *et al.*, 2004) and to highlight their limitations (Graves *et al.*, 2005). Some attention has also recently focused on developing methods that tailor scoring functions towards a specific need (Catana and Stouten, 2000).

Scoring functions in general can be categorised into three main groups; (1) scoring functions derived from first principles, (2) empirical scoring functions and (3) knowledge-based scoring functions.

3.4.2.1 Force field-based scoring functions

Force field-based scoring functions use first principles to quickly compute the scores of docking solutions, are transferable and also use terms which have a physical basis. Force field scoring functions tend to only measure potential energy therefore some force fields contain additional terms to account for desolvation and entropy using different models. The non-bonded interaction energy terms calculated in force fields are the vdw and electrostatics contributions, usually performed using a Lennard-Jones 6-12 potential and Coulombic function respectively. The Lennard-Jones potential is sometimes softened, to allow for "soft" docking and as in GOLD, where a 4-8 potential is used. Hydrogen bonding terms are accounted for using geometric dependent terms as shown with Q-fit (Jackson, 2002) and GOLD (Jones *et al.*, 1997). Continuum solvent models, using a fixed or scaled dielectric constant, are the common approach to dealing with the effects of solvent.

3.4.2.2 Empirical scoring functions

Empirical scoring functions use multivariate regression methods to fit coefficients of physical contributions to the binding free energy in order to reproduce measured binding affinities for a training set of known 3D protein-ligand complexes (Horton and Lewis, 1992; Bohm, 1994). LUDI (Bohm, 1994), was one of the earliest empirical scoring functions developed, and includes terms for hydrophobic contact, polar interactions and entropic fixation costs for loss of torsional, translational and rotational degrees of freedom upon binding. The FlexX (Rarey, 1996) scoring function uses a modified version of the LUDI function, and estimates free energy contributions from the number of rotatable bonds in the ligand, hydrogen bonds, ion-pair interactions, hydrophobic and π-stacking interactions of aromatic groups and

lipophilic interactions. Empirical scoring functions are fast and are reasonably good at predicting binding free energies. The disadvantage of these methods is that they are trained on protein-ligand complexes with good binding energies, and therefore they do not penalise steric clashes or same-charge interactions appropriately.

3.4.2.3 Knowledge-based scoring functions

Knowledge-based scoring functions are based on a classical statistical physics concept, where observed distributions of geometries are used to deduce the potential that give rise to the observed distribution. The function uses protein-ligand atom pair potentials derived from crystallographic complexes and assumes that these are represented correctly. A limitation of this type of function is the unavailability of sufficient data to allow computation of density distributions for different parameters. Also, this approach tends to treat all parameters independently when it may not be correct to do so. For example, a hydrogen bond with a large distance between donor and acceptor will not contribute towards binding, regardless of the directionality of the bond. The availability of large quantities of data can circumvent this issue as it would allow the consideration of joint distributions of parameters. One of the first studies which used this method was for the binding affinity prediction of HIV-1 protease complexes (Verkhivker *et al.*, 1995). Current popular knowledge-based scoring functions are Drugscore (Gohlke *et al.*, 2000) and DFIRE (Zhang *et al.*, 2005).

3.4.2.4 Consensus scoring

Consensus scoring has become increasingly popular for enhancing the performance of a scoring function (Charifson *et al.*, 1999; Wang and Wang, 2001; Clark *et al.*, 2002; Yang *et al.*, 2005). This method works by using individual scoring functions to synergistically score generated poses; for example one scoring function can be used to find poses, and a second scoring function can be used to refine the placement and correctly score the protein-ligand complex.

3.4.3 Docking search procedures

Numerous procedures have been employed to perform conformational and configurational searches. The most widely used are simulated annealing, matching, molecular dynamics, evolutionary algorithms (which include GAs), tabu searches, incremental construction and systematic searches. These are discussed below.

3.4.3.1 Matching algorithms

Matching algorithms are designed to align, or match, structural features of a ligand onto a protein binding site. Surflex (Jain, 2003) generates an idealised binding site, a protomol, from a protein's binding site, to which ligands are optimally aligned, and which consists of molecular information which represent the most favourable interactions to the protein binding site. During a docking run, the ligand is fragmented, and each fragment is aligned to the protomol to maximise the molecular similarity to the protomol. All the fragments are then scored, and the molecule is re-assembled, either by incremental construction (described below) or by a "whole molecule" algorithm (which is significantly faster).

LigandFit (Venkatachalam *et al.*, 2003) is a shape-directed docking procedure which matches a ligand to the active site of a protein. A site detection algorithm first of all detects the position of the active site, or, where possible, an experimentally-validated active site is used to define the shape of the active site. A Monte Carlo search procedure (see below) is used to search the conformational space of the ligand. Each ligand conformation is evaluated against the active site and, if it passes a certain threshold, is docked in the active site via a shape alignment protocol. Protein-ligand interaction energy calculations further refine these poses. 14 of the 19 test cases produced structures very close to the experimentally-determined ones. The method is also very fast and the authors recommend its use in high-throughput virtual screening studies.

The earliest protein-ligand docking algorithms assumed rigidity of both molecules' bonds. DOCK, one of the earliest docking algorithms (Kuntz *et al.*, 1982), used a matching technique modelling rigid-body molecules. One of the more recent versions of the program (DOCK 4.0) incorporates incremental construction and random conformational searches to allow flexible ligand docking (Ewing *et al.*, 2001). Incremental construction is described in a later section. The matching orientation procedure in DOCK 4.0, which is user-defined, can be either automatic or manual. For a manual matching operation, the user must specify geometric parameters such as the distance tolerance, and these are used to build orientations that match these parameters. In automatic matching the program performs nested cycles of matching, until a user-defined number of valid orientations has been reached. This technique is carried out on ligand fragments which are subsequently added to an anchor fragment and pruned. The purpose of pruning is to cut down on the degree of the systematic conformational search which increases exponentially with the number of fragments. This is done by analysing binding positions according to score and only keeping the best scoring and most diverse positions for the next stages of the algorithms. DOCK 4.0 still allows the option of rigid-body docking, where multiple conformations of the ligand are docked independently. DOCK 4.0, unlike previous versions, allows the user to specify the number of conformations (N) a rotatable bond of a particular flexible ligand can undertake. A ligand with 4 rotatable bonds will therefore have N^4 possible conformations to be docked. This gives the user control over the degree of conformational search to be explored.

3.4.3.2 Simulated annealing

The principles of simulated annealing originate from the annealing of physical objects in the manufacturing industry where a molten substance's temperature is slowly reduced until a crystal is formed. Mimicking this process in docking, the temperature of a system is reduced until a stable docked structure is achieved. The temperature (which corresponds to the degree of random motion) change is achieved by changing a control parameter, and the free energy is represented by a scoring function. To get

to the global minimum, careful temperature control is essential and large increases in temperature must be applied periodically, so as to overcome high-energy barriers, and to avoid being stuck in a local minimum. The Metropolis Monte Carlo simulated annealing algorithm in particular has been widely used. MCDOCK (Liu and Wang, 1998) is one of the more popular MC (Monte Carlo) -simulated annealing algorithms, which performs geometry-based docking and energy-based docking, followed by a final energy minimisation on the docked structure. The scoring function used in the second step utilises Lennard-Jones and electrostatic non-bonded interaction terms. The simulated annealing step of this algorithm first samples several geometries at a high temperature. The lowest energetic structure is then chosen, which is then used as a starting state for more structures, at lower temperatures. During these steps, several structures with low energies are saved, and the lowest is once again minimised. This version of MCDOCK was tested on 19 ligands, and the rmsd values obtained were between 0.25 Å and 1.84 Å for all 19 cases.

3.4.3.3 Tabu search

A tabu search, a heuristic search procedure, was first applied to the docking problem within the PRO_LEADS software (Westhead *et al.*, 1997). It involves the generation and maintenance of a tabu list, which contains a number of previously visited solutions, and which are considered to be "tabu", i.e. the algorithm is restricted from revisiting them. The algorithm begins with the generation of a random solution. The variables of the solution are then randomly adjusted using user-defined moves, and these moves are scored using an energy function. These moves are then ranked by the value of their energy functions, and each move is individually examined. If a solution resulting from a certain move is not sufficiently different (which is determined by an rms threshold value between two solutions) from the solutions in the tabu list, then it is considered tabu and is discarded. The best non-tabu move is selected (if found) and added to the tabu list. A move is always accepted, regardless of its tabu status, if it has energy lower than any solution that has been generated so far. If no non-tabu moves have been found, and if none of the moves have generated solutions of the lowest energies, then the algorithm terminates. This particular tabu search also has an

extra method which restarts the whole procedure using a new initial solution, if the best solution has not changed for a number of iterations. This helps the search *escape* any local minima. PRO_LEADS was validated against a test set of 50 complexes (Baxter *et al.*, 1999). Using a docking protocol that compromised between accuracy and average docking time per complex, 79% of the ligands docked with rms values of less than 2.0 Å. Using the same docking protocol, 10 000 ligands from the ChemBridge Prime database were selected for virtual screening experiments, to be docked into three receptor molecules (thrombin, factor Xa and ER), along with a number of known ligands. The authors reported a good separation between the docking energies of the two subsets.

3.4.3.4 Incremental Construction

The incremental construction technique is different from the other search techniques so far described in that the ligand is placed in the active site incrementally, i.e. the ligand is divided into fragments which are docked independently and then fused. FlexX (Rarey *et al.*, 1996) is one of the more widely used incremental algorithms, which considers the receptor as a rigid body, and allows for ligand flexibility. The scoring function used to assess the solutions is one similar to that developed by Böhm (1994).

The FlexX algorithm consists of three phases. The first phase consists of the selection of the base fragment. This is a fragment that is connected to the receptor. What needs to be considered in the selection of the fragment is that a larger fragment with more interaction groups increases the probability of finding the correct binding mode, but that this also increases the number of possible conformations of the base fragment. The authors obtained the best results if the fragments are small. After the base fragment has been selected the rest of the ligand is split into fragments.

The second phase of FlexX is the base placement algorithm, which uses a pose clustering technique to generate a set of placements for the base fragment in the active site, and Böhm's scoring function is used to assess the placements. The final phase of the algorithm involves the incremental construction of the ligand (complex

construction). This is approached as a tree search problem, where a node represents a placement of the ligand. The binding energies of the growing ligands are assessed at each stage so that the fragments are only placed in the most energetically favourable poses. Extended placements are then optimised where necessary, ranked (based on the scoring function mentioned earlier) and clustered (to remove similar placements). FlexX was tested on 19 complexes, and docked all 19 within a 0.5 Å to 1.2 Å rms deviation from the experimentally determined structures.

FlexX forms the basis of another docking algorithm, FLEXE (Claußen *et al.*, 2001), which considers protein side chain variations, point mutations and loop movements (to a certain degree). Each variation is represented by a single structure, and all the variable structures together are known as an ensemble (Knegtel *et al.*, 1997). The entire ensemble representation is based on a "united protein" description which is created by the superimposition of the different structures in the ensemble. Similar parts between all the structures in the ensemble are merged whereas the dissimilar parts are treated as variations of the united structure. 67% of the ten ensembles FLEXE was validated on had solutions of RMSD less than 2.0 Å in the top ten predictions. FLEXE is now known as FlexX-Ensemble.

Hindle *et al.*, (2002) introduced FlexX-Pharm, which employs the original FlexX flexible docking tool, but allows for pharmacophore-type constraints to be considered during the incremental construction stage. In many docking studies, some knowledge may already exist concerning particular features adopted by the protein target site when docked to a ligand. FlexX-Pharm considers two types of constraints, interaction and spatial constraints, when building a particular ligand. Only if the specified interaction between a particular group in the active site and a ligand fragment in a certain position is present would a partial solution be kept. A spatial constraint constricts a ligand to a particular position in the active site, and it is defined by an element type and a sphere in which the element must remain. FlexX-Pharm produced good results, especially with complexes where FlexX was not successful.

Q-fit (Jackson, 2002) is another tool that uses incremental construction, combined with probabilistic sampling, to dock fragments into a binding site. Q-fit has been utilised during the research for this thesis and is therefore described in more detail in Chapter 4.

3.4.3.5 Genetic algorithms

The basics of a genetic algorithm are described in section 2.2.1. In this section genetic algorithms will be discussed in the context of protein-ligand docking. One of the earliest GAs developed for flexible ligand docking is the extension to the DOCK programs developed by Oshiro et al. (1995). This program explores the conformational and orientational space of a ligand within a receptor, using a GA-based technique. A chromosome (which represents one potential solution, or pose, of the ligand), encodes both the orientation and the conformation of the ligand. The conformation is represented by values of all the torsion angles about rotatable bonds of the ligand. The representation of the ligand orientation within a chromosome is dependent on the GA approach being used. The program allows for two different approaches: a sphere-based GA method and an explicit-orientation-based method. The former orients the ligand by matching the ligand atoms with spheres that represent the active site. In this method, the orientation of a ligand is represented by pairs of integers, where each pair is a matching between a sphere and an atom number, and is encoded in a gene within a chromosome. The explicit-orientation-based method consists of two stages: the first stage determines restrictions on the ligand orientation space so that only the active site region is explored. In the second stage, the GA is used to find the lowest-energy conformation. In this approach, the orientation representation in the chromosome is a translation vector and Euler angles, describing the ligand's position in the search space, and the conformation is represented by torsion angles about the ligand's rotatable bonds. 10%-20% of a population in a particular generation is discarded, and two-point crossover is carried out on 60%-70% of the remaining population. Each bit in a chromosome has a certain probability of being mutated (0.0065 was the probability used by the authors). Based on the AMBER potential function, the ligand-enzyme interaction energy is calculated to score a particular solution. Both GA-based methods succeeded in finding low-energy solutions that are close to the experimental structures.

Autodock (Morris *et al.*, 1998) is an automated docking algorithm whose searching procedure was modified from a simulated annealing search to a hybrid of a genetic algorithm and a local search. The reason for this extension is that the success of docking with simulated annealing employed by Autodock was limited to ligands with eight rotatable bonds or less. The hybrid consisted of a traditional genetic algorithm combined with a local search method (LS) for energy minimisation. The LS procedure is adaptive, i.e. energies from previously generated poses during the LS influence the step size of the steps that follow.

In this GA the chromosome consists of real number genes; three Cartesian coordinates for translating the ligand along the three axes, four variables (quaternions) that define ligand orientation, and a real number representing each torsion angle in the ligand. Five stages are involved in forming a single generation: mapping and fitness evaluation, selection, crossover, mutation and elitist selection. Mapping converts the individual's genotype to its phenotype (i.e. from the genes that make up the chromosome to the coordinates of the ligand and its calculated energy function). The energy function (which is the sum of intermolecular interaction energy between ligand and protein and the intramolecular interaction energy within atoms of the ligand) signifies the fitness of a particular ligand pose. The selection of the individuals allowed to reproduce is performed through proportionate selection. Individuals with above average fitness will generate proportionately more offspring. Selection is followed by two-point crossover and mutation. Also, an elitist strategy is used to pass the top individuals to the next generation. The algorithm is implemented over several generations until a termination criterion is met. In Autodock, this occurs when a maximum number of generations or a maximum number of energy evaluations is reached.

At every generation, a user-defined fraction of the population can undergo a local search (LS). Here, LS searches the genotypic space around a particular individual and any improved adaptations observed are encoded in the genotype, which is inherited in the offspring. This concept of encoding, in which changes acquired in the phenotype are passed to the next generation is known as Lamarckian, after Jean Batiste de Lamarck's discredited theory that phenotypic characteristics acquired during an

individual's lifetime are inherited. Hence the GA in Autodock is referred to as the Lamarckian Genetic Algorithm (LGA).

The different search procedures in Autodock were tested on seven protein-ligand complexes, where the LGA search was found to be the most successful and efficient in predicting the docked structures (mean rmsd from experimental structures = 0.88 Å).

While Autodock considers the target structure as a rigid body, GOLD (Jones *et al.*, 1997) allows for some flexibility within the active site of the target. The GOLD algorithm is based on the GA described by Jones *et al.* (1995). The chromosomes in GOLD are represented by bit strings. Two binary strings, one for the protein, one for the ligand, are used to represent the torsional angles. Two integer strings represent hydrogen bond mappings between the protein and the ligand. A GA run begins with the generation of an initial population randomly. The fitnesses of the individuals are calculated based on the scoring energy function. An operator (such as crossover or mutation) is chosen using roulette wheel selection, which is also used to select the parents whose genotypes will be manipulated by the operators. The resulting children then replace the least fit members of the population, and the iteration process continues. The termination procedure is employed when 100 000 operators have been applied.

As well as modelling hydrogen bond interactions, GOLD also incorporates hydrogen bonding energy in the scoring function. The authors have noted the importance of hydrogen bonds observed from crystal structures which show that ligands appear to interact at a number of key hydrogen bonding sites to conform to the shape of the binding site. A recent version of GOLD accounts for water mediation and displacement by switching water molecules on and off and allowing their rotation around their three principal axes (Verdonk, *et al.*, 2005).

DARWIN (Taylor *et al.*, 2000), another GA-docking tool, is interfaced with the CHARMM force field program (Brooks, 1983) to score flexible ligand placements within a protein. Parallelisation of CPUs is employed to run several CHARMM

programs instantaneously to allow multiple energy evaluations and thus speed up the process of a DARWIN run.

The chromosome in a DARWIN GA contains the binary encoded variables of a ligand (a solution) that represent the starting orientation and position of the ligand, as well as the torsional bonds. The population size and the number of generations are user-defined. An elitist strategy (or the "survival rate"- termed by the authors) passes the fittest chromosomes to the next generation, whereas the least fit are deleted (defined by the term "death rate"). Mutation and crossover operators are carried out on the parent chromosomes to fill in the rest of the population (the population size remains the same throughout a run). DARWIN's search strategy performed well in finding solutions with fitnesses at least as good as the experimental structure.

GAs have also been applied to the docking of fragmented ligands. The SEED-FFLD algorithm (Cecchini *et al.*, 2003) docks ligands in two phases. First the SEED program (Majeux *et al.*, 1999) docks the ligand fragments into the binding site of the receptor. The FFLD program (a GA) then docks the whole ligand into the binding site, using the docked fragments from the SEED to determine its placement. The FFLD program therefore only explores the conformational space of the ligand. FFLD, as with Autodock, uses a hybrid technique that consists of a global and a local search. The stages which constitute a single generation in this program are: evolution, mapping, fitness evaluation, local search and similarity testing. The first stage comprises of using the genetic operators (one-point crossover and mutation) to the population to generate a new population. The binding energy of the new population is evaluated. A local search is performed on the top 10% of the individuals in the new population to refine the fitness. Finally the parent population is compared to the new population and parent chromosomes are replaced by new chromosomes, taking energy values and structural similarity between the chromosomes into consideration. By reducing structural similarities between the chromosomes the diversity of the population is retained and convergence to local minima is avoided. The termination criteria used by FFLD are determined by the maximum number of generations reached or number of energy evaluations performed.

3.4.3.6 Systematic methods

Systematic docking approaches attempt to carry out an exhaustive search of conformational, positional, and orientational spaces of a ligand relative to a protein binding site. the program Glide (Friesner *et al,* 2004) approximates a complete systematic search of a ligand's degrees of freedom by first implementing a rough positioning and scoring phase for narrowing down the search space. This generates ligand poses that are then minimised in the field of the recptor using a standard molecular mechanics energy function in conjunction with a distance-dependent dielectric model. Finally three to six lowest energy poses undergo Monte Carlo simulations to examine nearby torsional minima. This step is necessary in some cases to correctly orient peripheral groups and alter internal torsional angles. A scoring function is finally used to select the correct pose from the minimised poses. The scoring function implemented uses a modified and expanded version of Chemscore scoring function (Eldridge *et al*, 1997), which the authors term as GlideScore to predict binding affinity and for rank-rodering ligands in database screens. To select the correct pose, a composite scoring function is used, that combines GlideScore, the ligand-receptor molecular mechanics interaction energy and the ligand strain energy. This scoring function is termed the E_{model}, and which was found to be superior to using GlideScore or the molecular mechanics energy alone.

Glide was tested on 282 co-crystallised protein-ligand complexes derived from the PDB. The results showed that the program obtained geometries that were less than 1 Å in nearly half of the test cases, and greater than 2 Å in 94 of them. These results were compared to other published results, and it was found that Glide is nearly twice as accurate as GOLD and more than twice as accurate as FlexX for ligands with up to 20 rotatable bonds.

3.4.4 A review of comparative studies of docking methods

Numerous comparison studies have been published to understand and evaluate the performance of different docking methods relative to each other. It is not possible to provide a review of all, so this section will provide a review of the more significant papers.

Onodera et al., (2007) compared three docking tools, AutoDock, GOLD and DOCK by testing these on 116 target proteins. DOCK was shown to have the best screening performance in the enrichment rates compared with the others, whereas GOLD was the best in docking pose prediction. Another study, comparing the performance of Glide, GOLD and DOCK for virtual screening was carried out (Zhou et al., 2007). The test case targets in this particular study are deemed as being pharmaceutically interesting, and are tested with active compounds. Glide XP was shown to achieve better enrichment rates compared with the other two methods, while GOLD outperforms Dock.

One of the more comprehensive studies attempted to compare eight of the most widely used docking tools, in terms of their ability to both, predict poses of x-ray crystal structures, and to discriminate known inhibitors from randomly selected, drug-like molecules (Kellenberger et al., (2004)). The authors found that the same three algorithms (Glide, GOLD and Surflex) were successful in both properties. The strengths and weaknesses of each of the docking tools were also noted, based on physicochemical properties of the ligands and protein binding sites. A similar study comparing the performance of five docking programs (FlexX, DOCK, GOLD, LigandFit and Glide) against 14 protein families (and comprising of 69 targets) also found GOLD and Glide to be the most reliable (Kontoyanni et al., 2004). This paper also attempted to relate the results obtained with each docking tool with the nature of the binding site. For example GOLD was found to perform well with mildly or mostly hydrophilic targets, whereas Glide was not as discriminatory to the nature of the active site.

A more recent study conducted by Warren et al., (2006), attempted to assess the current state of docking algorithms, by evaluating 10 docking tools and 37 different scoring functions. The datasets consisted of eight protein targets tht are of seven protein types, and ligands that were very similar to pharmaceutical companies' compound libraries. The authors found that the docking algorithms were successful at generating correct poses, although scoring functions were less successful at differentiating crystallographic poses from the rest of the generated poses. It was also found that docking programs can identify active compounds from a set of decoys, though this was not possible across all protein targets. A study of the docking programs' and scoring functions' ability to predict compound affinity showed that none were able to do so successfully.

Effectively comparing docking algorithms can be problematic, in particular with respect to standardising the methods and the analysis of the docking programs' results to warrant a fair comparison. Issues include the use of rmsds as a reliable measure of success, the lack of consideration of crystal packing interactions and not always ensuring that all search problems are given equal levels of complexity (Cole et al., 2005). Finding an algorithm's optimum set of parameters suited to a particular problem can also bias comparative studies; attempts have been made recently at understanding docking parameters' effects on performance (Andersson et al., 2007).

3.5 Aims of this work

Though much progress has been achieved in the protein-ligand docking field, limitations attributed to the accuracy of scoring functions continue to hinder the development of accurate and robust algorithms.

As mentioned in section 3.2.3.1, scoring functions tend to weight individual energy terms before combining them to give a total energy, in a process that can be described as a weighted sum approach and which was discussed in section 2.3.1. Section 2.3.2

introduces Pareto concepts for multiobjective optimisation as an alternative way of dealing with problems comprising of different components, or objectives. In this work multiobjective optimisation is applied to the scoring function element of protein-ligand docking. The thesis will focus on understanding scoring functions further by applying multiobjective optimisation to a force field-based scoring function. Rather than using the total interaction energy to assess the quality of poses, a multiobjective approach, using individual components of the scoring function, will be used. In this way the influence of individual energy terms in docking a ligand correctly into a binding site can be examined energy. The role of individual energy terms when docking a particular test case could be compared to results obtained by a single objective optimisation algorithm (most current docking tools fall under this category). For cases when a single objective optimisation algorithm fails to find the correct solution, the total energy of the best solution can be broken down into its individual energy terms, and the relative contribution of each terms to the total energy can be compared to the energy terms of correct solutions obtained from a multiobjective algorithm. From such an analysis, it may be possible to realise whether the relative contributions of individual energy terms are important in finding correct solutions. These experiments would also indicate which of the terms, if any, is most important in docking a ligand with a correct pose.

As far as we are aware, no studies have been performed to study, in a multiobjective manner, the balance of energies which constitute a scoring function. An attempt to understand the importance of individual interaction energies in relation to others has been demonstrated by Brenk *et al.*, (2006) who have designed very simple binding sites dominated by only a few energy terms, and where other docking approximations do not apply. The binding site used by the authors allowed for the study of the balance between electrostatic energy and desolvation energy, which was performed by the retrospective docking of known binders and non-binders and for which the scoring function being implemented was found to predict accurately. The binding site was also used for prospective docking of a large compound database. The binding affinities of the top scoring and lowest scoring ligands were calculated and the crystal structures of the binders were elucidated, which helped in giving an insight into the accuracy and weakness of the scoring function. Similar studies by the same group

have been carried out on cavities that examine ligand binding in hydrophobic and slightly polar environments (Graves *et al.*, 2005, Wei *et al.*, 2002). In contrast this thesis will attempt to understand the roles of individual energy components algorithmically, within the realms of the scoring function.

The aim of this thesis is therefore to gain a greater understanding of the influence of individual terms that comprise scoring functions, and their effect on the performance of a docking algorithm. More specifically, this will be achieved through the following objectives:

- To develop a docking algorithm that optimises the individual energy terms in a scoring function independently, with an initial focus on electrostatic, hydrogen and vdw energy terms.
- To compare docking results obtained from multiobjective optimisation to single objective optimisation in order to see whether docking failures of single objective optimisation algorithms can be attributed to the incorrect optimisation of individual energy terms.
- To gain an understanding on whether individual energy terms have differing influences on different complexes.
- To test the algorithm on a dataset comprising of a single protein in complex with different ligands.
- To test the algorithm on a large dataset of different protein-ligand complexes.
- To extend the algorithm so that it incorporates a third objective, thus performing multiobjective optimisation on three objectives.

In the following chapter (Chapter 4), the development of a single-objective genetic algorithm that performs rigid-body protein-ligand docking is described, and the results obtained when the algorithm is tested on one dataset are presented. Chapter 5 describes the adaptation of the algorithm from single-objective into a multiobjective optimisation docking algorithm. Chapter 6 describes results obtained when the algorithm is tested on two datasets. Chapter 7 describes enhancements to the algorithm, followed by results comparing the enhanced version with the original version of the algorithm. Chapter 8 shows results from testing the algorithm on various and larger datasets. Chapter 9 describes the adaptation of the algorithm to

incorporate solvation energy as a third objective in optimisation, and finally Chapter 10 is a discussion of the major observations and future directions of the project.

4 Docking using single-objective optimisation

As was discussed in Chapter 3, the aim of this thesis is to apply the process of multiobjective optimisation to protein-ligand docking and, in particular, to a scoring function. The scoring function selected for this is Goodford's force field based GRID scoring function (Goodford, 1985) and which is discussed in section 4.4.

Since an algorithm that employs multiobjective optimisation follows the standard form of a GA, a good starting point in the process of producing a multiobjective optimisation docking tool is to first develop a single objective, standard GA (or SGA). An SGA has several of the elements that are necessary for multiobjective optimisation, such as crossover, selection, and mutation. By first implementing an SGA, these operators can be tested and their parameters optimised, before modifying the algorithm into a multiobjective optimisation program. The type of docking implemented here is rigid-body docking.

4.1 The chromosome and its genes

As was discussed in Chapter 2, the population in a GA is made up of a collection of chromosomes, each of which represents a solution to the problem which is being optimised. In the case of protein-ligand docking, the chromosome represents the pose of the ligand in the search space- which consists of a section of the protein binding site. For the SGA that was developed here, the ligand pose is coded into the chromosome using real value representation- the magnitudes of the translation and rotation of the ligand, with respect to a reference structure, are represented as floating point numbers in the chromosome's genes. Since the algorithm performs rigid-body docking, rotatable bonds have not been considered and these are not encoded in the chromosomes' genes. Within the algorithm a chromosome therefore consists of an array of floating point numbers, where each floating point is position-specific to the gene it represents. Figure 4.1 contains a schematic of the chromosome. The first three floating point numbers of the array, (trX, trY and trZ) represent the translation of the

ligand, in Angstroms along the x, y and z axes respectively. Rotation of the ligand along the x, y and z axes is represented by rotX, rotY and rotZ radians respectively, and which occupy the fourth, fifth and sixth positions of the chromosome. These genes represent the six degrees of freedom which are necessary for generating different poses of a ligand in order to implement rigid-body docking. They also represent *changes* that are carried out on a ligand's reference pose.

4.2 Mapping the genes to the ligand

The coordinates of the ligand, as extracted from a pdb file, represent its position relative to the protein it is bound to, and has usually been inferred experimentally. The experimentally determined ligand will be referred to in this thesis as the ligand's crystal structure, because all test cases applied here have been inferred by x-ray crystallography. The pose of the crystal structure also forms the comparison upon which the quality of a solution is assessed, by comparing the pose of a solution to that of the ligand crystal structure's pose. It is therefore important to remove any bias towards the ligand crystal structure, and to ensure that the use of this information is kept to a minimum during a run of the algorithm. For this reason the genes encoded by a chromosome are applied to a reference pose that has been modified from the ligand crystal structure rather than to the ligand crystal structure pose. This is done by first translating the ligand so that it is at the origin of the GRID box (see section 4.4) and then rotating it along the x, y and z axes by three randomly generated numbers. A pose is generated by the program by applying the rotations and translations stored in the chromosome to the reference ligand pose.

4.2.1 Rotation

The rotation procedure rotates the ligand along the x, y and z axes. The ligand reference pose is first translated from the GRID box origin to the global origin. It is essential to rotate at the global origin otherwise the ligand will be both rotated and translated erroneously. This is followed by the construction of rotation matrices using

the three rotation genes, which are hard-coded into the program, and are shown below.

$$\begin{bmatrix} 1 & 0 & 0 \\ 0 & \cos rotX & \sin rotX \\ 0 & -\sin rotX & \cos rotX \end{bmatrix}$$

$$\begin{bmatrix} \cos rotY & 0 & -\sin rotY \\ 0 & 1 & 0 \\ \sin rotY & 0 & \cos rotY \end{bmatrix}$$

$$\begin{bmatrix} \cos rotZ & \sin rotZ & 0 \\ -\sin rotZ & \cos rotZ & 0 \\ 0 & 0 & 1 \end{bmatrix}$$

By concatenating these matrices together (performed by multiplying them together), a unified matrix is produced, which, when multiplied by the three coordinates of an atom of a ligand, will rotate the atom by rotX, rotY and rotZ radians. This step simplifies and speeds the rotation process. Once all of the ligand's atom coordinates have been multiplied by the unified matrix, the newly-rotated ligand is translated back to the GRID origin and this is followed by the translation procedure.

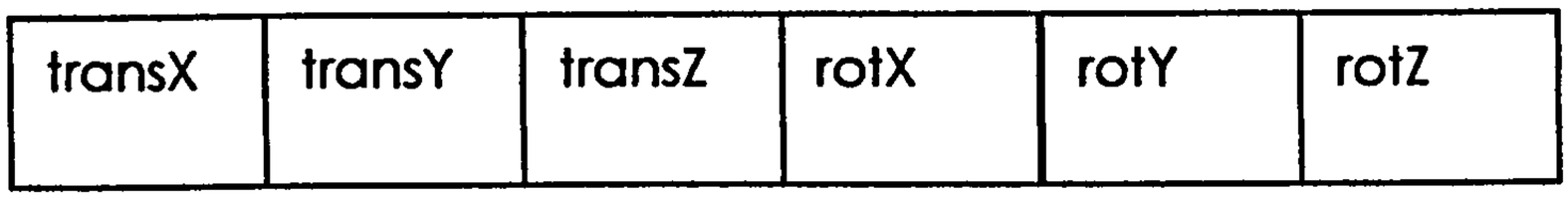


Figure 4-1: The structure of the GA chromosome. transX, transY and transZ are the magnitudes of the translation vectors of the ligand. rotX, rotY and rotZ represent the magnitudes of the rotations about the three axes.

4.2.2 Translation

The translation of the ligand allows for rigid movements along the x, y and z axes. The translation genes (trX, trY and trZ) move the ligand from its position at the GRID box origin by trX, trY and trZ Å along their respective axes. The algorithm performs this operation by simply adding the value of each gene to its corresponding coordinate of every atom in the ligand.

In this way a chromosome's genes are mapped onto a particular pose of a ligand. After this operation, the scoring function is used to assess the quality of the pose generated by a given chromosome, and this is described in section 4.4.

4.3 The Genetic Operators

4.3.1 Selection

The roulette wheel selection method is implemented for the selection of pairs of chromosomes from the parent population to undergo crossover. As was described in section 2.4.2, it is necessary to have a bias during the process of selection towards chromosomes with good fitness values, to ensure that more advantageous genes are passed on to the next generation. This is possible through roulette wheel selection by

creating a hypothetical roulette wheel that is divided into segments, where each segment represents a chromosome, and whose size is proportional to its fitness (see Figure 4.2). One of the main issues with this method is the problem with scaling (Hancock, 1994). Very fit individuals may take over the majority of the roulette wheel and this could flood the population with these individuals, leading to premature convergence. To avoid this situation, the linear scaling approach (Goldberg, 1989) was adopted to scale the individual fitnesses of all the chromosomes in the population. Linear scaling adjusts the fitness values so that the relationship between the raw fitness values (f) and the scaled fitness values (f') is linear, as shown in equation (1).

$$f' = af + b \qquad \text{Equation 4.1}$$

To maintain the scaling, the average scaled fitness (f'_{avg}) must equal the average raw fitness (f_{avg}). Also, in later generations of a GA run, the situation may arise where the average fitness may be close to the best fitness. This would result in average members contributing the same number of offspring as the best members and the optimisation process will stagnate. By introducing the following relationship into the scaling procedure,

$$f'_{max} = C_{mult} \cdot f_{avg} \qquad \text{Equation 4.2}$$

the number of offspring the fittest individual (f'_{max}) can contribute is controlled by C_{mult}, which is the number of expected copies desired for the best population member.

As a result of scaling, in later generations the situation may arise where, due to a few bad strings and a relatively close best fitness and population average fitness, the worst members are scaled to negative values. In these cases the worst member is always mapped to a scaled fitness of 0 ($f_{min} = 0$).

The following equations were derived and implemented in the scaling function, prior to roulette wheel selection.

4.3.1.1 Scaling

The fitness function used by the GA is the interaction energy between a given ligand pose and the protein. The more negative (i.e. the smaller) a value is the more desirable is the pose's conformation and the fitter is its chromosome. These negative energy values must first of all be converted to positive values, so that the fittest members have the largest, most positive values and the less fit ones have the lowest fitness values. To do this, the signs of the energy values are simply inverted, so that all negative energy values are now positive and vice versa. This is followed by subtracting the worst inverted value (which may be negative) from all the other values in the population, so that all chromosomes have positive values. These methods result in a population where the fittest individual has the highest value, and the worst individual a value of 0 ($f_{min} = 0$).

As mentioned earlier, both f_{avg} and f_{avg} need to remain equal. Therefore if,

$$y = a + bx \qquad \text{Equation 4.3}$$

describes the linear relationship between the raw and scaled values, then

$$f'_{avg} = a + b \cdot f_{avg}$$

Because f_{avg} is equal to f'_{avg}:

$$f_{avg} = a + b.f_{avg}$$

$$a = f_{avg}(1-b) \qquad \text{Equation 4.4}$$

$$f_{avg} = a/(1-b) \qquad \text{Equation 4.5}$$

From 4.2

$$C_{mult} \cdot f_{avg} = a + b.f_{max}$$

$$a + b \,.f_{max} = C_{mult} \,.f_{avg}$$

From 4.5

$$a = f_{avg} \,.\, (1\text{-}b)$$

$$f_{avg} \,.\, (1\text{-}b) + b \,.f_{max} = C_{mult} \,.f_{avg}$$

$$b(f_{max}\text{-}f_{avg}) = (C_{mult}\text{-}1) \,.f_{avg}$$

$$b = \left(f_{avg} \,.\, (C_{mult}\text{-}1)\right) \Big/ (f_{max} - f_{avg}) \qquad \text{Equation 4.6}$$

From 4.4 and 4.6

$$a = f_{avg} \,.\, \left(1 - \left((f_{avg} \,.\, (C_{mult} - 1)\right) \Big/ (f_{max} - f_{avg})\right)$$

$$a = f_{avg} \,.\, \left(f_{max} - f_{avg} - f_{avg} \,.\, (C_{mult} - 1)\right) \Big/ (f_{max} - f_{avg})$$

$$a = f_{avg}(f_{max} - C_{mult} \,.f_{avg}) \Big/ (f_{max}\text{-}f_{avg}) \qquad \text{Equation 4.7}$$

a and b are calculated in the above manner when the f_{min} is not negative. If f_{min} is negative then a and b are calculated as follows.

$$b = 1 - a \Big/ f_{avg} \qquad \text{Equation 4.8}$$

f_{min} is scaled down to 0, therefore y = 0.

$$0 = a + b.f_{min}$$

$$a = -b.f_{min}$$

Substituting from 4.8

$$\mathrm{a} = -f_{min}(1 - (a \,/\, f_{avg}))$$

$$a = (-f_{min} \cdot f_{avg}) \,/\, (f_{avg} - f_{min})$$

To calculate b (and substituting from 4.5)

$$0 = f_{avg} \cdot (1-b) + f_{min}$$

$$b = f_{avg} \,/\, (f_{avg} \cdot f_{min})$$

4.3.1.2 Algorithmic Details for Roulette Wheel Selection

Conceptually the roulette wheel selection method follows the figure shown in 4.2. In practice this is achieved by first generating a random number between 0 and 1. The proportionate scaled fitness of every chromosome (which is the scaled fitness of chromosome/total scaled fitness) in the sorted population is then individually summed, each time checking that the sum does not exceed the generated random number. Once this occurs, then the chromosome that was added last is taken as the selected one, and is returned by the algorithm.

The random number generation implemented here is the C library routine rand(). This function generates pseudo-random numbers between 0 and a RAND_MAX value. The rand() function is likely to follow a linear congruential generator, which defines a relationship between the last random number generated and the current one, hence resulting in pseudo-random number generation.

The function rand() has the advantage of being fast and straightforward to implement. For the purpose of the GA, it is therefore deemed adequate.

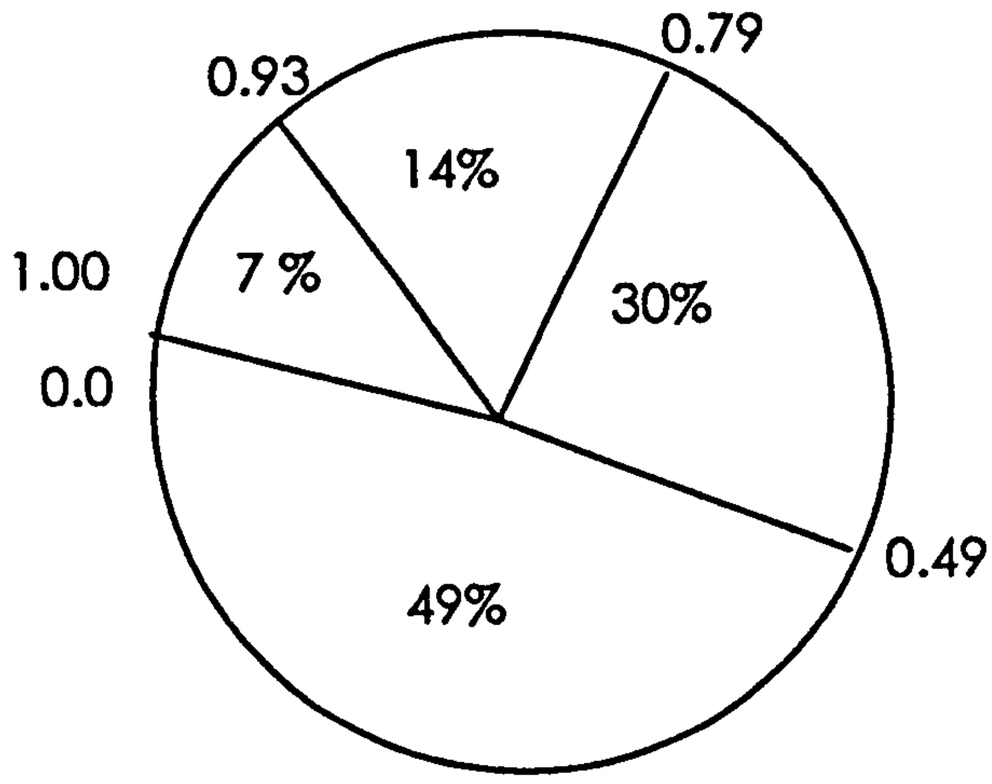


Figure 4-2 Hypothetical roulette wheel used as the selection operator. Every chromosome is given a segment on the wheel. The size of the segment is proportional to the fitness of the chromosome. The numbers surrounding the wheel represent each segment's proportion of the wheel. The selection of a chromosome, or segment, begins by generating a random number between 0 and 1. The segment into which that number falls results in the selection of its representative chromosome. For example if the random generated is 0.39, then the largest segment (49%) is selected, since 0.39 within its range in the roulette wheel. Similarly if the random number generated is 0.95, then the chromosome represented by the smallest segment (0.7%) is selected. Since the larger segments, representing the fittest chromosomes, cover a larger range of the total, they are more likely to be selected.

4.3.2 Crossover

The roulette wheel method selects chromosomes within the population that will undergo crossover, and that will therefore pass their genes onto the next generation. The method of crossing over that has been implemented here is single point crossover. Two parent chromosomes are selected by roulette wheel selection. These are then split at a random breakpoint and the resulting segments are swapped (see Figure 4.3). The newly-created chromosomes replace their parents, and this process of parent selection and crossover continues until the entire population has been

regenerated. An elitist strategy is also applied, in that the best two chromosomes of the parent generation are passed into the new generation unaltered.

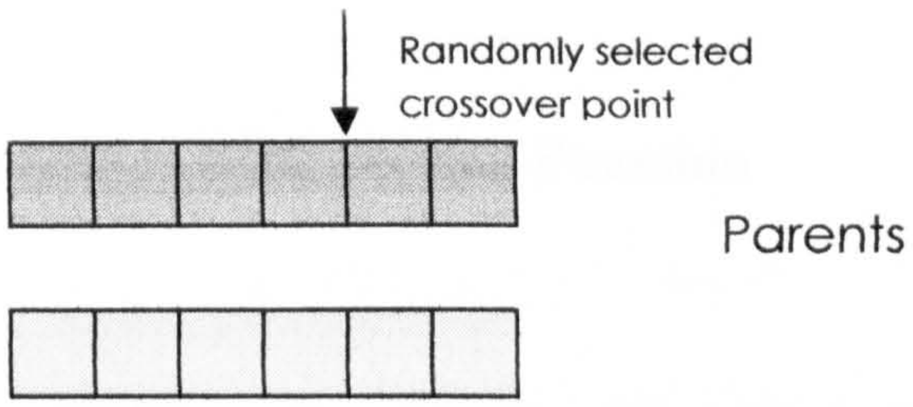


Figure 4-3 A single point crossover operation. The two parent chromosomes are split at a random point and the resulting portions are swapped to result in the two offspring.

4.3.3 Mutation

The final genetic operator, mutation, allows for discrete changes in the values represented by the genes. Mutation is performed on the population that has resulted from the crossover operator. A user-defined probability (the mutation rate) determines whether a given gene is to be mutated. The gene then undergoes a change by a particular step size. The step size is determined randomly, and cannot exceed a certain threshold. The threshold for a translation gene is 2.0 Å and for a rotation gene is 360°. When a gene is randomly selected for mutation, a step-size not exceeding the threshold for that type of gene is also randomly generated, and the gene is adjusted by

that step-size. The mutation function therefore loops through every gene in the population, randomly selects a gene and mutates this by the randomly-generated step size.

4.4 The GRID Scoring Function

The GRID scoring function (Goodford, 1985) is a molecular mechanics force field used throughout this study to describe the interaction energy between ligand atoms and the protein target. It is used to create molecular interaction grid maps for specified regions of a protein target.

GRID calculates the non-bonded interaction energy for various different probe types for an orthogonal grid which covers a user specified region of the protein. The GRID program requires a user defined centre co-ordinate point to determine the position of the simulation box on the protein target. Furthermore, the dimensions of the simulation box are also defined by the user to obtain the required box size. The default grid resolution is set at 0.5 Å and each functional group probe is placed at every grid point defined within the GRID box.

The GRID empirical energy function consists of van der Waals, electrostatics and hydrogen bond functions. Therefore the non-bonded interaction energy E_{xyz} can be calculated from these three terms shown in the equation below:

$$E_{xyz} = \sum E_{vdw} + \sum E_{el} + \sum E_{hb}$$

Equation 4.9

where E_{vdw} is the Lennard-Jones potential for calculating the van der Waals energy, E_{el} is the electrostatics function and E_{hb} is the hydrogen bonding function.

The Lennard-Jones potential in GRID is calculated using a 12-6 function given by the equation below:

$$E_{vdW} = \varepsilon\left(\frac{A_{ij}}{r^{12}}\right) - \left(\frac{C_{ij}}{r^{6}}\right)$$

Equation 4.10

The details of this energy function have been discussed previously in section 3.2.3.2.4.2.

The electrostatics interaction E_{el} are evaluated pairwise between probe and protein. However, the value of E_{el} is critically sensitive to the spatial dielectric behaviour of the environment. GRID assumes *a priori* that the environment surrounding solution, ζ, has a bulk dielectric of 80. The protein phase, ε, has a dielectric that reduces ζ towards 4 in the centre of the protein. The depth of each protein atom (S_q) in the protein phase is calculated by counting the number of neighbouring protein atoms whose nuclei lie within a distance of 4 Å. The depth of the probe (S_p) at each xyz position is assessed in a similar manner. Equation 4.11 below describes E_{el} :

$$E_{el} = \frac{pq}{K\zeta}\left[\frac{1}{d} + \frac{(\zeta - \epsilon)/(\zeta + \epsilon)}{\sqrt{d^2 + 4s_p s_q}}\right]$$

Equation 4.11

where p and q are the electrostatic charges on the probe and pairwise protein atom that are separated by a distance, d, and K is a combination of geometrical and natural constants.

Default cut-off distances of 3.5 Å, 8 Å, and 12 Å are applied for the hydrogen bond, van der Waals and electrostatics calculations respectively. In addition, the maximum positive (unfavourable) interaction energy permitted by any single grid point is restricted to 5.0 kcal/mol.

The hydrogen bond function implemented in GRID is a direction dependent 8-6 function consistent with the hydrogen bond parameters of GRID as described by equation 4.12:

$$E_{hb} = [C_{ij}/d^8 - D_{ij}/d^6] \cos^m \theta, \qquad \text{Equation 4.12}$$

where

$$C_{ii} = -3E_{min}(2R_{min})^8,$$

$$D_{ii} = -4E_{min}(2R_{min})^6$$

E_{min} is the minimum in the potential energy well when two identical atoms of type, *i*, are interacting and R_{min} is half the distance between the atoms at this point. The energy from a hydrogen bond is angle dependent within this function. If a receptor donates a hydrogen bond, θ is the angle DHP where D is the protein donor, H is the hydrogen atom of the donor and P is the probe. If the receptor accepts a hydrogen bond, θ, is the angle ALP where A is the protein acceptor, L is the lone pair of electrons on the acceptor and P is the probe. The term m is a constant set at 4 and E_{hb} is set to zero when $\theta < 90°$.

The number of hydrogen bonds that can be accepted and donated is specified in the GRID parameters file for both the probes and protein. To maximise interactions hydrogens and lone pairs are rotated to orientate them for optimal interaction and only the most energetically favourable hydrogen bonds are selected. Hydrogen atoms and lone pair electrons are computed from the heavy atom co-ordinates of the protein according to the method described by Jackson *et al.*, (1998).

4.4.1 The Probe Map Files

Calculating interaction energies takes up valuable computational time. If every ligand pose's raw interaction energy was calculated during the GA run there would be a substantial increase in execution time. One of the way of dealing with this issue is to

calculate interaction energies at different points on the protein binding surface prior to running the pose-generating algorithm (the GA in this case), and to store these in look-up tables, which are then accessed during the run of the algorithm. The look-up tables are known as the probe map files and are generated by a program called Liggrid. The GRID scoring function has its own set of probes, and these are functional groups with individual parameters and constitute groups of atoms that represent different parts of a ligand, following a united atom approach. To create the probe map files, a grid box is placed on the protein binding site surface, and the interaction energies are calculated between all the points on the grid box and the protein. Each probe type is placed at every grid point, its interaction energy with the protein is calculated and stored in the probe map file. Therefore before a GA run, a ligand will have all of its atoms assigned probe types. This can be done manually, by observing the local environment of each atom in the ligand, or using an automated program, known as gmol2. During a GA run, when a given pose's fitness is being assessed, the position of each atom in the ligand is used to determine which energies to look up in the relevant probe map file. Figure 4.4 is a schematic of probe map file generation.

But as one would expect, the ligand atoms do not always fall exactly at a unit grid box vertex. By taking into account the energy values of the eight vertices of the unit box which contains the atom, and the distances between the atom and the vertices, trilinear interpolation can be applied to calculate the interaction energy of the probe with the protein at that particular point.

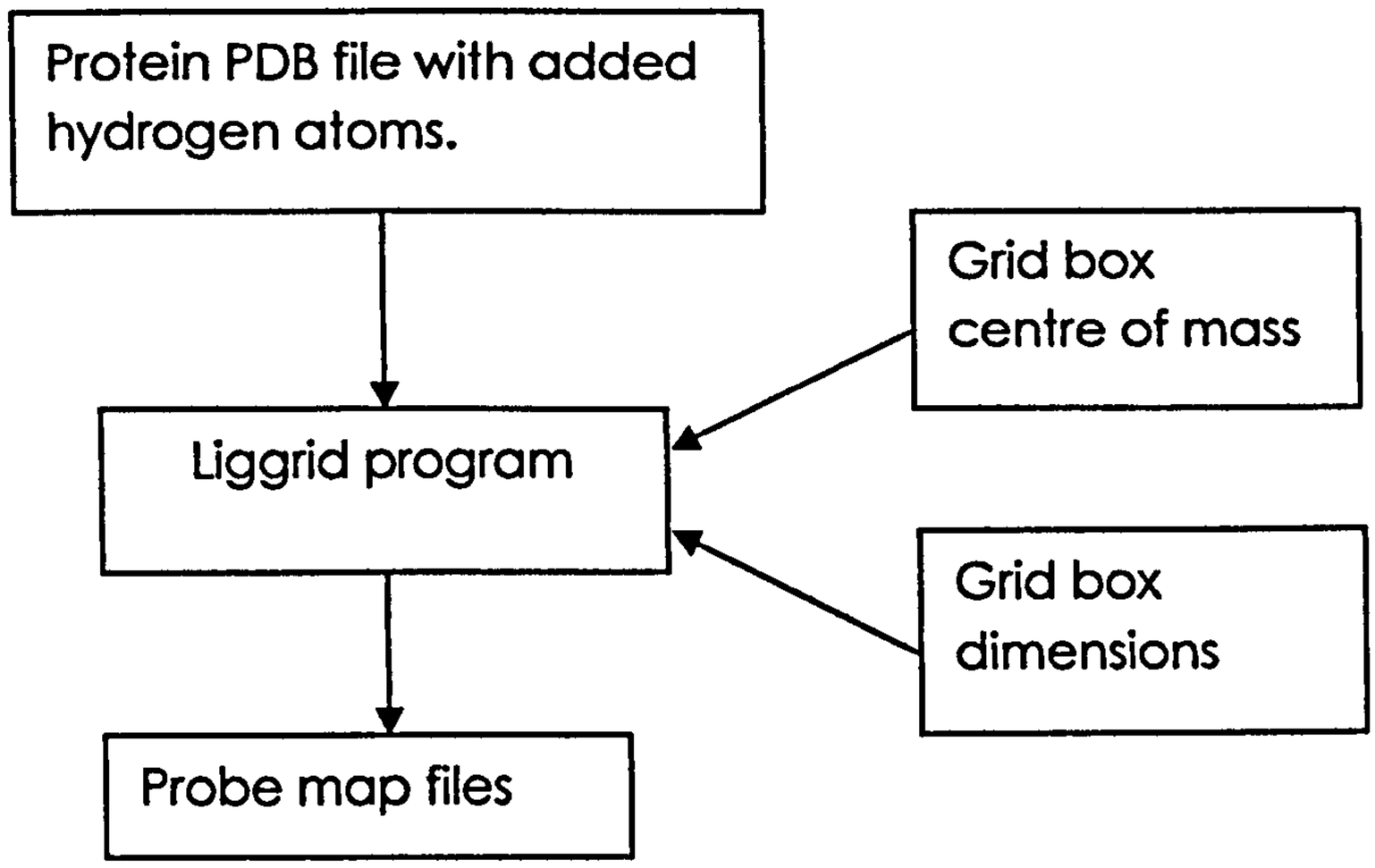


Figure 4-4 The generation of probe map files using the Liggrid program. The protein PDB file, a specified grid box centre of mass and box dimensions are the required input for the program.

4.4.2 Trilinear Interpolation

First the coordinates of the bottom left point of the unit grid box containing the atom need to be determined.

$$low(i) = round(coord(i)/grdspc) + 1 \qquad \text{Equation 4.13}$$

Where *coord(i)* is the x, y or z coordinate of the atom, grdspc is the grid spacing, i.e. the dimensions of a unit grid box (a parameter of Liggrid) and *round*() rounds the operation enclosed between the brackets to the nearest whole number.

Next the distance between the atom and *low(i)* known as the fractional distance (*crd(i)*), is calculated.

$$crd(i) = coord(i)/grdspc - floor(coord(i)/grdspc) \qquad \text{Equation 4.14}$$

where *floor*() rounds down the float resulting from the operation in brackets.

As mentioned earlier, the interaction energy at each vertex of a unit grid box is stored linearly in the probe map files. This is a listing of the interaction energies of all the vertices of the grid box. To find out the energy values of the eight vertices surrounding the probe, the index i.e. the position in the probe map file of a vertex needs to be determined. This is done in the following way.

$$indx1 = grdpts(1).grdpts(2).(ord(3)-1) + grdpts(1).(ord(2)-1)+ ord(1) \quad \text{Equation 4.15}$$

where *grdpts(1)*, *grdpts(2)* and *grdpts(3)* are the dimensions (number of unit grid boxes) along the x, y and z axes respectively, and *ord(1)*, *ord(2)* and *ord(3)* are the x, y and z coordinates of a vertex.

Equation 4.15 is repeated to obtain the indices of all eight vertices (*indx1...indx8* shown in Figure 4.5). Using these indices, the energy values are obtained from the probe map files, referred as *raw1...raw8*, which correspond to the energy values at *indx1...indx8* respectively.

The interaction energy contribution (*a1...a8*) at each vertex is determined as follows.

$$a8 = raw8$$

$$a7 = raw7 - a8$$

$$a6 = raw6 - a8$$

$$a5 = raw5 - a8$$

$$a4 = raw4 - a8 - a7 - a6$$

$$a3 = raw3 - a8 - a7 - a5$$

$$a2 = raw2 - a8 - a6 - a5$$

$$a1 = raw1 - a8 - a7 - a6 - a5 - a4 - a3 - a2 \qquad \text{Equation 4.16}$$

Finally, using the energy values derived above and the fractional distances, trilinear interpolation is performed, which returns the overall interaction energy of the probe at that point.

$$probeEn = a1.crd(1).crd(2).crd(3) + a2.crd(1).crd(2) + a3.crd(1).crd(3) + a4.crd(2).crd(3)+a5.crd(1)+a6.crd(2)+a7.crd(3)+a8 \qquad \text{Equation 4.17}$$

where *probeEn* is the interaction energy of the probe.

The above calculations are performed on all the atoms of the ligand at run time, and the resulting interaction energy values of each probe (*probeEn*) are summed up to give the interaction of the entire ligand.

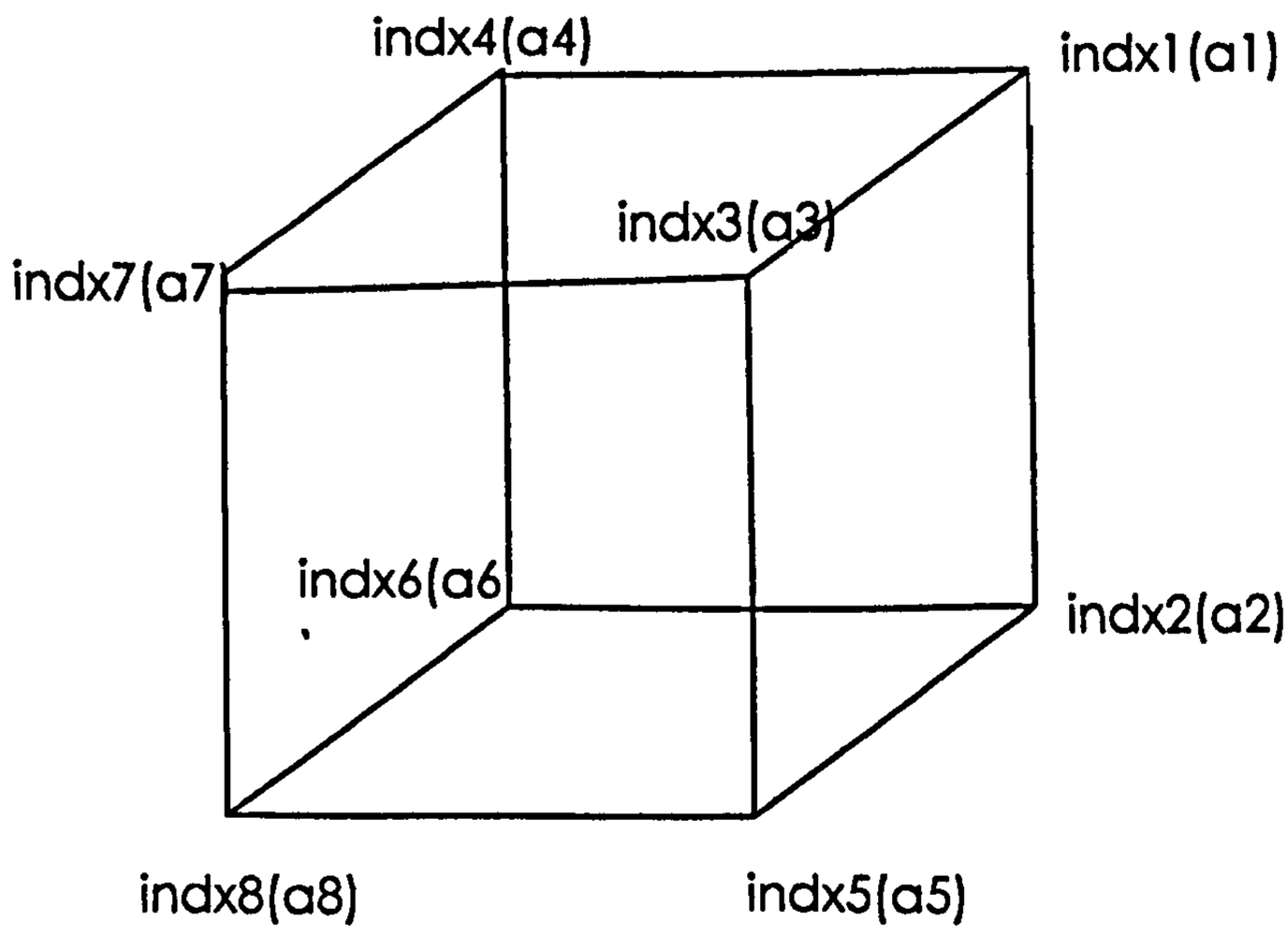


Figure 4-5 Box representing a unit of the Grid box that is placed on the protein binding site. *i1-i8* represent the vertices of the indices used in trilinear interpolation.

4.4.3 Bumps

The bumps file (.bmp) is one of Liggrid program's output files. It contains details of information on the grid box (size in Angstroms of a unit cell, box dimensions and the

global coordinate of the box origin). More importantly, it contains the grid box's ``bumps", which is basically a series of 'T' and 'F' characters, one for each cell of the grid box. A 'T' at a particular vertex indicates an overlap with the protein at that point, whereas an 'F' means there is not. Whenever a new pose is generated, a count of the number of atoms which fall within cells containing bumps (labelled 'T') is taken. If the count is higher than a threshold value, which has been set at 40%, then that pose is rejected. The percentage value represents the portion of the ligand allowed to 'bump' before rejecting a pose. This is time-saving since it avoids scoring poses which sterically clash with the protein, and which will inevitably have high, unfavourable scores.

4.5 The Genetic Algorithm Structure

The genetic operators described in section 2.2.1 together evolve the population until some termination criterion has been met. Each operator has a specific role in the SGA structure, which is illustrated in Figure 4.6. The overall strategy of the SGA is generational, meaning that the offspring population created by the genetic operators replaces the parent population entirely, with the exception of the top two chromosomes of the parent population which are passed to the offspring population without change. This gives an elitist element to the algorithm and ensures that top solutions are never lost from the population unless better solutions are found.

A GA run begins when the initial population is formed by generating chromosomes that consist of randomly generated genes. These chromosomes are scored by the GRID scoring function by firstly applying the rotations and translations incorporated in the genes to the reference ligand, as described in section 4.2. Next, the reproduction operator selects chromosomes which the crossover operator uses to generate new chromosomes. The mutation operator is then applied, and the resulting population is regarded as the new generation. This generation is scored, and the cycle continues, until the termination criterion has been met. The termination criterion in this case stops a run of the algorithm at 30 000 generations.

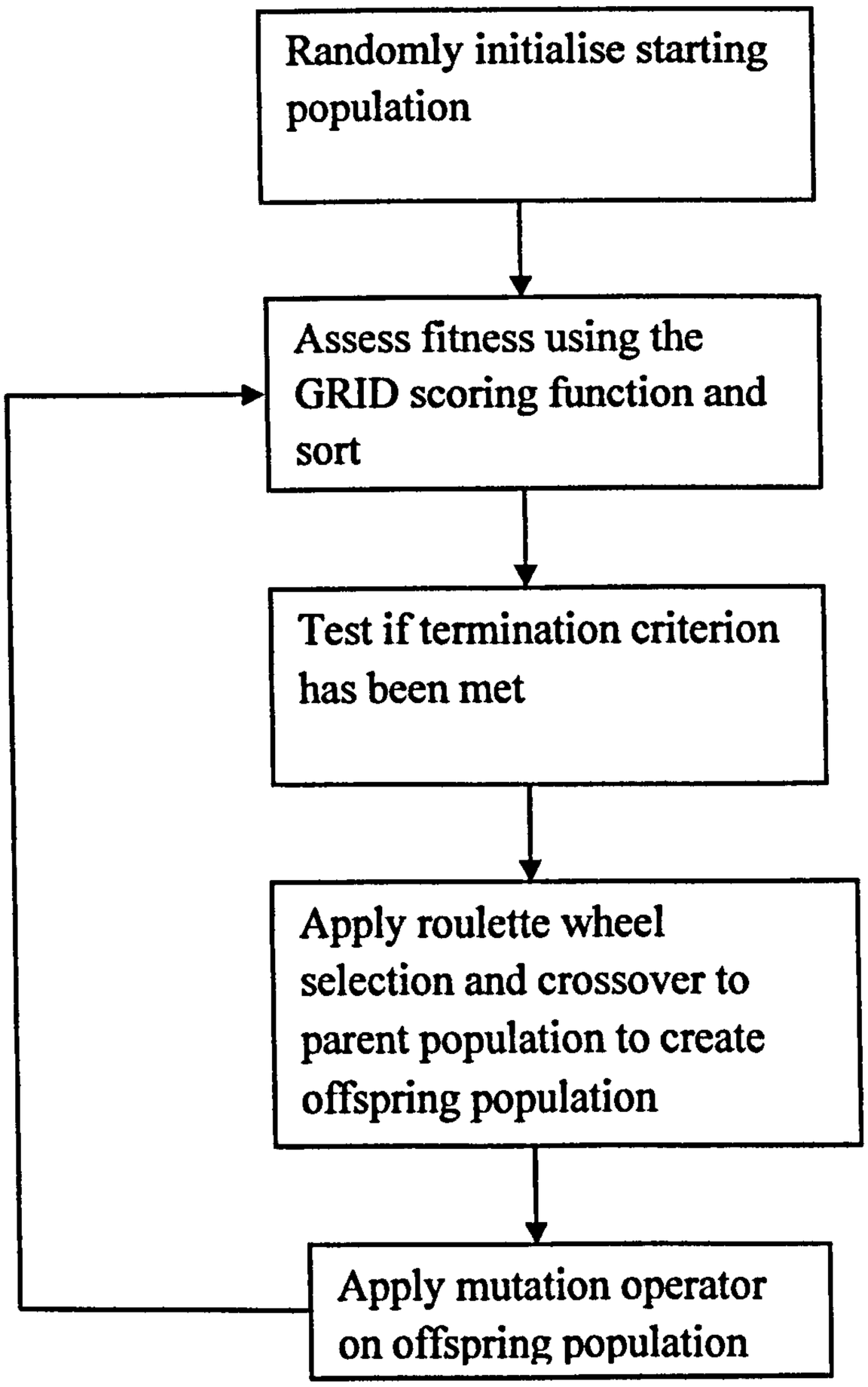


Figure 4-6 Schematic of the SGA.

4.6 Q-fit overview

The Q-fit method is a fragment based rigid-body ligand docking algorithm (Jackson, 2002). It uses a grid-based deterministic search algorithm, the basis of the algorithm relies on probe 3D energy grid maps created from the protein target. These molecular interaction field maps are created with the molecular mechanics force field GRID (see section 4.4) to calculate interaction energies against the protein for all functional groups contained within the ligand.

Q-fit uses a probabilistic model with a geometric matching methodology to simulate ligand binding. The Boltzmann principle is used to describe ligand conformations most likely to bind to a protein.

Q-fit determines a ligand binding conformation on an energetics basis. The program calculates the interaction energy of a ligand within an active site using functional probes (representation of a ligand atom/functional group in the GRID scoring function). Therefore, the energetically most probable conformation can be identified by calculating each individual atom's contribution with respect to any specific atomic configuration.

Any given ligand can be represented by a set of functional probe groups which relate to a set of molecular interaction field grid maps of a protein. Figure 2.1 is a schematic of the method utilised to place a small ligand fragment in the most energetically favourable position within a protein active site. Given the pre-calculated 3D grid maps of atom preferences the program reads those appropriate for the atom probe types in the given ligand.

For each map the grid point interaction energies are then sorted in order of decreasing interaction energy with the receptor so that the most favourable (i.e. negative) interaction energy is at the top of the list and the least favourable at the bottom (Figure 4.7A).

The top N locations for each probe type are pooled into a single list and again sorted in order of decreasing energy. This creates a set of M interaction points (N x no. of probe types representing the ligand) that are energetically sorted to provide the most favourable binding site locations.

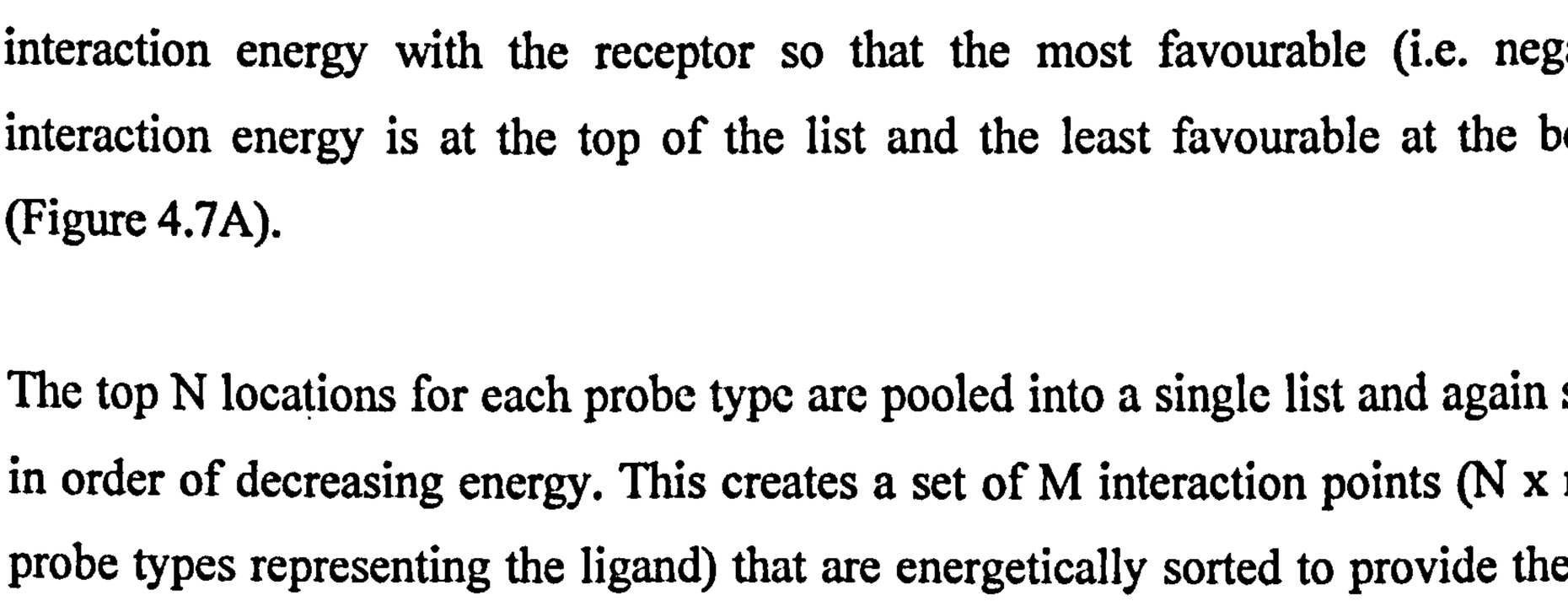


Figure 4-7 Schematic representation of steps (A-C) involved in placing a small ligand fragment in the most energetically favourable position within a protein active site. Adapted from (Jackson, 2002).

The M interactions site list from the receptor protein is then searched for low energy triplets that correspond to triplets found in a ligand (Figure 4.7 (B and C)). The search procedure is based on a pose clustering (Rarey *et al.*, 1996). The M interaction sites are systematically searched to find a receptor pair (m_i, m_j) that is coincident to a ligand pair (l_i, l_j) such that the distances ($|m_i - m_j| - |l_i - l_j| < \delta$). The value of δ is set to 3.0 Å and is the maximum difference in distance between matched ligand atom-atom, and grid interaction site grid-grid distances. Then a further point match is identified that creates the most favourable triplet interaction energy and is also coincident to the ligand triplet (l_i, l_j, l_k) such that the distances ($|m_i - m_k| - |l_i - l_k| < \delta$) and ($|m_j - m_k| - |l_j - l_k| < \delta$). Therefore these triplets are classified as δ-compatible and are stored for the transformation stage which maps the ligand to the receptor.

The ligand triplet solutions from pose clustering are transformed onto the receptor interaction sites using a least squares fitting routine (McLachlan, 1979) Following placement an inter-molecular clash filter (where a protein atom overlaps with a probe) is applied. In this way ligand fragments consisting of probe triplets are docked in the protein binding site. The output from Q-fit is a file containing a list of poses generated by the algorithm ranked in order of increasing interaction energy. All poses within the ranked list are energy minimised using a downhill Simplex algorithm (Nelder and Mead, 1965) implemented according to Gschwend and Kuntz (Gschwend and Kuntz, 1996).

4.7 SGA Parameters

Before testing the SGA on a set of protein-ligand complexes, its parameters were optimised using three rigid-body protein-ligand complexes: trypsin with benzamidine (3ptb), pentosyltransferase with guanine (1ulb), and l-arabinose binding protein with

alpha-L-arabinose (1abe). The optimal consensus parameters obtained from these cases, and which were applied on the rest of the test cases, are listed in table 1.

The probe map files are generated using Liggrid (as described in section 4.4.1) and these are used by both Q-fit and the SGA for scoring the generated poses during a run. The input required by Liggrid (and as shown in figure 4.4) are the box dimensions and centre of the box. The box dimensions represent the search space and these are assigned manually, depending on the size of the ligand, by allowing at least 7 Å between the outside of the ligand and the edge of the box. The centre of mass of the box is generated by taking the x, y and z coordinates of the centre of mass of the ligand crystal structure and randomly adjusting these, in order to remove the introduction of any bias from experimental data. This process of generating input parameters for liggrid applies to all test cases described in the thesis.

Parameter	Value
Population	150
Number of generations	30000
Mutation rate	20%
C_{mult}	1.2
Rotation step size	2π rads
Translation step size	2.0 Å

Table 4.1 SGA paramters

4.8 Results from Dataset 1

The SGA was tested on Dataset 1, a dataset that was also used to test Q-fit and which is described in section 4.6. The purpose of this section is to test whether the SGA is reasonably able to find good solutions, i.e. low-energy solutions which have rmsds below 2.0 Å from the crystal structure, the accepted threshold for determining the

success of a docking run. Both the SGA and Q-fit use the GRID scoring function to assess solutions and guide the search during a run. Because both algorithms use the same scoring function then it is feasible to compare the magnitudes of the interaction energies of solutions from the two algoritms. By comparing the top-ranked SGA solution with the top-ranked Q-fit solution it is possible to gain insight into the performance of the SGA in terms of its ability to successfully find solutions which have minimised interaction energies. This comparison presumes that Q-fit has found solutions which are close to, or at, the global minima. This is a reasonable assumption to make where the rmsd of the top-ranked solution is below the 2.0 Å threshold.

Looking at the table of results (Table 4.2), it can be seen that the SGA obtained solutions with rmsds below 2.0 Å for six of the ten complexes in Dataset 1. The four complexes which the SGA did not find correct, top-ranked, solutions for are 3tpi, 1ulb, 4dfr and 3ptb. For two of these, 4dfr and 3ptb, Q-fit also did not find correct solutions in its top ranks. Looking at the interaction energies obtained for 4dfr, it can be observed that Q-fit was able to find a solution with much lower energies than the SGA, though the rmsd of this solution is high. With 3ptb, the interaction energy of the Q-fit solution is slightly lower than that of the SGA, and both solutions have similar rmsds that are higher than 2.0 Å. With 3tpi, the SGA obtained a top-ranked solution with a higher interaction energy than Q-fit's solution, and also has a high rmsd. With 1ulb, the SGA obtained a solution with a lower interaction energy than the Q-fit solution (-28.04 kcal/mol versus -35.00 kcal/mol for the SGA solution). Interestingly the rmsd of the higher energy Q-fit solution is below the 2.0 Å threshold, whereas that of the SGA solution is not. It would be expected that a lower energy solution should have an rmsd that is closer to the crystal structure. This anomaly may imply that the crystal structure is not close to the global minimum, and that the scoring function is not capable of associating low-energy solutions with poses that are close to the crystal structure. For the rest of the complexes, both algorithms obtained solutions with good rmsds for the top-ranked solutions. Generally Q-fit's solutions have slightly lower energies than the SGA's, with the exception of 1dbb, for which the SGA's top-ranked solution has an energy of -27.92 kcal/mol and Q-fit's solution an energy of -26.93 kcal/mol.

These results indicate that the SGA can, overall, find solutions which have low rmsds and are therefore correct, and that the algorithm and the genetic operators employed are capable of conducting an effective search for these solutions. Athough Q-fit did outperform the SGA in several of the test cases, in terms of both energies and rmsds, it is worth noting that Q-fit performs the additional step of locally minimising its final solutions which serves to further minimise their interaction energies. This step is not performed in the SGA. As with most optimisation algorithms that are controlled by different parameters, it is quite likely that further tweaking and experimenting with the various parameters, such as increasing population size, or running the algorithm over longer generations, will lead to improvements in the overall performance of the SGA. However, since the purpose of this work is to study protein-ligand docking in a multiobjective optimisation context, it was decided that the current performance of the SGA is adequate and provides a good foundation for adapting it to a multiobjective algorithm. All adjustments needed to improve performance can be carried out on the ensuing multiobjective optimisation algorithm. The development of the algorithm, along with the refinements implemented, is discussed in the following chapters.

		Interaction energy (kcal/mol)	Rmsd (Å)
2gbp	Q-fit	-44.01	0.48
	SGA	-39.00	0.64
1ldm	Q-fit	-36.57	0.72
	SGA	-35.88	1.05
2phh	Q-fit	-34.26	0.52
	SGA	-31.89	0.59
3tpi	Q-fit	-49.65	1.23
	SGA	-16.59	10.47
1stp	Q-fit	-26.72	1.10
	SGA	-25.19	0.95
1dbb	Q-fit	-26.93	1.2
	SGA	-27.92	0.47
4dfr	Q-fit	-32.99	7.1
	SGA	-26.6	8.44
3ptb	Q-fit	-35.26	2.2
	SGA	-34.29	2.3
1abe	Q-fit	-36.38	0.62
	SGA	-35.50	0.82
1ulb	Q-fit	-28.04	0.64
	SGA	-35.00	4.2

Table 4.2 Energies and rmsds of top ranked solutions obtained by docking Dataset 1 using Q-fit and SGA.

5 Conversion of SGA to a Multiobjective Genetic Algorithm

Protein-ligand docking was described in Chapter 3. It is clear from that discussion that the quality of different ligand poses is assessed by one of many different types of scoring functions. Many of these scoring functions calculate and add the different interaction energy types together in a function known as the master equation, to produce, as output, a single energy value. Based on this value, the quality of the ligand is assessed in relation to other generated poses. These scores are also used to drive the conformational and orientational search towards a global minimum.

As was mentioned in Chapter 3, the aim of this thesis is to apply multiobjective optimisation to the protein-ligand docking problem. Because several of the features implemented by a single-objective, standard genetic algorithm (SGA) are also common to multiobjective genetic algorithms (see section 2.4), a standard genetic algorithm (SGA) was developed in Chapter 4 with the purpose of converting it to a multiobjective algorithm. As the results of testing the SGA on Dataset 1 have shown, the SGA is adequate at finding near-optimal solutions, and the decision was therefore taken to adapt the algorithm for multiobjective optimisation. This chapter describes the new features and the changes that were performed on the SGA to do this, and the chapter which follows (Chapter 6) describes the datasets used in testing the algorithm, as well as the results obtained.

5.1 Structure of algorithm

The multiobjective optimisation algorithm that was developed here follows the structure of an elitist non-dominated sorting GA, or NSGA-II for short (Deb *et al.*, 2000), and which was described in section 2.5.4. This highly elitist method ensures that only the best solutions are passed on to the next generation - and that no good solution is ever lost. The only way that a good solution is eliminated from the population is if a better solution is found. By keeping elites in a population the

probability of finding good solutions is increased. The version of the NSGA-II described here differs in two aspects from the original NSGA-II developed by Deb *et al.*, (2000). The roulette wheel selection method is used (described in section 4.3.1.2) as the genetic operator which contributes towards the formation of the offspring population, rather than binary tournament selection. The second difference is that a fitness-sharing niching procedure is used to select solutions from within the same rank, rather than the crowded tournament selection operator. The requirement for niching can arise in several situations, and which will be described below. The reason for modifying the NSGA-II algorithm is because both of the methods implemented here perform the same functions as the originals, i.e. both roulette wheel and the binary tournament selection find a selection of good solutions for recombination, and the crowded tournament selection operator and niching are both capable of finding the solutions which are in the least crowded sections of a front. Also at the time of development it was more convenient to use a modified NSGA-II in order to minimise changes made from previous versions of the algorithm.

A flowchart detailing a single generation of the algorithm is shown in figure 5.1. In summary, the initial chromosome population (N) is created at random, using C's random number generator function rand(). The individual chromosomes are scored using the different objectives, the Pareto ranks of the chromosomes are determined, and the population is sorted based on their ranks. The selection, crossover and mutation operators are applied to produce an intermediate population. This is combined with the parent population ($2N$) and the combined population is Pareto ranked. The combined population of size $2N$ is reduced to N by selecting the highest ranked N chromosomes. The details of these methods are described below.

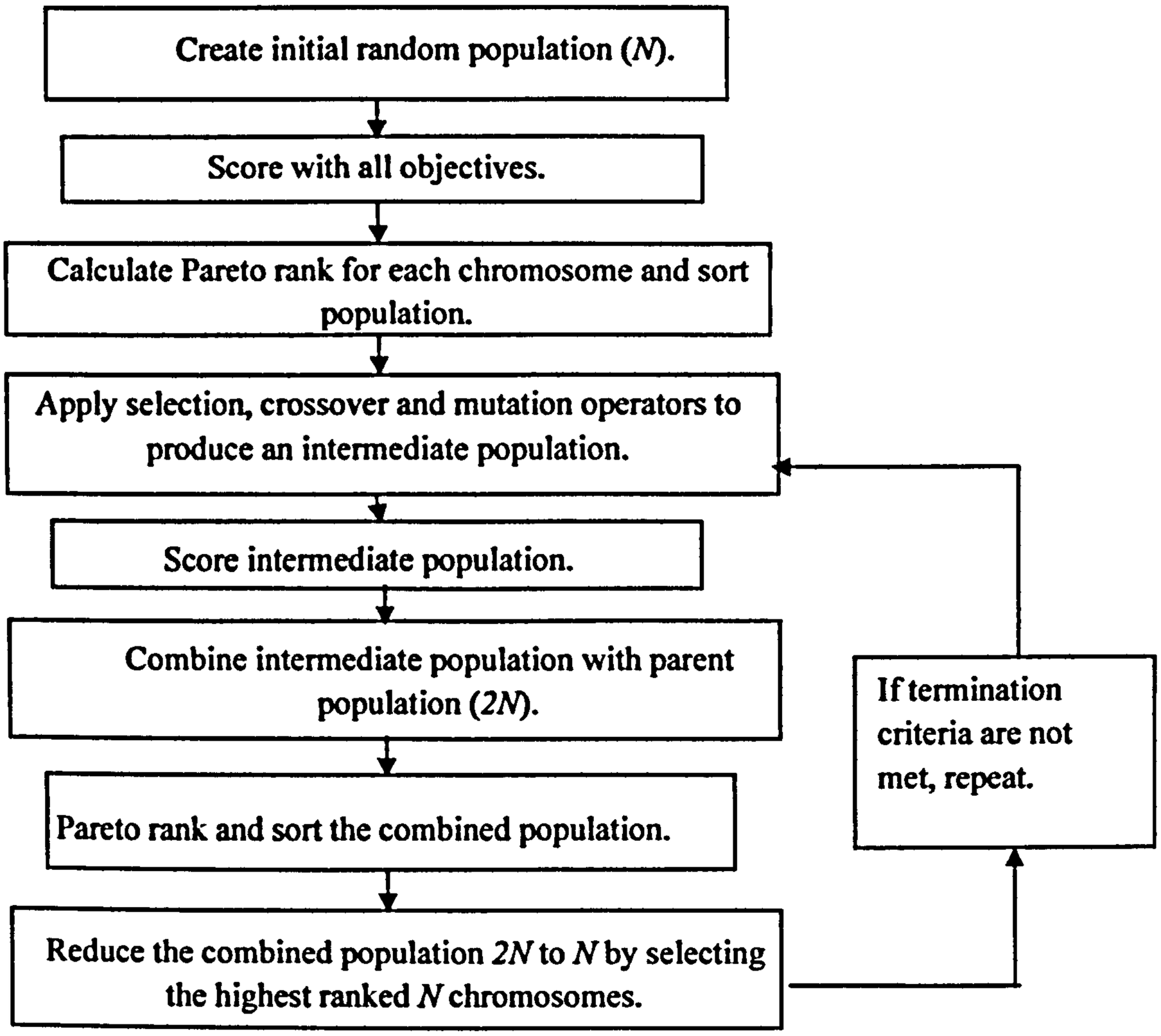


Figure 5-1 Schematic of the MOGA which follows a NSGA-II structure.

5.2 NSGA-II: details of the algorithm

The NSGA-II was adapted from the SGA described in Chapter 2, and many of its features have not been modified. What follows is a description of the features introduced in the NSGA-II.

5.2.1 The objectives

As was stated in Chapter 3, the purpose of developing an NSGA-II for protein-ligand docking is so as to investigate the relative contributions of individual interaction types for protein-ligand docking optimisation. The GRID scoring function (Goodford, 1985), as described in section 4.4, incorporates electrostatic, vdw and hydrogen bond donor and acceptor calculations in its assessment of interaction energy. The NSGA-II was developed to optimise two objectives. These were:

- Electrostatic and hydrogen bond donor and acceptor energies combined (referred as electrostatic and hydrogen bond energies from here on).
- Vdw interactions.

Although both of these objectives are energy types and both have the same units (kcal/mol), they are not necessarily commensurate. A certain ligand pose can have the preference to form predominantly electrostatic and hydrogen bond interactions within the protein active site, it may prefer to form mainly vdw interactions, or it may form interactions that are a relatively equal balance of both energy types. Also, different poses of a given ligand can have the same total energy, but these can have different balances of the two objectives; for example a change in orientation can decrease the electrostatic interaction energy but increase the pose's vdw interaction energy, withough changing the overall total energy. A multiobjective algorithm would be able to differentiate between these two situations, unlike a single objective approach. The possibility that a ligand may have preference to a particular interaction energy type within the protein binding site indicates that in these situations the two different objectives are conflicting, and that a multiobjective approach would be more able to quantify this than a composite function such as those that are used in single objective approaches.

It is believed that highly populated docked conformations of ligands (such as those found in crystal structures) correspond to minima in the free energy of a complex system comprising ligands, protein and solvent, that is, minima of a single objective function with a single defined balance of non-bonded and entropic contributions. The

accurate calculation of this free energy is, however, problematic. Docking studies tend to be limited to optimisation of highly empirical functions which try to capture some of the important contributions to free energy. In this context, the appropriate balance of the various terms is not clear, and the fact that the choice of different weightings leads to different solutions shows that the terms are competing and thus 'incommensurate'. A multiobjective approach allows investigation of all possible weightings between competing empirical energy terms.

It is also important to be aware of the fact that both these objectives are interlinked, in that the range of electrostatic and hydrogen bond energies is limited by the vdw interactions. Even though a protein and ligand atom may be forming favourable electrostatic and hydrogen bond interactions, this would be irrelevant if the two atoms' vdw interaction energy is high. It is therefore important to take this anomaly into account when splitting the interaction energy into different energy types. How this was performed with regard to the GRID scoring function is explained in section 5.2.1.1.

As was explained previously, multiobjective optimisation is achieved by Pareto ranking of chromosomes within a population. Since Pareto ranking requires that each individual chromosome in the population is assessed with the different objectives, each chromosome must be scored by the two objectives. As was shown in Chapter 4, chromosomes within a population generated by the SGA have their fitness assessed by the total GRID interaction energy. This, in turn, is performed through trilinear interpolation (described in section 4.4.2), from position-specific energy values extracted from probe map files which are generated by the Liggrid program. The same process of scoring chromosomes occurs with the NSGA-II, but rather than using the probe map files to estimate a single value representing the total interaction energy (as with the SGA), two sets of values, representing the two objectives, are calculated separately. This is done by using two different sets of probe map files. By scoring a given chromosome, or orientation, using values generated from the two sets of probe map files, and manipulating these, two values representing the two objectives are obtained, and are then used in the Pareto ranking of the population. The process of separating out the different components of the energy function is feasible through the adjustment of the probes.dat file, and which is explained below.

5.2.1.1 Editing of probes.dat file for estimating vdw energies

One of Liggrid's input parameters is the probes.dat file, a text file that contains the parameters needed by the program to calculate the interactions for the ligand's probes. These are arranged in table format; the first three columns represent van der Waals parameters, the fourth column holds charge information (for electrostatic interaction energy calculations), and the final four columns contain hydrogen bond donor and acceptor parameters. These parameters are used by the Liggrid program to generate the probe map files. Because the parameters which calculate these individual energy types are in an amendable text file, by setting parameters for one energy type only and setting the other parameters to zero, it is possible to run Liggrid using the edited probes.dat file to obtain probe map files which contain energies for one interaction energy type only. Therefore, to obtain a set of probe map files which contain vdw interactions only, the probes.dat file was amended to create a vdw version of the probes.dat file, referred to as probesV.dat. The process of amending this file is illustrated in Figure 5.2. The probesV.dat file contains the first three columns containing the vdw parameters. The parameters for the other interaction energy types are set to zero, as figure 5.3 shows. During a run of the NSGA-II, the trilinear interpolation method will use the vdw-specific probe map files to estimate the vdw interaction energy that a particular atom of the ligand is making with the protein. The vdw interaction energies are summed over all atoms to form the vdw objective used in Pareto ranking.

5.2.1.2 Electrostatic and hydrogen bond energies

The calculation of the electrostatic and hydrogen bond energies at a particular grid point occurs during a run of the NSGA-II. To obtain the electrostatic and hydrogen bond energies at a particular grid point, both vdw-specific as well as the total energy probe map files are required. During the NSGA-II run, to find the electrostatic and hydrogen bond energies for a particular chromosome's pose, the electrostatic and

hydrogen bond energies for all grid points involved in calculating a given atom's energy (section 4.4.2) are first determined. The total interaction energy for the grid points are obtained from the probe map files which contain total interaction energies. This is followed by extracting the vdw energies for the same points from the vdw-specific probe map files. For every grid point, the vdw energy is subtracted from the total interaction energy of that grid point to then obtain the point's electrostatic and hydrogen bond energy. The electrostatic and hydrogen bond energies for these grid points are then used to estimate that particular atom's energy (section 4.4.2). This process is repeated for all atoms of that pose and these are summed together. The resulting value, representing the electrostatic and hydrogen bond energies for that particular chromosome, is used during Pareto ranking of the population.

The reason for finding the electrostatic and hydrogen bond energies by subtracting the vdw energy from the total energy rather than calculating the electrostatic and hydrogen bond energies in the same manner as the vdw energies (i.e. directly from the probe map files) is to prevent the existence of artificially favourable electrostatic and hydrogen bond energies where the vdw energy is high (for example when there is a clash with the protein). This has the effect of smoothing the energy surface at these high vdw energy points (such as at the surface of the protein).

Given the situation where an atom's position allows it to form favourable electrostatic and hydrogen bond interactions, but unfavourable vdw interactions, for example if it clashes with the protein then the electrostatic and hydrogen bond energies are set to zero, which prevents favourable electrostatic and hydrogen bond energies from erroneously biasing these orientations positively. Therefore when estimating the electrostatic and hydrogen bond energies for a particular atom, the vdw energies of the grid points surrounding the atom are first checked. If any of these values are at a certain threshold (5.0 kcal/mol), indicating a clash with the protein, then the corresponding electrostatic and hydrogen bond energies of the grid point is given a value of 0. In this way the overall energy of the atom at that position is unfavourable and an atom that is making unfavourable vdw interactions will not be at an advantage if it is forming favourable electrostatic and hydrogen bond interactions. Figure 5.3 summarises the process of estimating the values for the two objectives.

(A)

```
C1=  1.90  6.  2.07  0.000  0.00 0.00 00 00  00  Aromatic CH group
C2   1.90  7.  1.77  0.000  0.00 0.00 00 00  00  Methylene CH2 group
*** Made up ***
C3   1.95  8.  2.17  0.000  0.00 0.00 00 00  00  Methyl CH3 group
N:   1.65  6.  1.10  0.000 -6.50 1.50 00 01  97  sp3 N with one lone
pair
N:=  1.65  6.  1.80  0.000 -5.50 1.55 00 01  97  sp2 N with one lone
pair
```

(B)

```
C1=v 1.90  6.  2.07  0.000  0.00 0.00 00 00  00
C2v  1.90  7.  1.77  0.000  0.00 0.00 00 00  00
C3v  1.95  8.  2.17  0.000  0.00 0.00 00 00  00
N:v  1.65  6.  1.10  0.000  0.00 0.00 00 00  00
N:=v 1.65  6.  1.80  0.000  0.00 0.00 00 00  00
```

Figure 5-2 Sections of the two files, (a) probes.dat and (b) probesV.dat. (a) contains information on the parameters needed to calculate the different interaction energy types (see main text). The probesV.dat file contains only the vdw parameters. This file is created by retaining the first three columns (which contain the vdw interaction parameters) and setting the rest of the columns, which represent other interaction type parameters, to zero. Both of these files are used as input for the Liggrid program to generate the necessary probe map files needed by the NSGA-II.

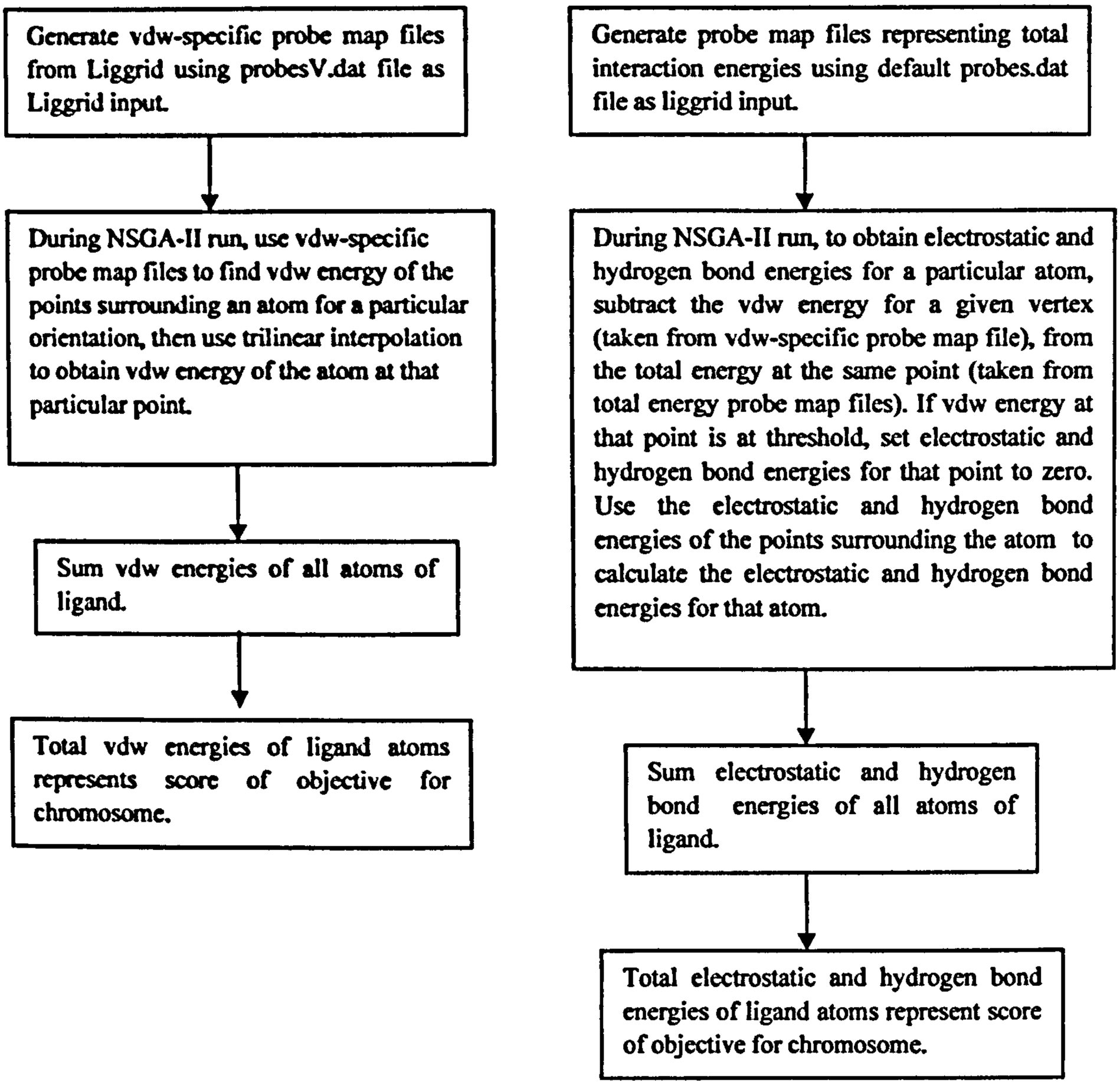


Figure 5-3 Schematic of the scoring methodology of the NSGA-II for the two objectives: vdw interactions and electrostatic and hydrogen bond energies.

5.2.2 The Pareto ranking function

Once the entire population has been scored for both objectives, the objectives are used to Pareto rank the chromosomes of the population. Pareto ranking was explained in detail in Chapter 2, and involves the comparison of all chromosomes against each other for a given objective. This function within the algorithm consists of two nested for loops that take each chromosome and compare its objectives against the other chromosomes' objectives. If any other chromosome within the population shows a lower energy value for *both* objectives then that chromosome's rank is incremented. In this way a count of the number of chromosomes that dominate a particular solution

is kept, which, at the termination of the loop, represents the Pareto rank of that chromosome. This process is illustrated in Figure 5.4.

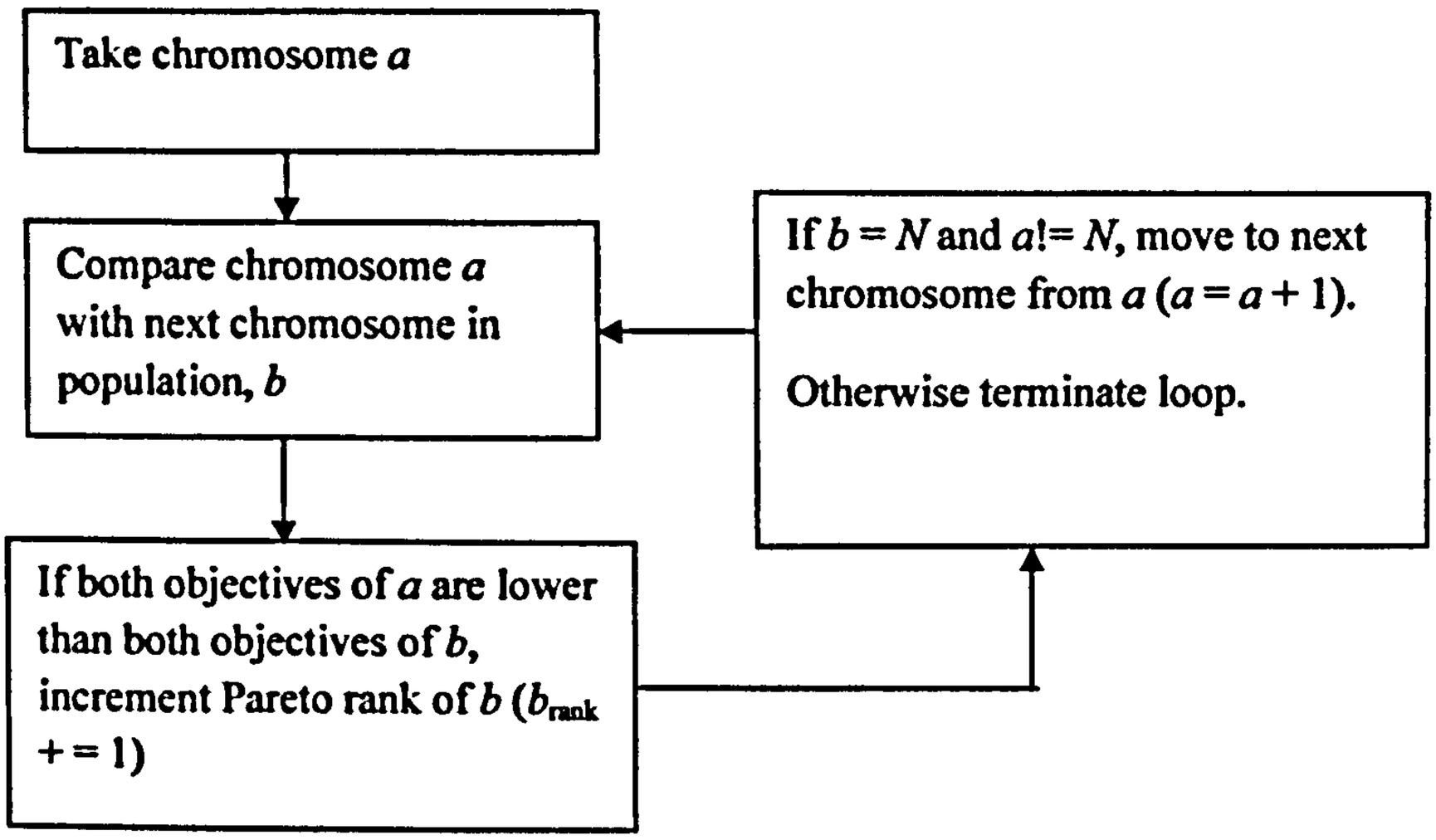


Figure 5-4 Pareto ranking of population.

After the entire population is ranked, the chromosomes within the population are sorted in ascending order. The process of sorting is necessary for the selection operator, which is explained below.

5.2.3 Producing the intermediate population

With the parent population ranked, the next stage of the algorithm is applying the genetic operators to the parent population to produce the intermediate population. This involves the selection of pairs of chromosomes for crossover to produce an intermediate population upon which the mutation operator is applied.

5.2.3.1 Selection of chromosomes for crossover

An effective selection procedure will bias the process of selection of chromosomes towards the fitter individuals in the population. At the same time it must ensure that less fit individuals are also selected, to maintain diversity in the population, and to prevent premature convergence. In the SGA described in Chapter 2, a roulette wheel selection procedure was implemented for the selection of chromosomes. Each chromosome is represented on the wheel by a segment with a size that is proportional to its fitness. With multiobjective optimisation it is the Pareto rank, rather than a raw score, that determines a chromosome's fitness. Therefore implementing roulette wheel selection in the NSGA-II requires that the segments on the wheel represent Pareto ranks rather than fitness values. For example, given a population of size 100, where the chromosomes are distributed over 35 ranks (including the non-dominated rank), then the roulette wheel will consist of 35 segments. The size of the segment is dependant on the rank it represents- the higher ranks will have larger segments and vice versa. This process is analogous to rank selection in SGAs.

For the NSGA-II, the roulette wheel selection procedure was adapted from that implemented in the SGA so that, at every generation, the number of segments on the wheel is equal to the number of Pareto ranks. The linear scaling procedure, which was used with the SGA, has also been implemented. This procedure, as was described in section 4.3.1.1, will ensure that the size of the segments are proportionately distributed on the roulette wheel, so that top ranks are not over-represented, and therefore avoids scenarios of flooding the population with individuals from the top ranks.

To apply the linear scaling procedure to the roulette wheel, first the "fitness" of each rank is modified so that each rank is inversely proportional to its position. Therefore, from the previous example, the top Pareto rank will have a fitness of 35, the second highest 34, etc. with the lowest rank having a fitness of 1. This is achieved using this equation:

$$\text{rank}_x = (N+1) - x \qquad \text{Equation 5.1}$$

where (1...x) is the number of each rank, $rank_x$ represents the "fitness" of rank x, and N represents the population size. Once the fitnesses of all the ranks has been determined, the linear scaling process is applied to these. During selection, the roulette wheel is spun, returning a rank (and not a chromosome, as with the SGA). A rank may encompass several chromosomes, and since one spin of the roulette wheel must result in only one chromosome, a second process must be implemented to select a chromosome from within a rank. It is also important to ensure that outliers from within a rank have the possibility of being selected, as these are necessary to help maintain the general diversity within a population. Niching enables these factors, and is described in the following section.

5.2.4 The niching function

Niching is important for maintaining diversity within a population, and is discussed in more detail in section 2.4.4. The opportunity to use niching occurs at different points during the run of the NSGA-II. As discussed above, it is applied when the selection of a chromosome is required from within a rank which contains more than one. Niching is also applied when reducing the intermediate population from $2N$ down to N at the end of a generation. Before the roulette wheel operator for selection is applied, each chromosome within a rank is allocated a specific fitness value (f_i`- the niched fitness). This value is proportional to the chromosome's niche count ($m_i$`).

$$f_i` = f_i / m_i` \qquad \text{Equation 5.2}$$

f_i is an arbitrary constant that represents the highest possible fitness that could be allocated to a chromosome, and in the case of the NSGA-II has been set to 100.

The niche count (m_i`) for a chromosome i is calculated as follows:

$$m_i` = \sum_{j=1}^{N} sh(d_{ij}) \qquad \text{Equation 5.3}$$

where $sh(d_{ij})$ represents a sharing function evaluated between i and all the chromosomes in the rank (N), including i itself, as follows.

$$sh(d_{ij}) = 1 - (d_{ij}/\sigma_{share}),\ d < \sigma_{share} \qquad \text{Equation 5.4}$$
$$= 0, \text{otherwise}$$

σ_{share} is a user-defined value that denotes the niche radius. d_{ij} is the distance, whether in objective or decision space, between the chromosome i and all other chromosomes j in the rank. The $d < \sigma_{share}$ condition ensures that chromosomes with d greater than the niche radius σ_{share} do not contribute towards i's sharing function $sh(d_{ij})$.
This means that if a chromosome is by itself in one niche then it will have a niche count of 1 (m_i=1) which will allow it to receive its full fitness share.

The "distance" between chromosomes depends on whether decision or objective space niching is required. For objective space niching, d_{ij} is evaluated using a simple Euclidian distance calculation of the two objectives (*obj1* and *obj2*) for chromosomes i and j.

$$d_{ij} = \sqrt{((obj1_i - obj1_j)^2 + (obj2_i - obj2_j)^2)} \qquad \text{Equation 5.5}$$

When niching in decision space, the rmsd between the poses that are encoded by chromosomes i and j could be used.

$$d_{ij} = \sqrt{(1/n \sum r_{ij}^2)} \qquad \text{Equation 5.6}$$

where n is the number of atoms of the ligand and r_{ij} is the distance between two corresponding atoms encoded by chromosomes i and j in 3D space. In the NSGA-II, objective space niching, through trial and error, was found to be more effective than decision space niching and this is implemented throughout this work.

5.2.4.1 Selection of niched chromosomes from within a rank for crossover

The niched fitness values provide a discriminating measure to select a single chromosome from one rank for crossover. This is done using another roulette wheel selection procedure, this time the segments represent chromosomes within a rank, and the segments' sizes are proportional to the chromosomes' niched fitness values. This biases the search towards chromosomes which are less crowded relative to the rest of the rank's chromosomes, or to the outliers, and this helps retain population diversity. No scaling of the niched fitness values was carried out at this point. The roulette wheel is "spun" once, and returns the chromosome to which the genetic operator, cross-over, is applied.

5.2.5 The crossover operator

The techniques described above are repeated twice to select the two chromosomes needed for crossover. The single point crossover operation was described in the previous chapter and this is also applied here. Chromosomes resulting from this operation are checked for duplicates against all other chromosomes of the parent population and also against any chromosomes that have already been produced during the current generation. If either of the chromosomes produced by the crossover is found to be a duplicate then the pair of chromosomes is discarded and the whole process is repeated until a unique pair of chromosomes is found.

5.2.6 The mutation operator

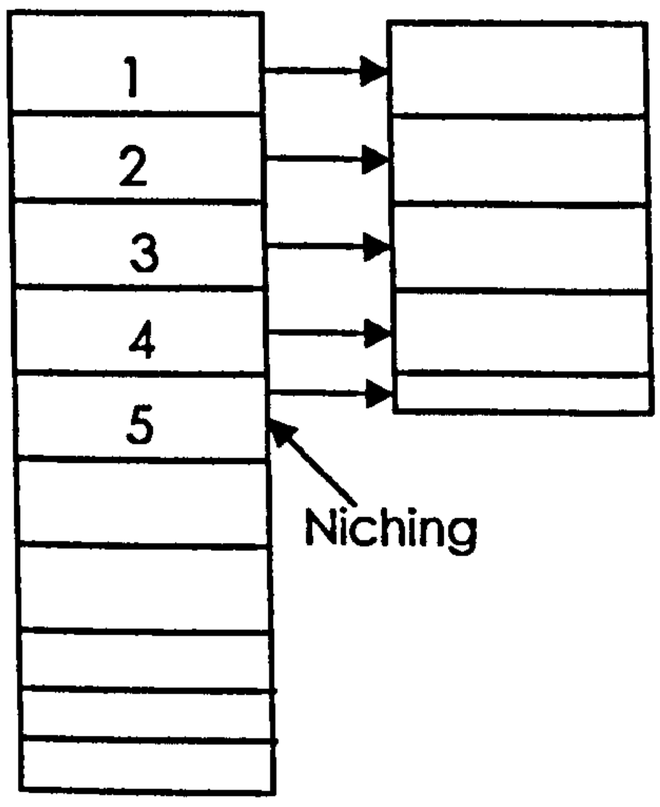
After the crossover operator has produced the intermediate population, the final genetic operator, the mutation operator, is applied. This is performed in exactly the same way as described in section 4.3.3. The duplicate checking function is also used here- if a mutation causes a chromosome to become identical to another chromosome in the parent or intermediate population, then that mutation is rejected, and the

mutation process is repeated (with different random numbers) until a unique chromosome is created.

5.2.7 From intermediate to offspring population

The highly elitist nature of NSGA-II only allows chromosomes into the offspring population if they show fitness values higher than any parent chromosome. Therefore the chromosomes in the intermediate population must be compared to the parent population. Firstly the objectives of the intermediate population's chromosomes are calculated. This is followed by combining the parent and intermediate population together. The combined population is Pareto ranked and sorted in ascending order. The final offspring population is formed by taking the top N chromosomes of the combined population (where N is the population size) to form the offspring population. The situation may arise though, where the number of chromosomes in the final rank of the offspring population "overflows" or exceeds the number of positions available. In this case the niching function is applied to the final rank's chromosomes, and the roulette wheel selection based on the resulting niched fitness values selects the final chromosomes of the offspring population (Figure 5.5).

This entire process is repeated with the offspring population now behaving as the parent population.



Combined population

Offspring population

Figure 5-5 Creating the offspring population (overleaf). All chromosomes within a rank of the combined population are passed over to the offspring population. But if the final rank contains more chromosomes than the offspring population size (e.g. rank 5 in figure), then niching in the rank allows for the selection of the chromosomes needed to complete the offspring population.

5.2.8 Termination criteria

Monitoring progress in an SGA, and hence deciding when to terminate a run is fairly straightforward, as the main requirement is to keep track of only one solution's status (the top-ranked solution). However implementing a termination criterion in a multiobjective optimisation algorithm would require a different strategy since it is the progress of the Pareto front, in multi-dimensions, that needs to be observed. One way of doing this is to check the current set of Pareto solutions against an archive containing solutions which have been generated a few generations ago. If none of the current set dominates the previous set, the run is terminated. However, this approach does not guarantee convergence to the true Pareto front since the two populations could be mapping different regions of the Pareto front. Establishing effective termination criterion in multiobjective optimisation is a well recognised problem and many implementations are based on running for a fixed number of generations determined through experimentation. This is the approach adopted here (Laumanns *et al.*, 2002, Roudenko *et al.*, 2004).

With the NSGA-II, the algorithm is run for a fixed number of generations, after which the program terminates. The determination of the fixed number of generations was achieved through trial and error. The algorithm was run for a very large number of generations on a few test cases. The number of generations was gradually reduced until the optimum number of generations, which finds the most advanced Pareto front was reached. With multiobjective optimisation methods it is difficult to ensure when the most advanced, or "true" Pareto front, has been reached. In the case of the NSGA-II, this was verified when solutions with low rmsds (of less than 2.0 Å) were obtained within the Pareto front. Solutions which are as close to the crystal structure as

possible are deemed as being "correct". Also, as the ensuing results chapters show, the performance of the NSGA-II is compared to that of Q-fit, a published docking algorithm which uses the total interaction energy, or a single objectivc, to guide thc optimisation process. For the majority of the results shown, Q-fit's top ranked solutions (those with the lowest total interaction energy), are among the Pareto solutions at the Pareto front. This is a good indication that the Pareto front has advanced optimally, i.e. the true Pareto front has been reached since the Pareto solutions also include optimal solutions from a different algorithms. Some cases are also presented in the following results chapters where the Q-fit solution has lower energy than any of the Pareto solutions. This indicates that the Pareto solutions have not advanced far enough, and for these cases the NSGA-II was run for a longer number of generations in an attempt to reach true convergence. If this still does not advance the Pareto front's position, it is considered that the Pareto front has converged to a local minimum.

Currently, for the purpose of algorithm development the number of generations for which to run the algorithm is specified manually. The NSGA-II was tested on two datasets, the results of which are discussed in the following chapter.

5.3 Chapter summary

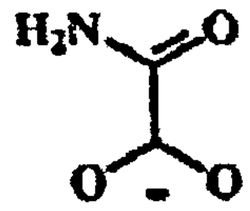
In this chapter thc conversion of a single-objectivc SGA to a multiobjcctivc NSGA-II was described. The new features which have been introduced, some of which are integral to multiobjective optimisation, are the scoring of each chromosome by two objectives rather than one, Pareto ranking of the population, a fitness-sharing niching technique and a highly elitist structure to the algorithm. In the next chapter, the NSGA-II is tested on two datasets, one of which is Dataset 1 which was used to validate the SGA in the previous chapter. Description of thc datascts, thc cxccution of the NSGA-II on this data, along with the results obtained are all discussed in the next chapter.

6 Initial Results of NSGA-II

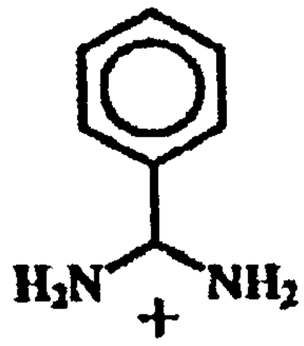
In this chapter, the performance of the NSGA-II is tested. The performance of the NSGA-II will depend on its ability to obtain solutions, within its Pareto set, which have rmsds that are 2.0 Å or less from the crystal structure. The 2.0 Å threshold is widely used in the field to judge the performance of a docking algorithm- poses within this threshold are assumed to be making similar interactions to the co-crystallised ligand, and the pose prediction technique is therefore assumed to be relatively accurate. Also any results obtained from testing the NSGA-II can be used to assess whether a multiobjective optimisation approach has an advantage over single objective optimisation, and whether it provides information that cannot be obtained from single objective optimisation.

The test set selected for validating the NSGA-II is that used in Jackson (2002), which consists of twenty proteins in complex with rigid-body ligands/ligand fragments. As the NSGA-II performs rigid-body docking, a rigid-body test set is ideal for initial proof-of-concept tests. This dataset represents a diverse set of fragments which bind non-covalently to their receptors and have few or no rotatable bonds (Jackson, 2002). These were in turn selected from the original GOLD data set (Jones *et al.*, 1997), which consists of 100 protein-ligand complexes, with the ligands noted as being "interesting" and "drug-like". The complexes are placed in four categories (*good*, *close*, *errors*, *wrong*) - depending on the accuracy of their prediction by the GOLD docking algorithm. Complexes where the binding mode, including all hydrogen bond and metal coordination interactions and other close contacts between the protein and ligand were reproduced correctly were placed in the *good* category. When the correct pose was predicted, but with a few displaced ligand groups then the complex was placed in the *close* category. The *errors* category contained predictions that were partially correct but with significant errors, and the predictions that were completely incorrect were placed in the *wrong* category. Ten of the complexes in the Jackson dataset were taken from the *good* or *close* categories (referred as Dataset 1), and ten from the *errors* or *wrong* categories (Dataset 2). By subjecting a given algorithm to

complexes which are considered to be at varying levels of difficulty for docking, the algorithm's capabilities can be fully tested. Figures 6.1 and 6.2 give the chemical structures of the ligand/ligand fragments. Tables 6.1 and 6.2 give the PDB code of the complexes and their protein-ligand constituents.

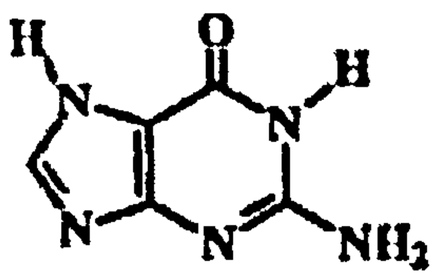


1ldm

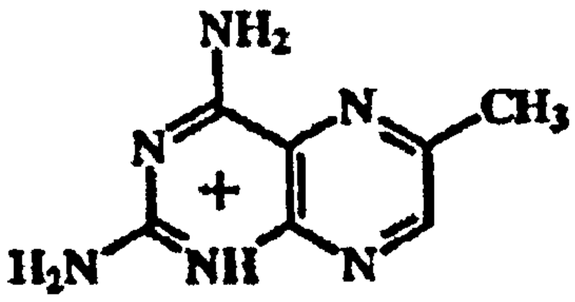


3ptb

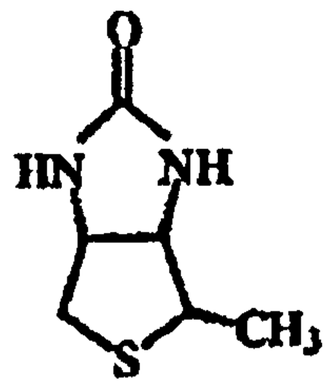
2phh



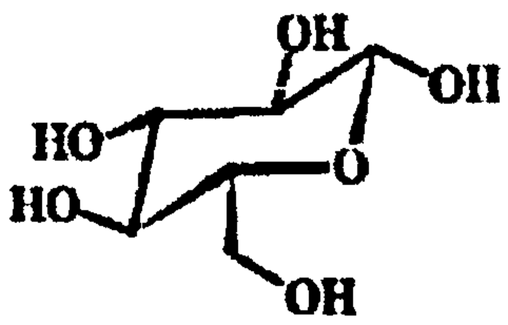
1ulb



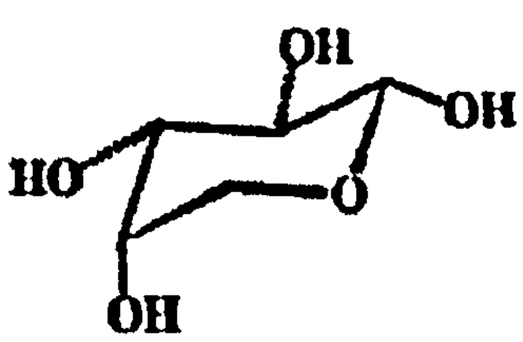
4dfr



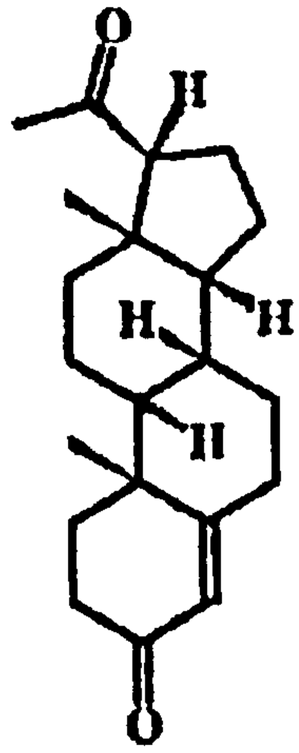
1stp



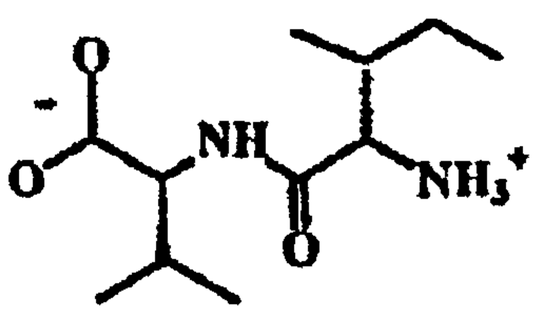
2gbp



1abe

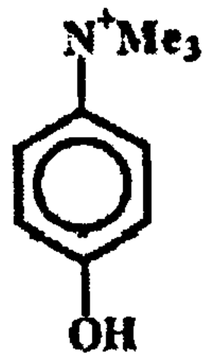


1dbb

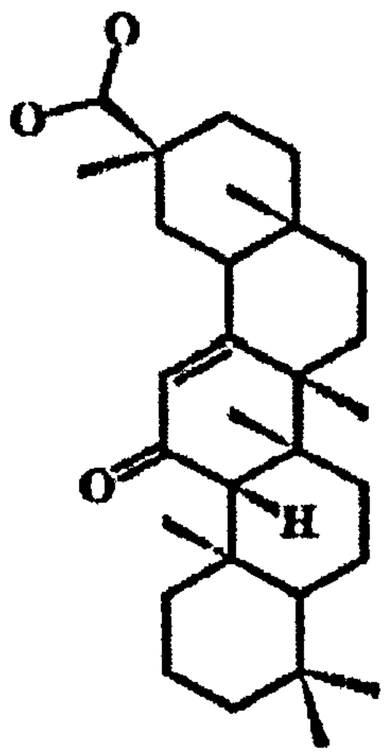


3tpi

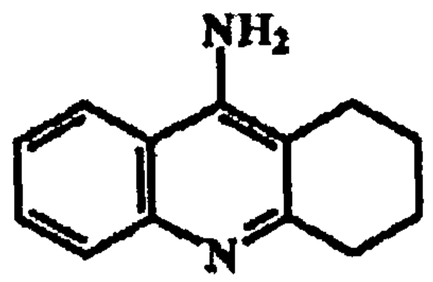
Figure 6-1 Molecular structures and PDB codes of ligands in Dataset 1



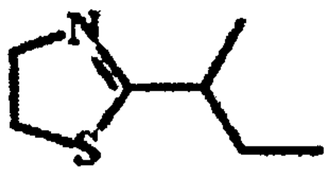
1ack



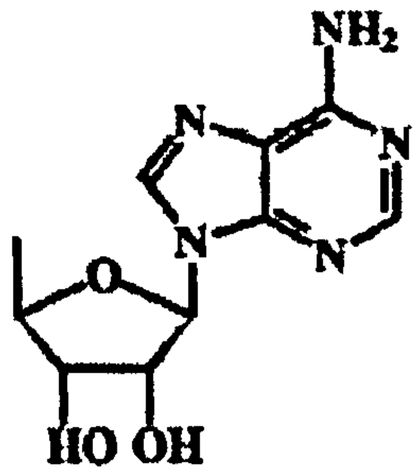
1hdc



1acj

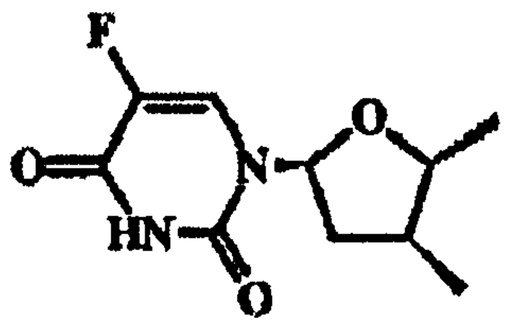


1mup

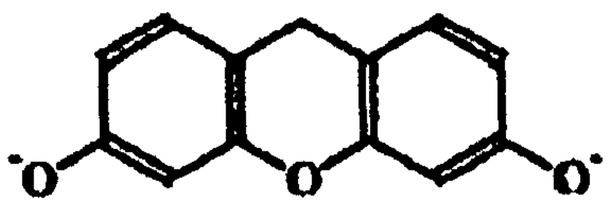


2ak3

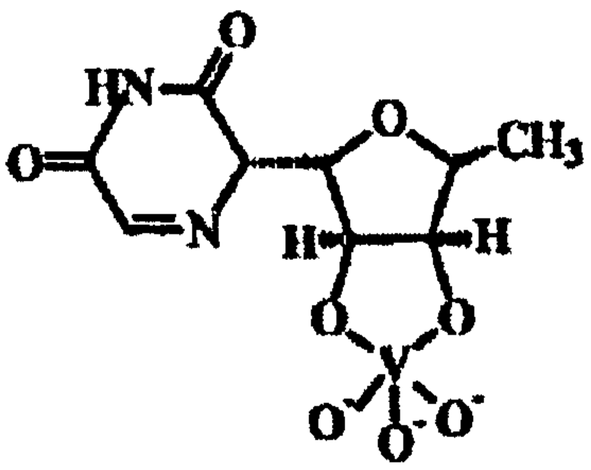
2mth



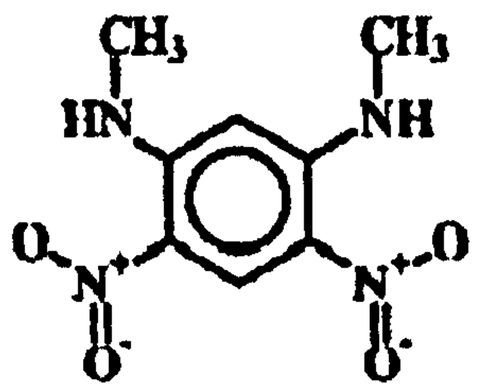
1tbd



4fab



6rsa



1baf

Figure 6-2 Molecular structures and PDB codes of ligands in Dataset 2

PDB code	Protein-ligand complex
3ptb	Beta-trypsin with benzamidine
1abe	L-arabinose-binding protein with l-arabinose
1ulb	Purine nucleoside phosphorylase with guanine
1stp	Streptavidin complex with biotin
1ldm	M=4= lactate dehydrogenase ternary complex with NAD and oxamate
2phh	P- Hydroxybenzoate hydroxylase with P-hydroxybenzoate
3tpi	Trypsinogen complex with ile-val
1dbb	Fab' fragment of the db3 anti-steroid monoclonal antibody with progesterone
4dfr	Dihydrofolate reductase with methotrexate
2gbp	D- Galactose D-glucose binding protein with Beta-d-glucose

Table 6.1 PDB codes and the proteins and ligands for complexes within Dataset 1

PDB code	Protein-ligand complex
1baf	Fab fragment of murine monoclonal antibody with tetramethyl dinitrophenyl
1acj	Acetylcholinesterase with tacrine
1ack	Acetylcholinesterase with edrophonium
1hdc	3-alpha, 20-beta-hydroxysteroid dehydrogenase with carbenoxolone
1mup	Major urinary protein complex with 2-(sec-butyl) thiazoline
1tdb	Major urinary protein complex with 2-(sec-butyl) thiazoline
2ak3	Adenylate kinase isoenzyme-3 (gtp: amp phosphotransferase)
2mth	Methylparaben insulin
4fab	fab fragment with fluorescein (dianion)
6rsa	Ribonuclease a complex with uridine vanadate

Table 6.2 PDB codes and the proteins and ligand for complexes within Dataset 2

6.1 Parameterisation of NSGA-II

The performance of an NSGA-II, as with an SGA, is controlled by several parameters. This includes all the parameters associated with an SGA (such as mutation rate, population size), and a few parameters specific to a NSGA-II. What follows is a description of the parameters.

6.1.1 Population size

The population consists of chromosomes representative of solutions in the search space. A large population will therefore have more genetic diversity than a smaller population (no duplicate chromosomes are permitted). A richer pool of genes increases the probability of good genes combining through crossover to obtain good solutions. (Also, with a larger population, a larger space in the GRID box is covered by chromosomes of the initial population. This speeds the process of reaching the favourable areas within the GRID box.) The downside of a large population is the increase in computation time resulting from applying the genetic operators, two-objective scoring and the Pareto ranking process to more chromosomes. Continually increasing the population size also ceases to improve results after a certain point. Therefore a balance between computational time and effective optimisation needs to be achieved. A multiobjective approach usually requires a larger population size than a single objective algorithm, because a group of optimal solutions (the Pareto solutions) are sought, rather than one single optimal solution. After several trial and errors, a population size of 200 was seen as the most appropriate.

6.1.2 Mutation rate

Three parameters are involved in mutation; the mutation rate, and the two step sizes (one for mutating the translation genes and one for the rotation genes). A high mutation rate allows the search to proceed randomly whereas with a low mutation

rate the NSGA-II might fail to discover unexplored areas in the search space. Similarly too large a step size introduces random changes in the genes whereas a very small step size may have little effect. Ultimately a balance needs to be reached. The optimum values found for this NSGA-II's purposes were 2.0 Å for the translational step-size and 360° for the rotational step-size.

6.1.3 *Fmult* and σ_{share}

The scaling procedure that was implemented in the SGA (section 4.3.1.1) has also been applied here. During multiobjective optimisation, chromosome selection is dependent on the Pareto ranks of the chromosomes, rather than their raw fitness values. The Pareto ranks are used in the selection procedure (in a process analogous to rank selection in an SGA) and are therefore scaled, using Goldberg's linear scaling procedure described in Chapter 4. (Rank selection involves ranking a population by its chromosomes' raw fitness values, and using a chromosome's position (or rank), as its representative fitness value for the selection operator.). The f_{mult} parameter was not modified for the NSGA-II, since the 1.8 value used for the SGA was seen to be just as effective. The reason for applying the scaling procedure here is because there is the possibility that the highest ranks could be favoured in the selection procedure, which lead to loss of diversity and premature convergence. This is especially an issue if there is a small number of Pareto ranks and when the top rank occupies a disproportionately large segment of the roulette wheel.

The σ_{share} value, the niche radius described in the previous chapter, has been set at 1.0.

Mutation rate	20%
Rotation step-size	360 degrees
Translation step-size	2.0 Å
Population size	200
Generation number	100 000
Niche radius (objective space)	1.0

Table 6.3 NSGA-II parameters when NSGA-II is tested with datasets 1 and 2 (all molecules are considered to be rigid) (overleaf)

6.2 Comparison to Q-fit

As was discussed in Chapter 4, the Q-fit docking tool uses the GRID scoring function to assess the quality of the poses generated during the search procedure. As the NSGA-II implements the same scoring function, it is feasible to compare the performance of the NSGA-II against that of Q-fit. The main difference between the algorithm is the search procedure rather than the scoring function, so it should therefore be possible to attribute differences in results between the algorithms to the number of objectives used to guide the search. Also comparing the results from these programs may highlight information which multiobjective optimisation provides that single objective optimisation does not. Also, by comparing the rmsds and energies of solutions from both algorithms, it may be possible to infer whether the global optimum has been reached, assuming the results from both algorithms are in agreement. Ultimately the Q-fit search procedure is led by the total interaction energy- and is optimised by a single objective, whereas the NSGA-II search procedure is guided by multiple independent objectives; this is a process which has not been, as far as we are aware, applied to protein-ligand docking, and by comparing these two methods directly it will be possible to observe whether multiobjective optimisation offers any advantages relative to single objective optimisation in terms of finding correct energies, and learning about the influence of the different objectives on a given complex.

6.3 Interpretation of Pareto plots

The most effective way of presenting the results obtained by the NSGA-II is to show Pareto fronts, containing Pareto solutions, and that have been plotted in objective space. These Pareto solutions will be shown as plots where the electrostatic and hydrogen bond interactions are on the x-axis, and the vdw interactions are on the y-axis. By retrospectively observing where correct solutions (i.e. solutions of rmsds of less than 2.0 Å) fall on the Pareto front, it is possible to learn which of the objectives, if any, are the most influential. The NSGA-II can potentially find several Pareto solutions, all of which are different poses of the ligand. In order to facilitate the

comprehension of these results, the Pareto solutions are clustered based on their rmsd values from the crystal structure of the ligand (this process is explained in more detail in the next section).

Figure 6.3 is a hypothetical example of a plot which shows that the correct Pareto solutions (those within 2.0 Å of the crystal structure) are predominantly influenced by electrostatic and hydrogen bond interactions. This is apparent by the relatively more favourable magnitudes of these interactions, which are in the region of -30 to -45 kcal/mol. Although the vdw interactions do play a role in finding these correct solutions, they are not as influential as the electrostatic and hydrogen bond interactions.

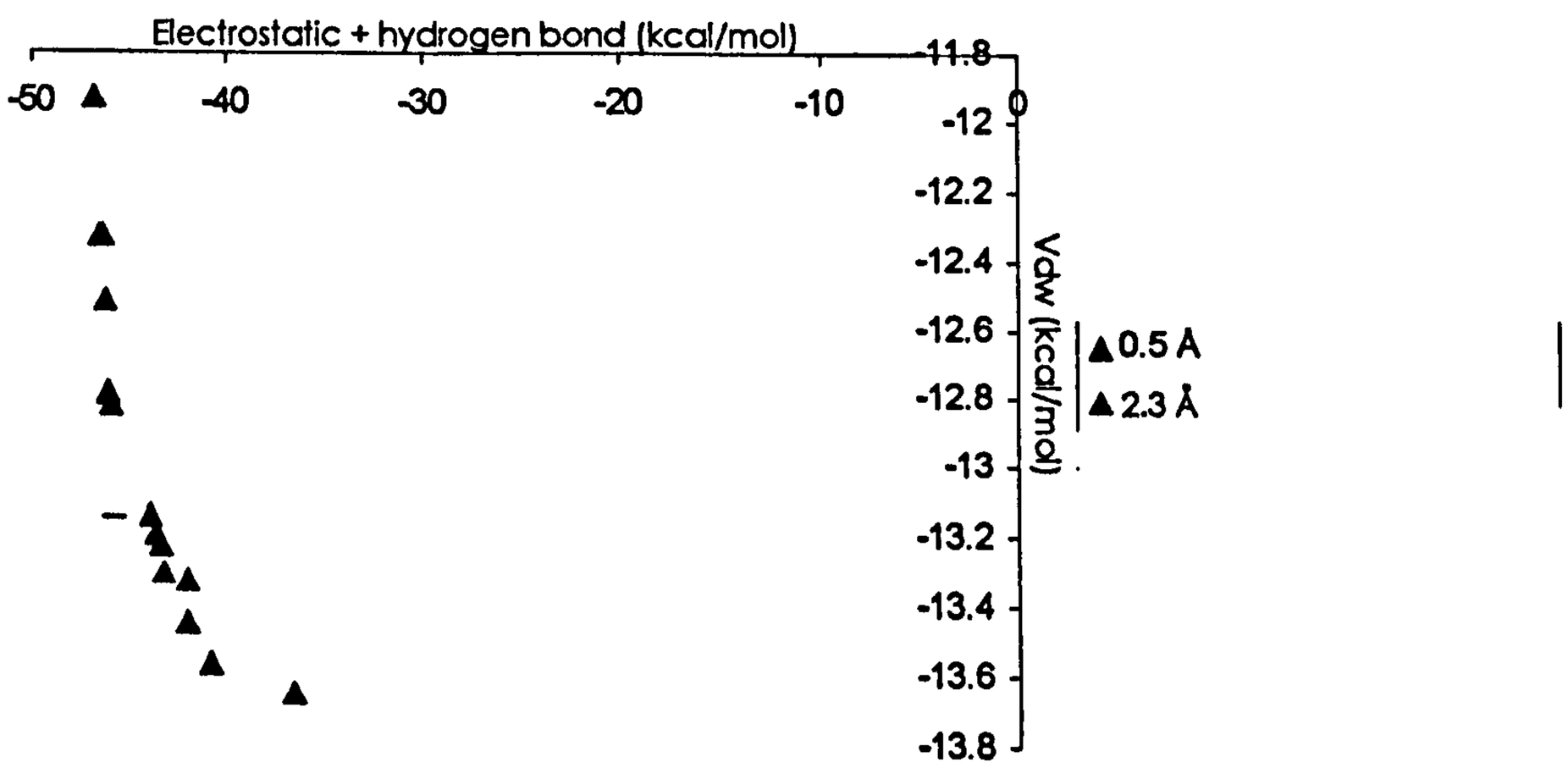


Figure 6-3 Pareto solutions predominantly dominated by electrostatic and hydrogen bond interactions in objective space

This next plot (Figure 6.4), shows several Pareto solutions which have been clustered into four different groups. Only one of these groups, cluster 0.4 Å, contains correct solutions below 2.0 Å. The position, in objective space, of solutions in this cluster shows that these are more influenced by vdw interactions. As the figure shows, the vdw interactions for these solutions are more favourable than the electrostatic and hydrogen bond interactions.

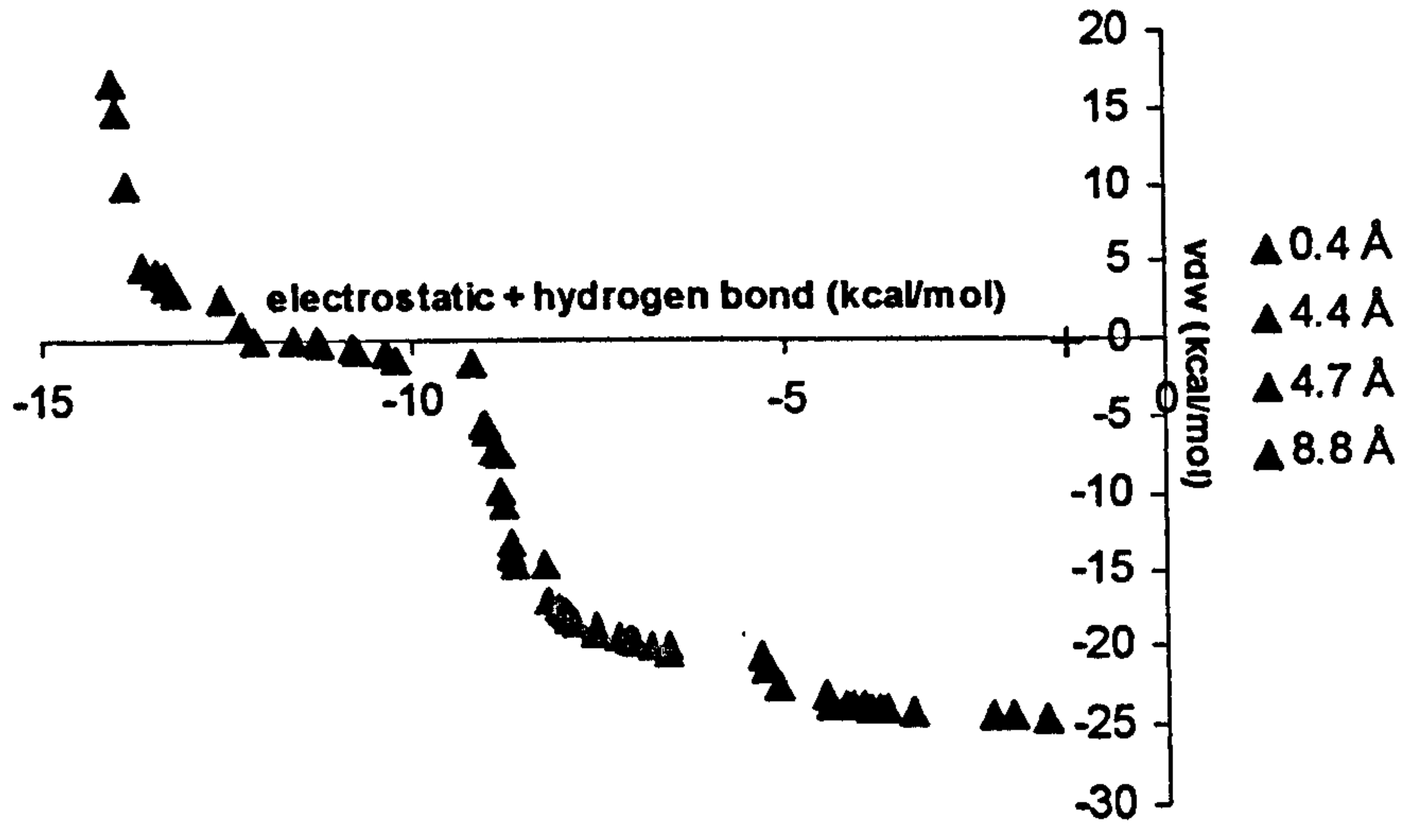


Figure 6-4 Pareto front where correct solutions are dominated by vdw interactions

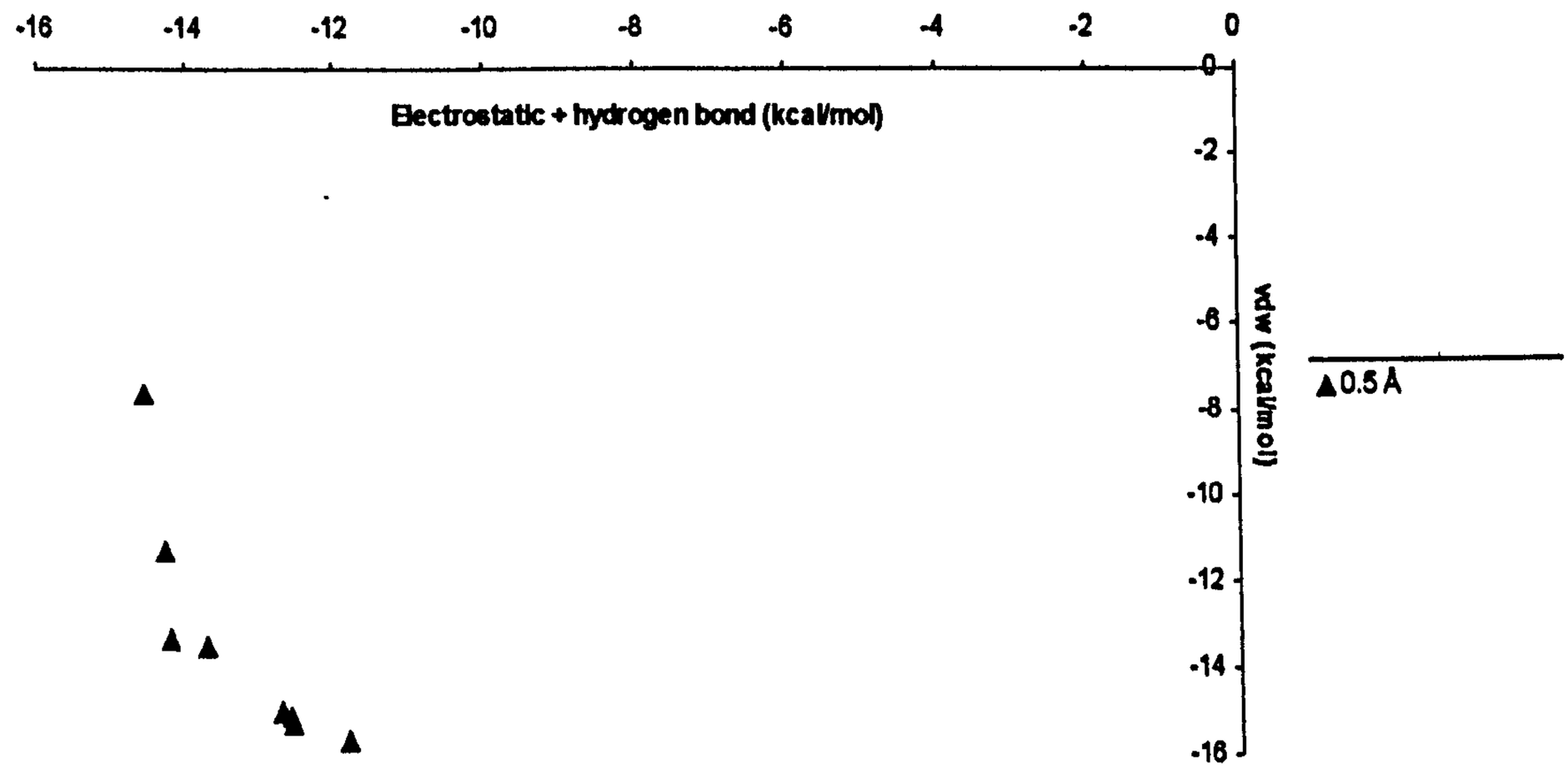


Figure 6-5 Pareto solutions which are relatively equally influenced by both objectives

Looking at Figure 6.5, it can be observed that the correct cluster is relatively equally influenced by both objectives. This is unlike the previous two plots, where one of the objectives is seen as relatively more favourable than the other. It is therefore possible from this plot to infer that neither of the objectives has a predominating influence in obtaining correct solutions when docking this particular case.

6.4 Robustness of algorithm

As with any stochastic procedure, the robustness of a multiobjective algorithm must also be investigated. The robustness of an algorithm specifies whether the optimisation process is consistent, and that the optimal solutions found at the end of a run have not been reached randomly. The NSGA-II's robustness can be tested by running the algorithm using different random number generator seeds, and comparing the Pareto fronts generated by the different runs. C's srand() function was used to seed the different runs with different integers. The test for robustness was carried out on different test cases, all of which confirmed the algorithm's robustness. One of these tests is represented in Figure 6.6 which shows the Pareto fronts obtained when docking 1ack, a complex from Dataset 2. As the figure shows, the three different Pareto fronts produced by the different runs seeded with 8, 2 and 1 have all converged to approximately the same areas in objective space. This confirms that the algorithm is robust enough to obtain similar solutions when the random numbers generated at all points in the run are different.

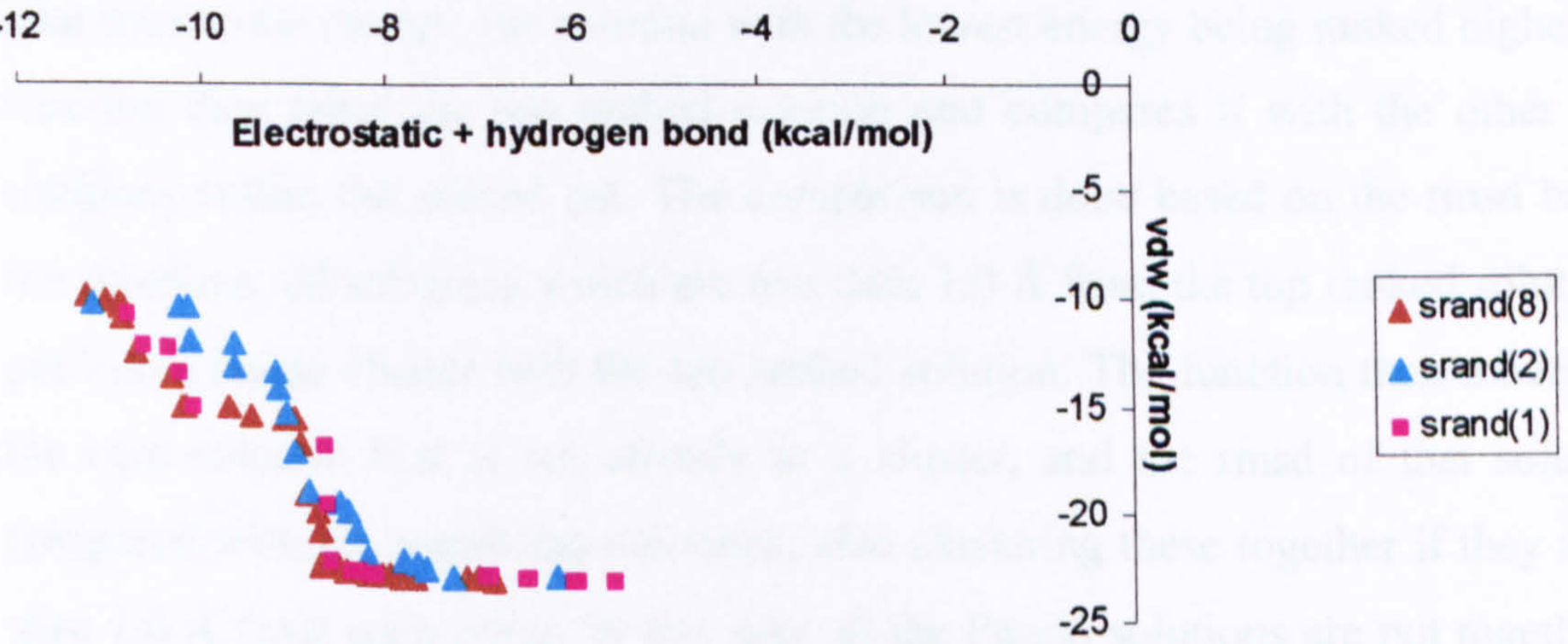


Figure 6-6 Pareto fronts obtained when NSGA-II was seeded with different integers. The test case used is 1ack.

6.5 Dataset 1 results

Using the parameters which have been discussed in the above section, the NSGA-II was tested on the 10 complexes from ADataset 1. The following figures show the

Pareto solutions obtained when each molecule was docked back into its co-crystallised protein active site. For comparison, results obtained from Q-fit are also shown. To obtain the energy terms (the two objectives) for Q-fit, the coordinates of the Q-fit solution are input into the NSGA-II which then scores the ligand pose and returns the values of the two objectives. The solution is plotted in objective space along with the corresponding Pareto solutions. In cases where the rmsd of the top-ranked Q-fit solution is below 2.0 Å then only the top-ranked Q-fit solution is shown on the plot. If the rmsd of the top-ranked solution is greater than 2.0 Å then that solution is plotted, along with the next solution within the ranked list that has an rmsd below 2.0 Å. The top-ranked Q-fit solutions are labelled as "top-rank" whereas the low rmsd Q-fit solutions are labelled as "best rmsd". As has been mentioned previously, multiobjective optimisation produces a number of solutions, each of which represents a unique pose. To be able to easily comprehend the results, and to simplify the process of observing these solutions in objective space, the Pareto solutions are clustered into groups based on their rmsds from the crystal structure using a first-pass clustering procedure. This function first sums the two objectives of each pose in the Pareto solution set. These solutions are then ranked based on their total interaction energy, the solution with the lowest energy being ranked highest. The function then takes the top ranked solution and compares it with the other Pareto solutions within the ranked list. The comparison is done based on the rmsd between the solutions; all solutions which are less than 1.0 Å from the top ranked solution are put into a single cluster with the top ranked solution. The function then moves on to the next solution that is not already in a cluster, and the rmsd of that solution is compared with the remaining solutions, also clustering these together if they are less than 1.0 Å from each other. In this way all the Pareto solutions are put together into clusters based on their orientations.

6.5.1 1abe

As figure 6.7 shows, for 1abe both Q-fit and the NSGA-II docked the molecule correctly. The top ranked Q-fit solution has an rmsd of 0.6 Å and, judging by its position in objective space, is influenced by electrostatic and hydrogen bond interactions more so than by the vdw interactions. The majority of the solutions in the Pareto set have an approximate rmsd of 0.3 Å. Comparing the positions of the top-ranked Q-fit solution and the Pareto solutions, it can be seen that the electrostatic and hydrogen bond energies of the Q-fit solution are more negative, and are therefore more optimised, than any of the Pareto solutions. In terms of the vdw interactions, some of the Pareto solutions have lower energies than the Q-fit solution, though the same Pareto solutions have higher electrostatic and hydrogen bond energies. The slightly lower electrostatic and hydrogen bond energies of the Q-fit solution indicates that such solutions can exist, and can be found by an algorithm's search procedure. As the NSGA-II did not find solutions which are as low implies that the Pareto front has not converged completely in terms of the electrostatic and hydrogen bond energies. Nevertheless this difference in electrostatic and hydrogen bond energies between the Q-fit and closest Pareto solution is relatively small – approximately 0.7 kcal/mol and the Q-fit solution only dominates two of the Pareto solutions- the remaining Pareto solutions are not dominated by this solution. It can therefore be concluded that the NSGA-II was successful at docking this complex.

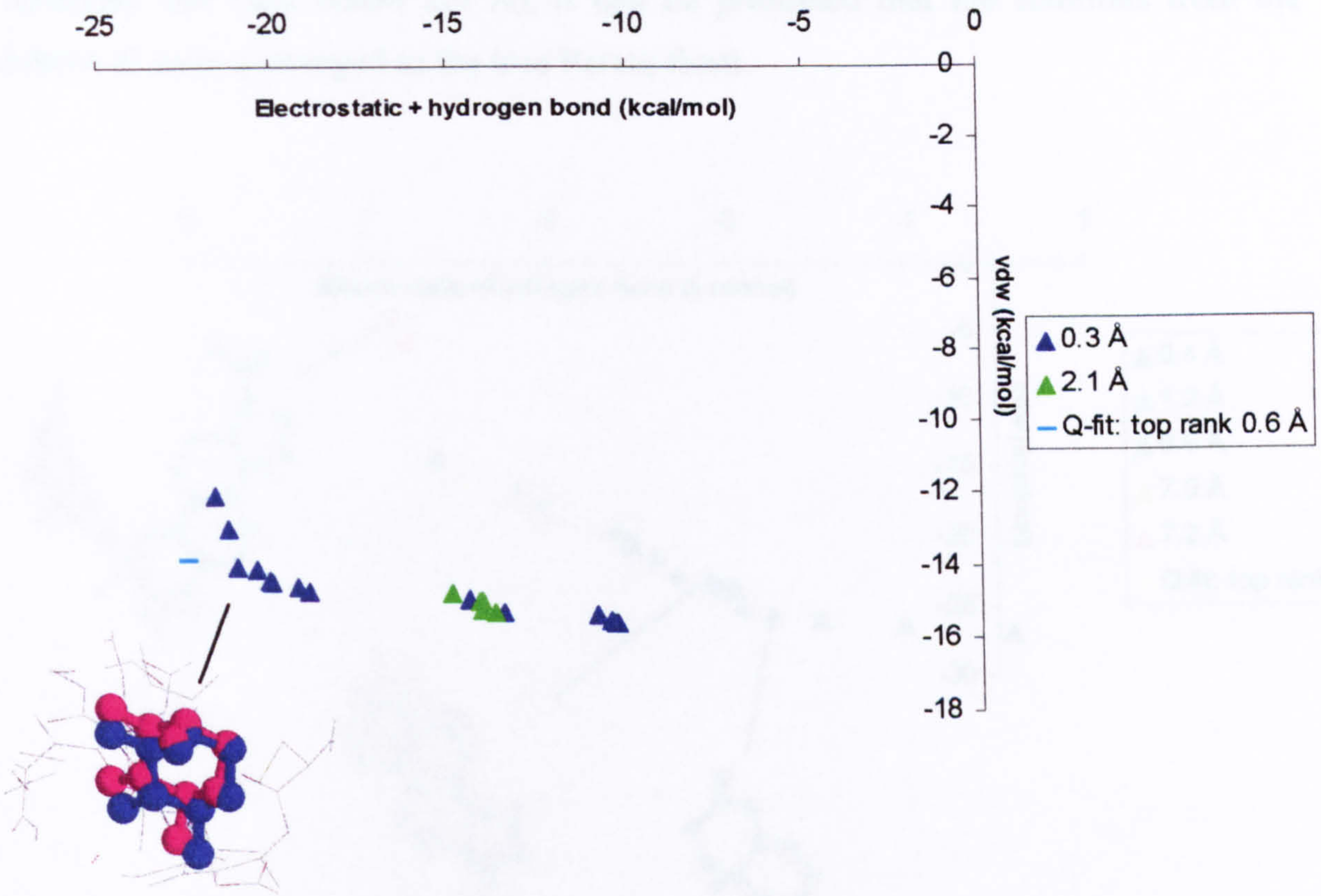


Figure 6-7 Pareto solutions obtained when docking 1abe. The top-ranked Q-fit solution is also shown.

6.5.2 1dbb

The NSGA-II produced a larger number of Pareto solutions for 1dbb than 1abe (Figure 6.8). These are also more varied in orientation, which is reflected in the number of clusters and the variety of approximate rmsds from the crystal structure. The top-ranked Q-fit solution is on the Pareto front, among Pareto solutions with low rmsds. The rmsd of the Q-fit solution is 1.2 Å, which is higher than some of the Pareto solutions which have approximate rmsds of 0.4 Å. The solutions within this cluster are spread across the Pareto front, and are varied in terms of the two objectives. The Pareto clusters with rmsds above 2.0 Å (the green, orange and red triangles) in general have more optimised electrostatic and hydrogen bond energies, but higher vdw energies. Since the top-ranked Q-fit solution is on the Pareto front, and because the rmsds of the lowest energy solutions from both algorithms are

relatively low (and below 2.0 Å), it can be presumed that the solutions from the NSGA-II have converged to the true Pareto front.

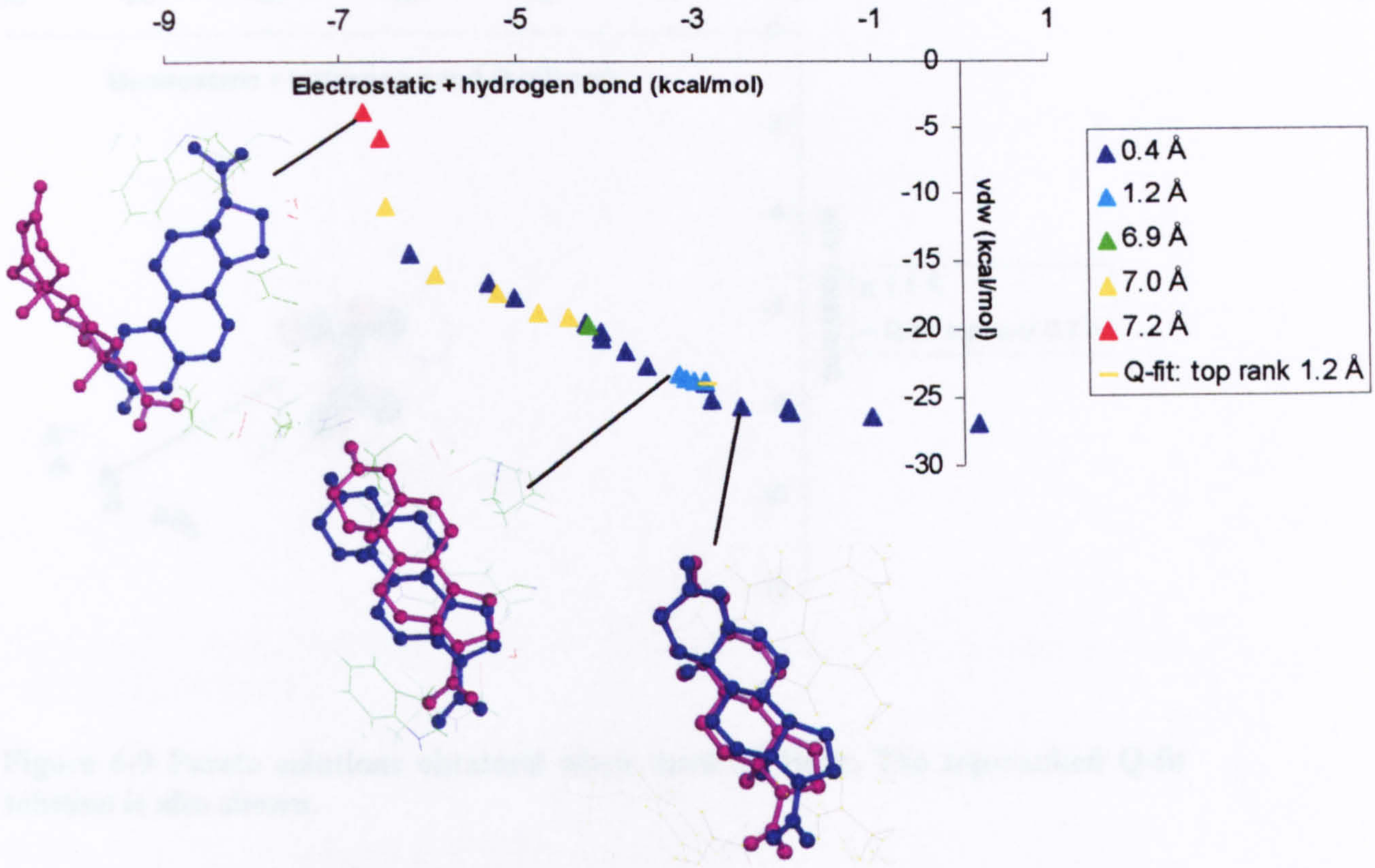


Figure 6-8 Pareto solutions obtained when docking 1dbb. The top-ranked Q-fit solution is also shown.

6.5.3 1ldm and 1stp

The Pareto solutions obtained when docking 1ldm all fall into a single cluster of an approximate rmsd of 1.1 Å (Figure 6.9). These solutions are more influenced by the electrostatic and hydrogen bond energies than by the vdw interactions. The top-ranked Q-fit solution is among the Pareto solution set, and has an rmsd of 0.7 Å.

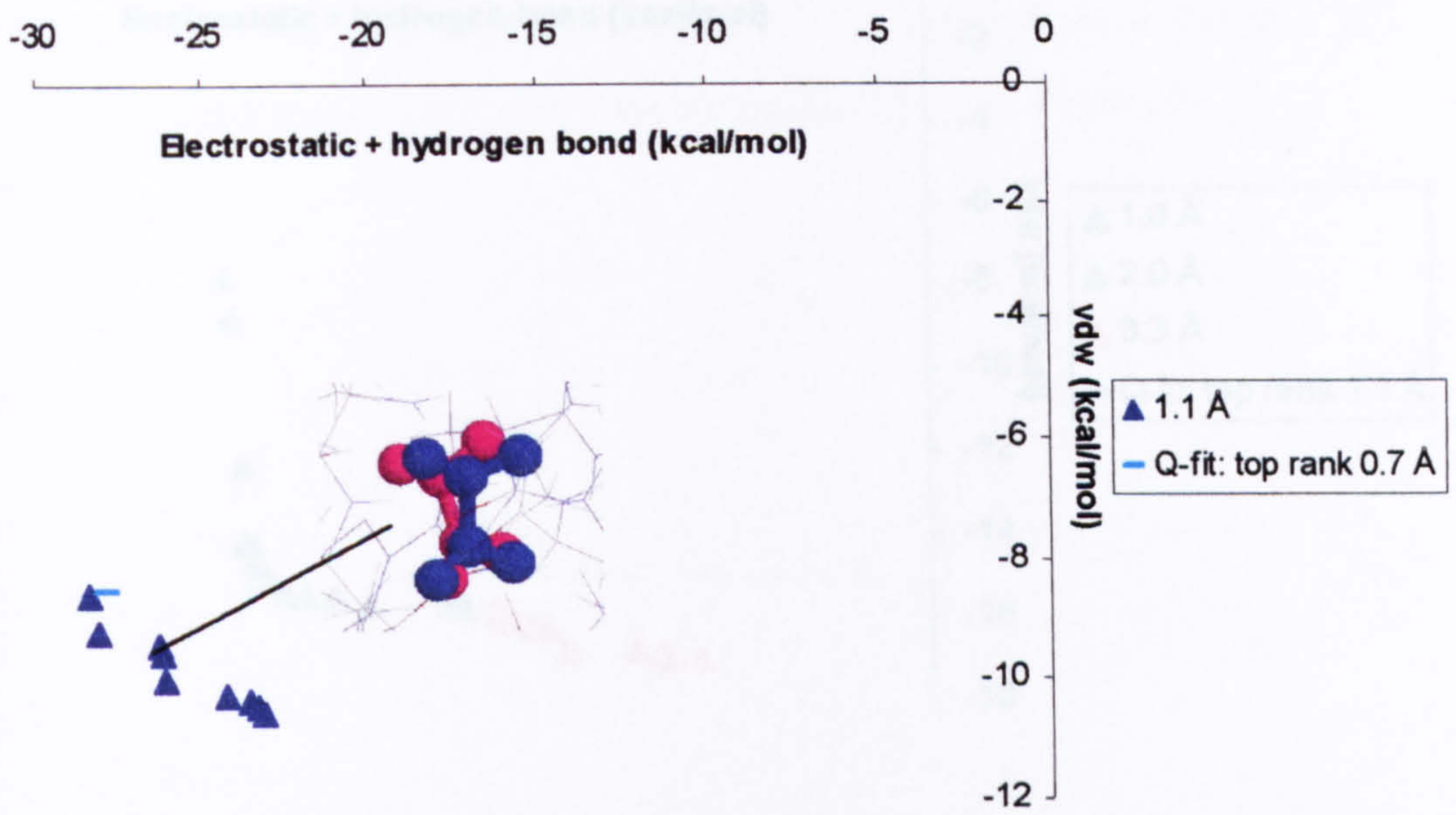


Figure 6-9 Pareto solutions obtained when docking 1ldm. The top-ranked Q-fit solution is also shown.

Unlike 1dbb, the Pareto clusters obtained when docking 1stp are not scattered across the Pareto front (Figure 6.10). The positions of the three clusters obtained fall into three clear groups in objective space. The group with the highest rmsd (3.3 Å) generally has low vdw energy but its electrostatic and hydrogen bond energies are higher. Conversely solutions in the 2.0 Å Pareto cluster have higher vdw energies but lower electrostatic and hydrogen bond energies. As the figure shows, The Pareto cluster with the lowest rmsd (1.0 Å) is in between the 2.0 Å and 3.0 Å clusters; its solutions have lower electrostatic and hydrogen bond energies than the 3.0 Å cluster, and also lower vdw energies than the 2.0 Å cluster. This cluster appears to be slightly more influenced by vdw energies than by electrostatic and hydrogen bond energies. The top-ranked Q-fit solution obtained when docking 1stp has an rmsd of 1.1 Å and is among the solutions of the 1.0 Å cluster.

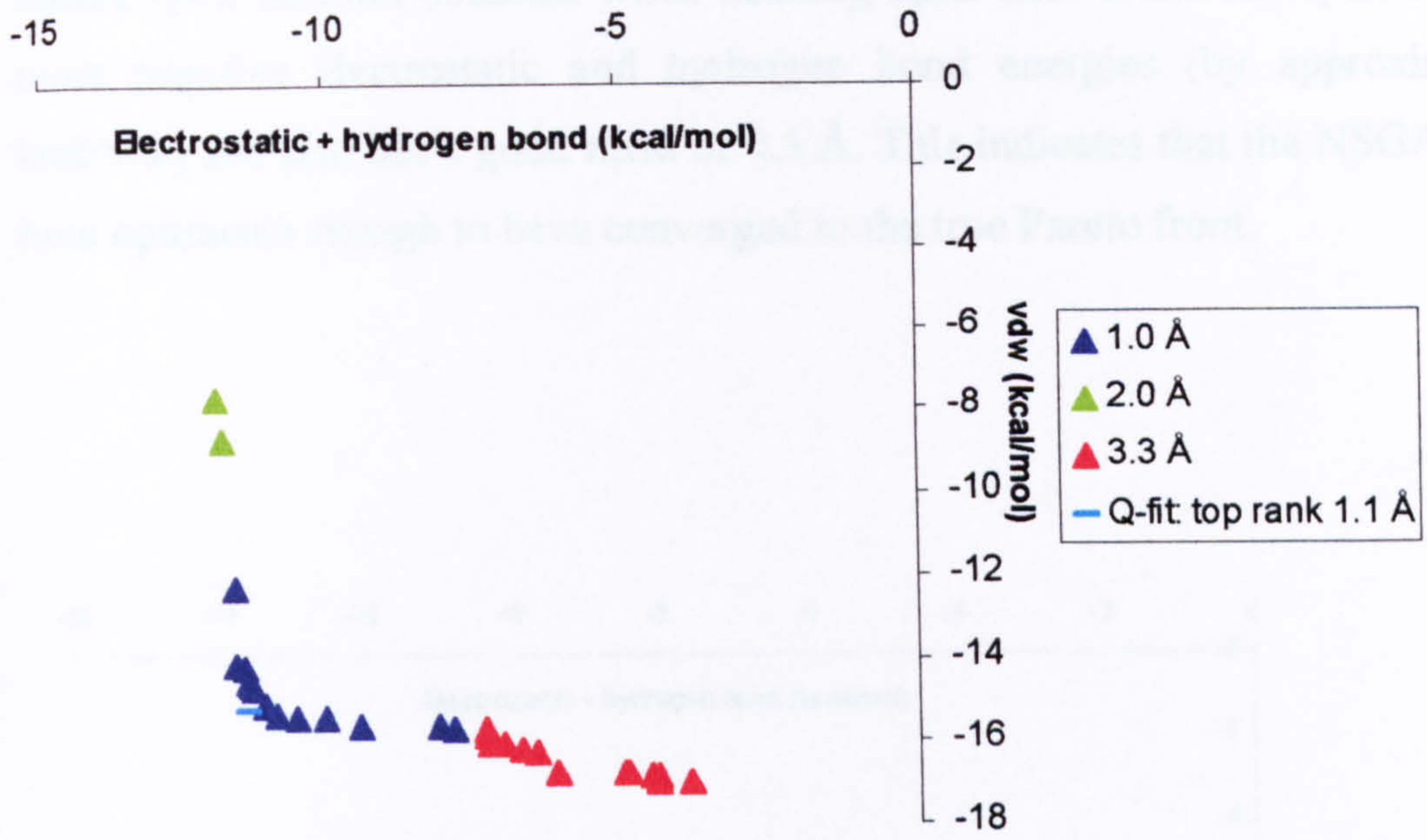


Figure 6-10: Pareto solutions obtained when docking 1stp. The top-ranked Q-fit solution is also shown

6.5.4 1ulb, 3tpi, 2gbp and 2phh

For 1ulb, 3tpi, 2gbp, and 2phh the NSGA-II obtained Pareto solutions which fall into single clusters (Figures 6.11, 6.12, 6.13, 6.14). The clusters produced by 1ulb, 3tpi and 2gbp are correct, and have rmsds of 0.5 Å, 1.0 Å and 0.6 Å respectively. In comparison, the 2phh Pareto solutions have high rmsds of approximately 4.5 Å. The top-ranked Q-fit solutions for 1ulb, 3tpi and 2gbp have rmsds which are below 2.0 Å. The position of the top-ranked Q-fit solution in objective space for 1ulb is on the Pareto front, among the Pareto solutions. This may indicate that, since both algorithms have found good solutions that are at the same points in objective space, the Pareto solutions obtained for 1ulb have converged to the true Pareto front. Figure 6.12 shows that for 3tpi, the Q-fit solution is slightly dominating the Pareto solutions, though these also have good rmsds. Regarding 2gbp, both algorithms produced solutions with good rmsds, but, as can be observed from Figure 6.13, the top-ranked Q-fit solution has slightly more negative electrostatic and hydrogen bond energies. This implies that the Pareto front has not fully optimised the electrostatic and

hydrogen bond energies. Comparing the Pareto solution set positions and the top-ranked Q-fit solution obtained when docking 2phh shows that the Q-fit solution has more negative electrostatic and hydrogen bond energies (by approximately 1.5 kcal/mol) and also has a good rmsd of 0.5 Å. This indicates that the NSGA-II has not been optimised enough to have converged to the true Pareto front.

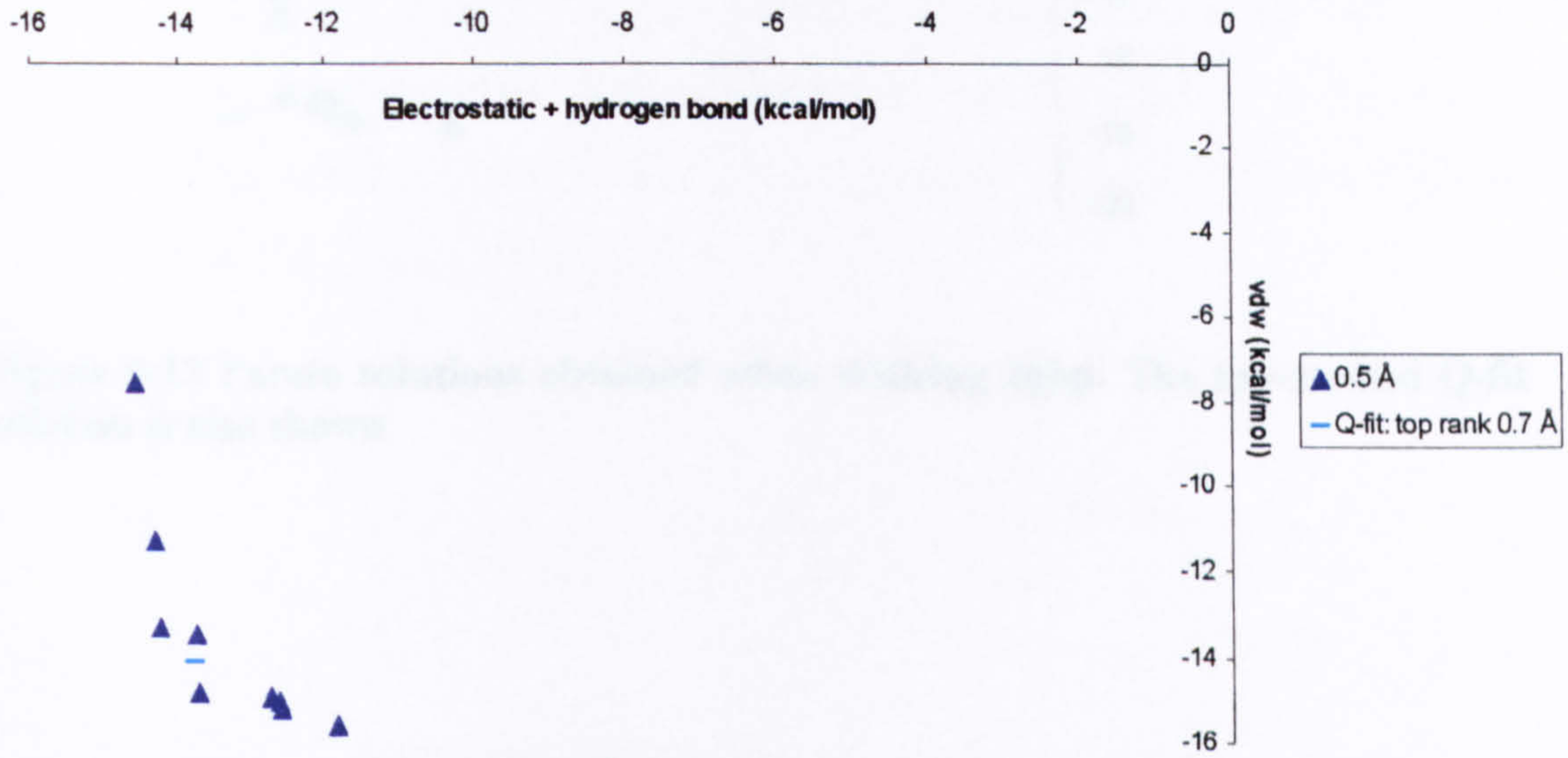


Figure 6-11 Pareto solutions obtained when docking 1ulb. The top-ranked Q-fit solution is also shown

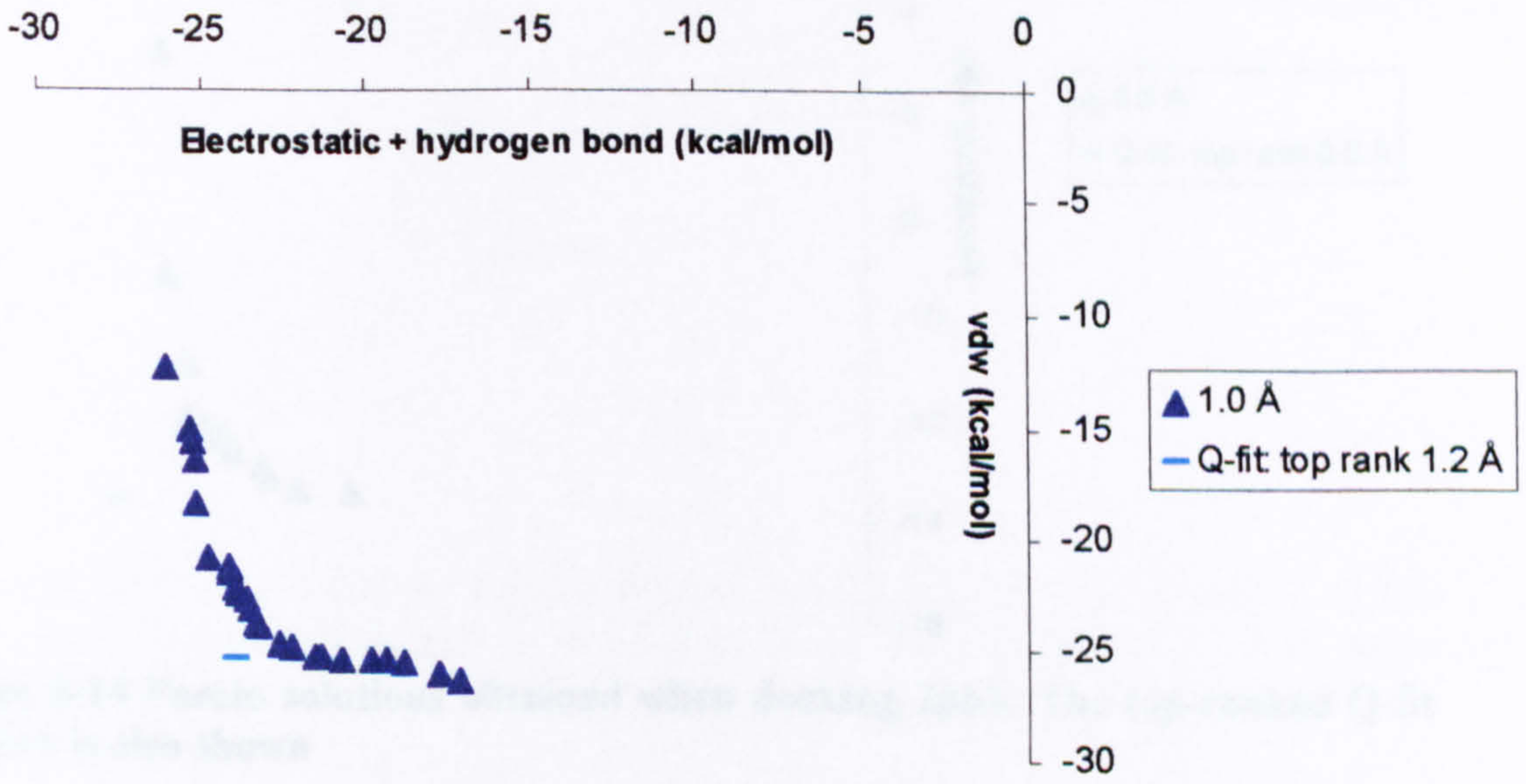


Figure 6-12 Pareto solutions obtained when docking 3tpi. The top-ranked Q-fit solution is also shown

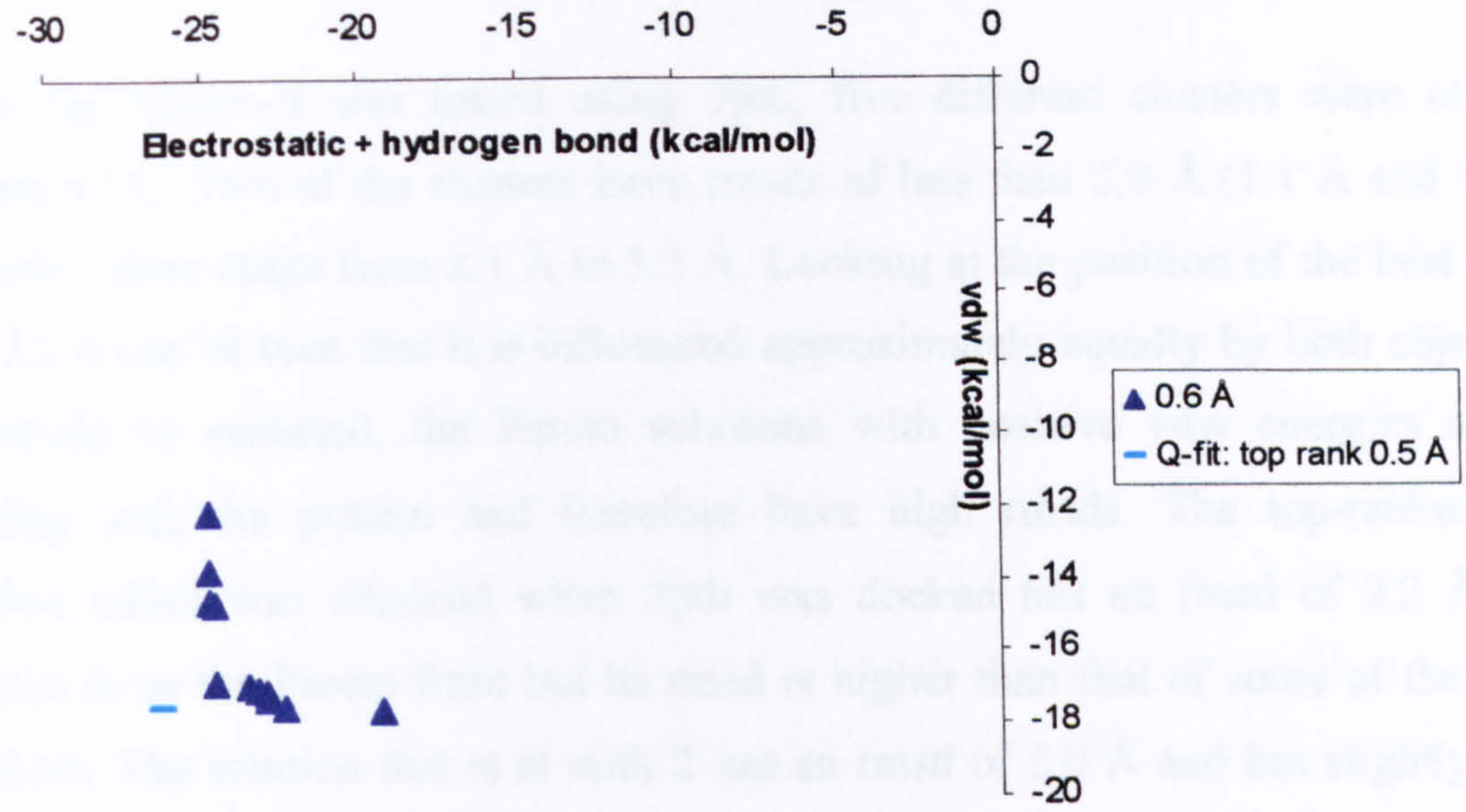


Figure 6-13 Pareto solutions obtained when docking 2gbp. The top-ranked Q-fit solution is also shown

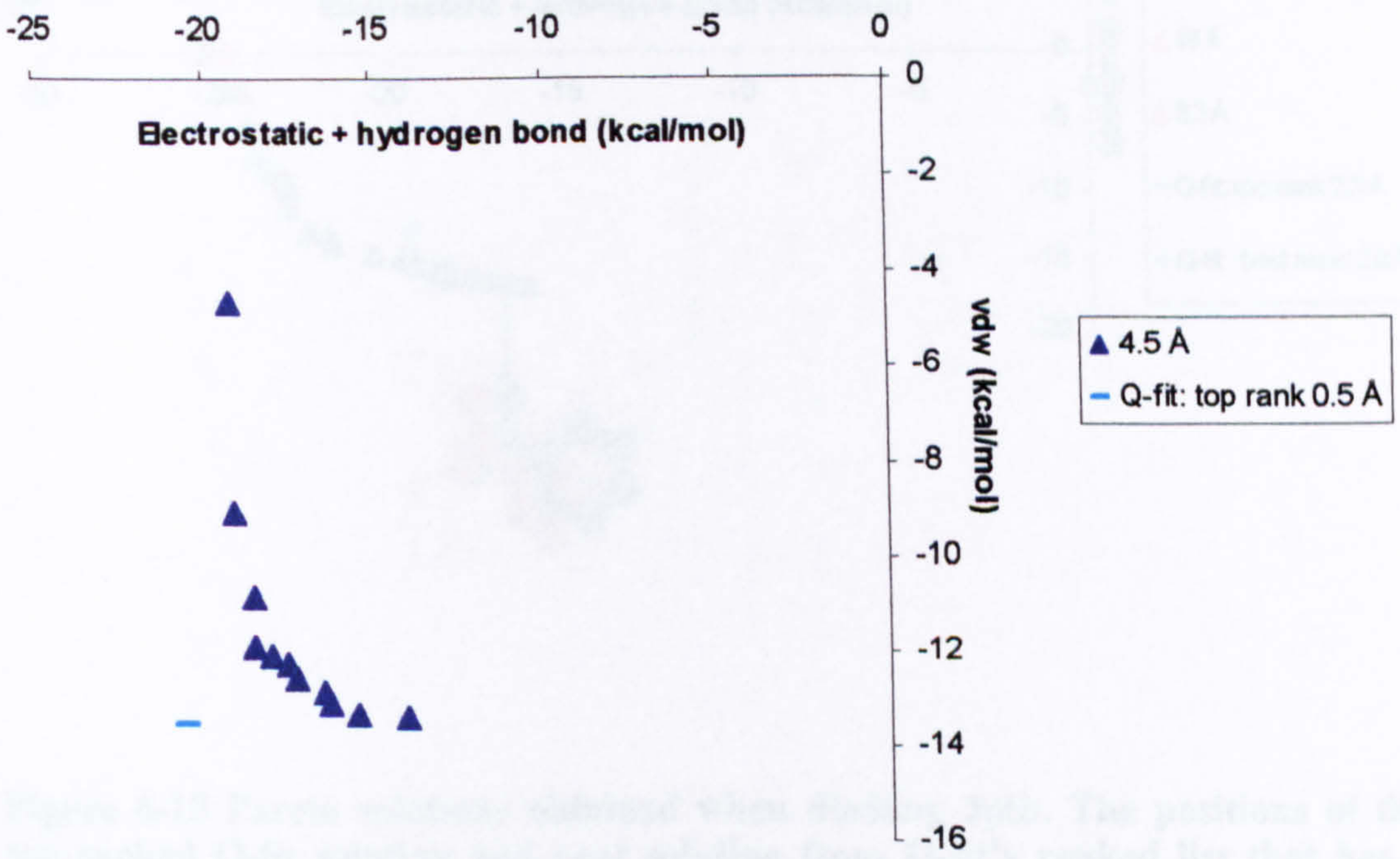


Figure 6-14 Pareto solutions obtained when docking 2phh. The top-ranked Q-fit solution is also shown

6.5.5 3ptb

When the NSGA-II was tested using 3ptb, five different clusters were obtained (Figure 6.15). Two of the clusters have rmsds of less than 2.0 Å (1.1 Å and 1.9 Å). The other three range from 2.1 Å to 3.3 Å. Looking at the position of the best cluster (1.1 Å), it can be seen that it is influenced approximately equally by both objectives. As would be expected, the Pareto solutions with positive vdw energies may be clashing with the protein and therefore have high rmsds. The top-ranked Q-fit solution which was obtained when 3ptb was docked has an rmsd of 2.2 Å. This solution is on the Pareto front but its rmsd is higher than that of some of the Pareto solutions. The solution that is at rank 2 has an rmsd of 2.0 Å and has slightly worse energies than the Pareto solutions.

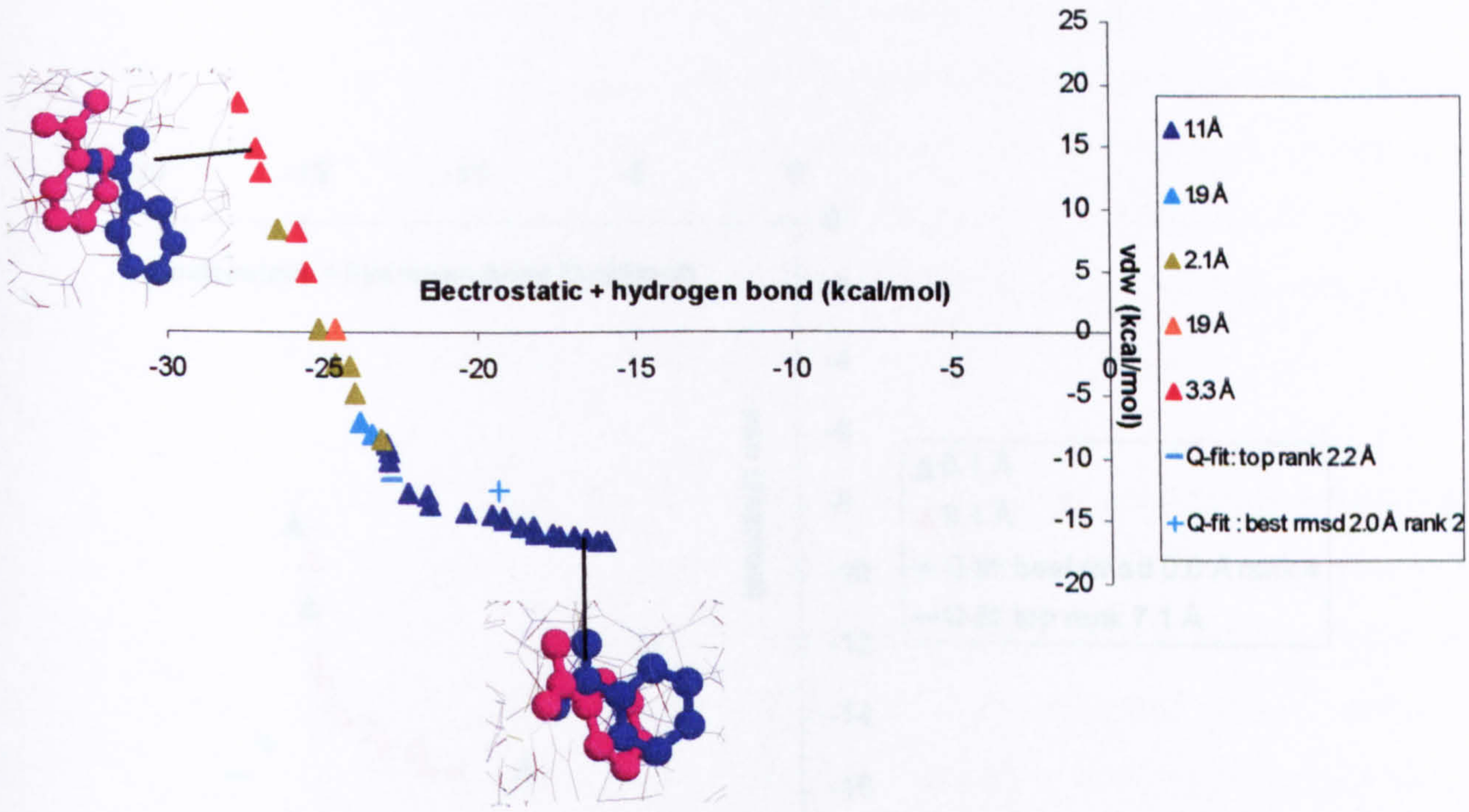


Figure 6-15 Pareto solutions obtained when docking 3ptb. The positions of the top-ranked Q-fit solution and next solution from Q-fit's ranked list that has a good rmsd are also shown.

6.5.6 4dfr

The NSGA-II was not able to successfully dock 4dfr (Figure 6.16). Both Pareto clusters obtained have rmsds of high rmsds of 8.1 Å and 8.4 Å. The top-ranked solution from Q-fit also has a high rmsd, of 7.1 Å, but the fourth ranked solution in the Q-fit list has a good rmsd of 0.6 Å. Both these solutions have lower energies than the Pareto solutions. Two things can be concluded from this test case. Firstly the NSGA-II failed because it did not fully optimise its solutions and therefore the true Pareto front has not been reached. Secondly the fact that the two Q-fit solutions are close to each other in objective space but have very different rmsds indicates that this is a difficult complex to dock. A small change in energy results in a large change in orientation. It can therefore be inferred that the NSGA-II solutions have converged to a local minimum.

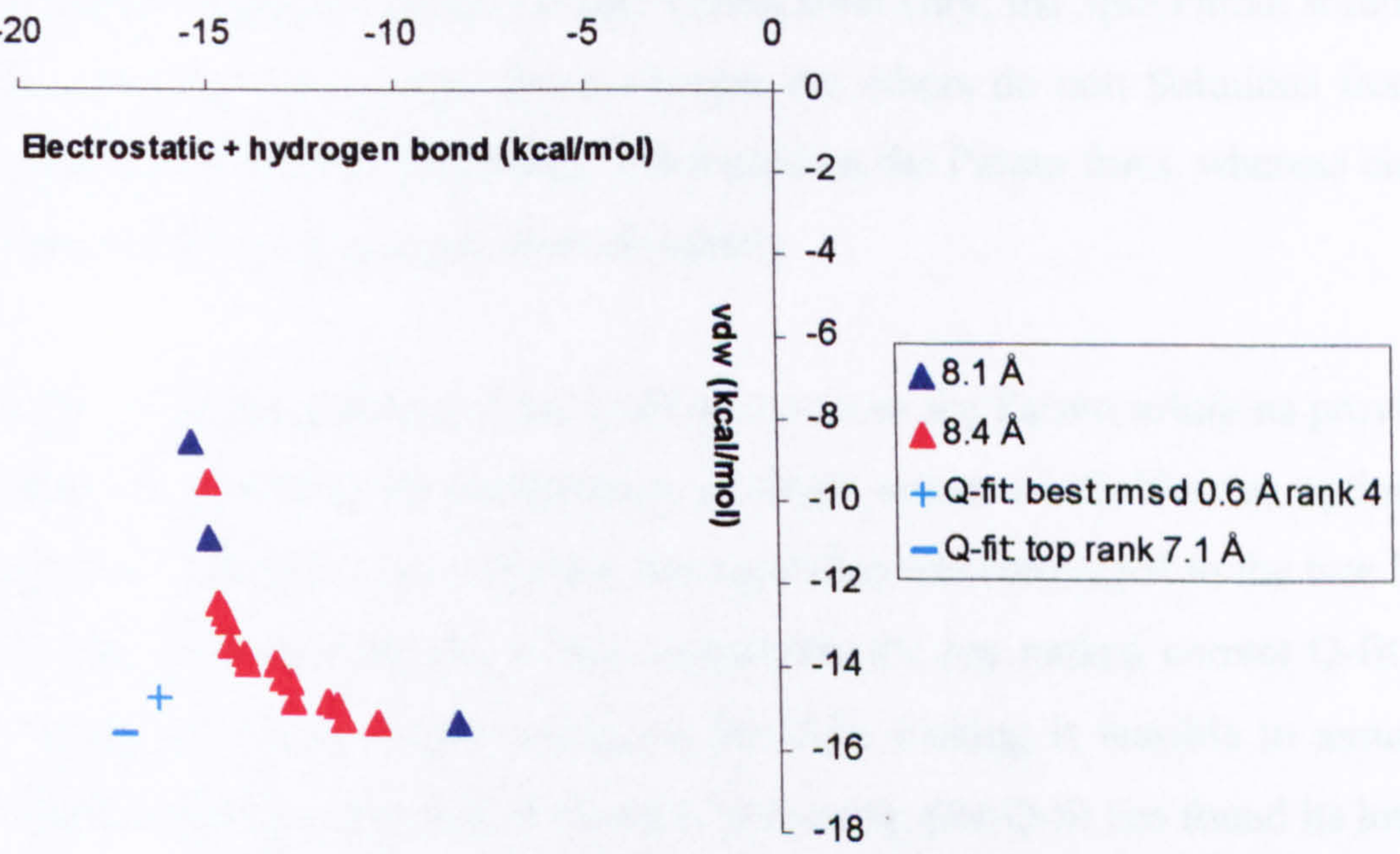


Figure 6-16 Pareto solutions obtained when docking 4dfr. The positions of the top-ranked Q-fit solution and next solution from Q-fit's ranked list that has a good rmsd are also shown.

6.5.7 Summary of results obtained with Dataset 1

The NSGA-II was successful in docking most of the complexes in Dataset 1. The Pareto sets for eight out of the ten complexes contain solutions which have good rmsds. As the plots have shown, the balance of energy types between different complexes varies. For example the correct Pareto solutions for 1ldm, 2phh and 2gbp are influenced by electrostatic and hydrogen bond interactions, whereas the Pareto solutions from 3tpi and 1ulb appear to be equally influenced by both objectives. The number of clusters produced also varies between complexes. Single clusters were produced for 1ldm, 1ulb, 3tpi, 2gbp and 2phh, all of which have good rmsds, with the exception of 2phh. The spread of solutions in the clusters varies, most are spread at evenly over both objectives, 1ldm covers a relatively narrower range than the other single-cluster complexes in both objectives. The remaining complexes, 1dbb, 4dfr, 3ptb, 1stp and 1abe produced more than one cluster. The highest number of clusters obtained for a complex was with 3ptb, where five clusters were produced. The spread of these complexes' clusters on the Pareto front vary; the 3ptb Pareto solutions spread into positive vdw energy space whereas the others do not. Solutions from different clusters for 1dbb are irregularly distributed on the Pareto front, whereas clusters from 1stp and 1abe are grouped more discreetly.

Comparing the position of the Q-fit solutions to the Pareto solutions provides a good basis for assessing the performance of single versus multiobjective optimisation and gives an indication as to whether the algorithm has converged to the true Pareto front or not. With the majority of the complexes, the top ranked correct Q-fit solution is among the correct Pareto solutions therefore making it feasible to assume that the Pareto front has optimally converged (assuming that Q-fit has found its lowest energy solution). Q-fit did manage to obtain correct solutions within its top ranks for the two complexes which the NSGA-II did not find good solutions for, 2phh and 4dfr,. These Q-fit solutions are also dominating the Pareto solutions. This indicates that, since good solutions exist, the NSGA-II did not optimise the solutions for these two complexes fully and that they have not converged to the true Pareto front. Overall however, the NSGA-II has been successful at docking this dataset.

6.6 Dataset 2 results

The NSGA-II was also tested on the 10 complexes from Dataset 2. The following figures show the Pareto solutions obtained when each molecule was docked back into its co-crystallised protein active site. As with Dataset 1, results obtained from Q-fit are also shown, and these are labelled as described previously. The Pareto solutions obtained for complexes from this dataset have also been clustered based on their orientations in relation to the crystal structure.

6.6.1 1acj

Docking 1acj with the NSGA-II resulted in 8 clusters (Figure 6.17). One of these has an rmsd below 2.0 Å, (0.7 Å) and, looking at its position in objective space it can be seen that this correct cluster is more influenced by vdw than by electrostatic and hydrogen bond interactions. The 0.7 Å solutions are spread between 0 and -1 kcal/mol for the electrostatic and hydrogen bond interactions, whereas all of these solutions' vdw interactions are approximately -25 kcal/mol. The rest of the clusters (which incidentally have high rmsds) are spread across both axes, and have decreasing electrostatic and hydrogen bond interactions and increasing vdw interactions. The Q-fit solution with the best rmsd is ranked 3rd and has an rmsd of 0.6 Å. Its position in objective space is among the correct Pareto solutions, and therefore has low vdw interactions and relatively high electrostatic and hydrogen bond interactions. The top-ranked Q-fit solution has an rmsd of 2.8 Å and this solution, as the figure shows, has lower electrostatic and hydrogen bond interactions than the correct solutions. This implies that the interactions are not balanced correctly and that undue influence is given to the electrostatic and hydrogen bond interactions, resulting in the high rmsd of these solutions. Figures 6.18 and 6.19 show the poses of an electrostatic and hydrogen interaction influenced pose (from cluster 5.6 Å) and a vdw influenced pose (cluster 0.7 Å).

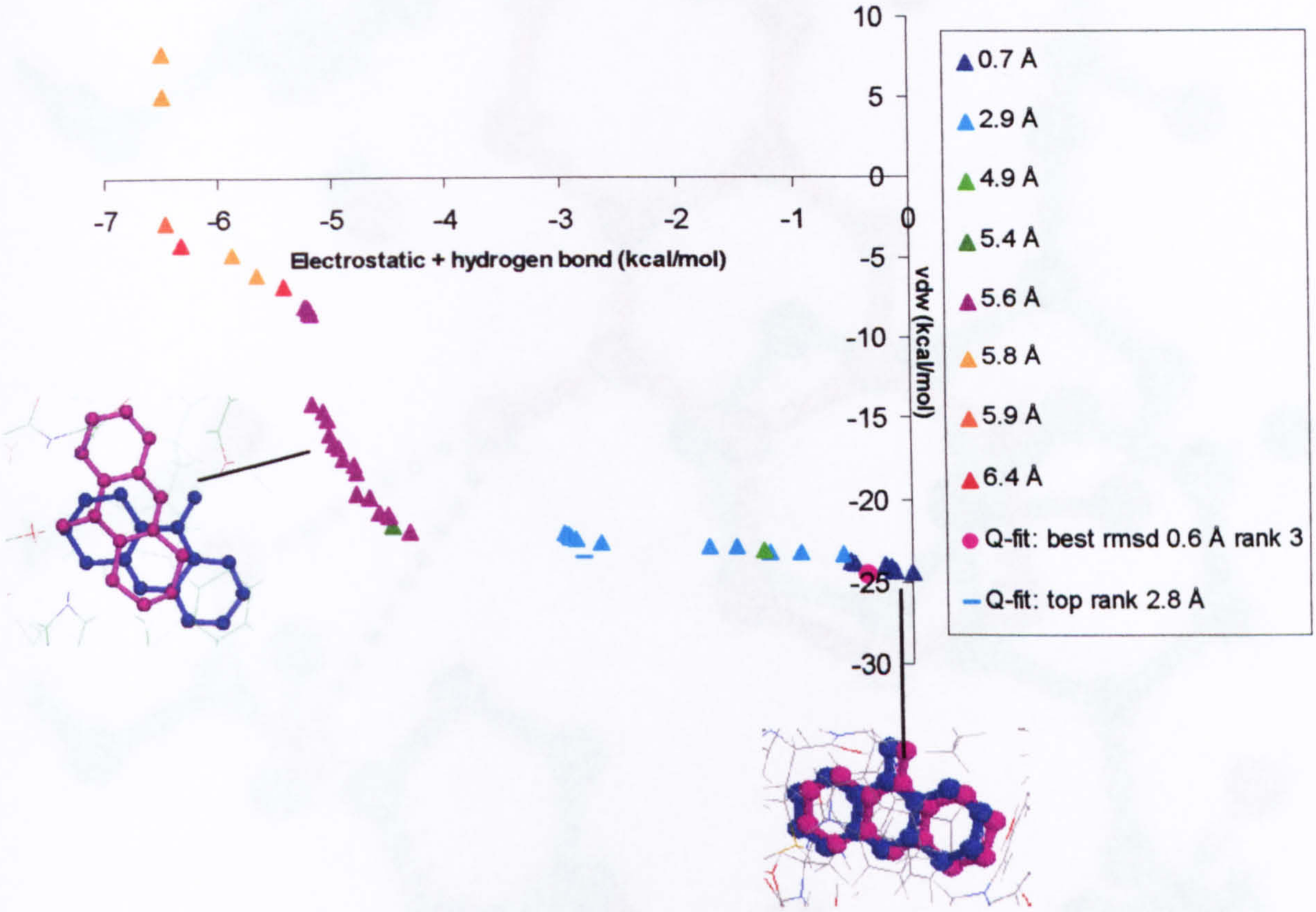


Figure 6-17 Pareto solutions obtained when docking 1acj. The positions of the top-ranked Q-fit solution and next solution from Q-fit's ranked list that has a good rmsd are also shown.

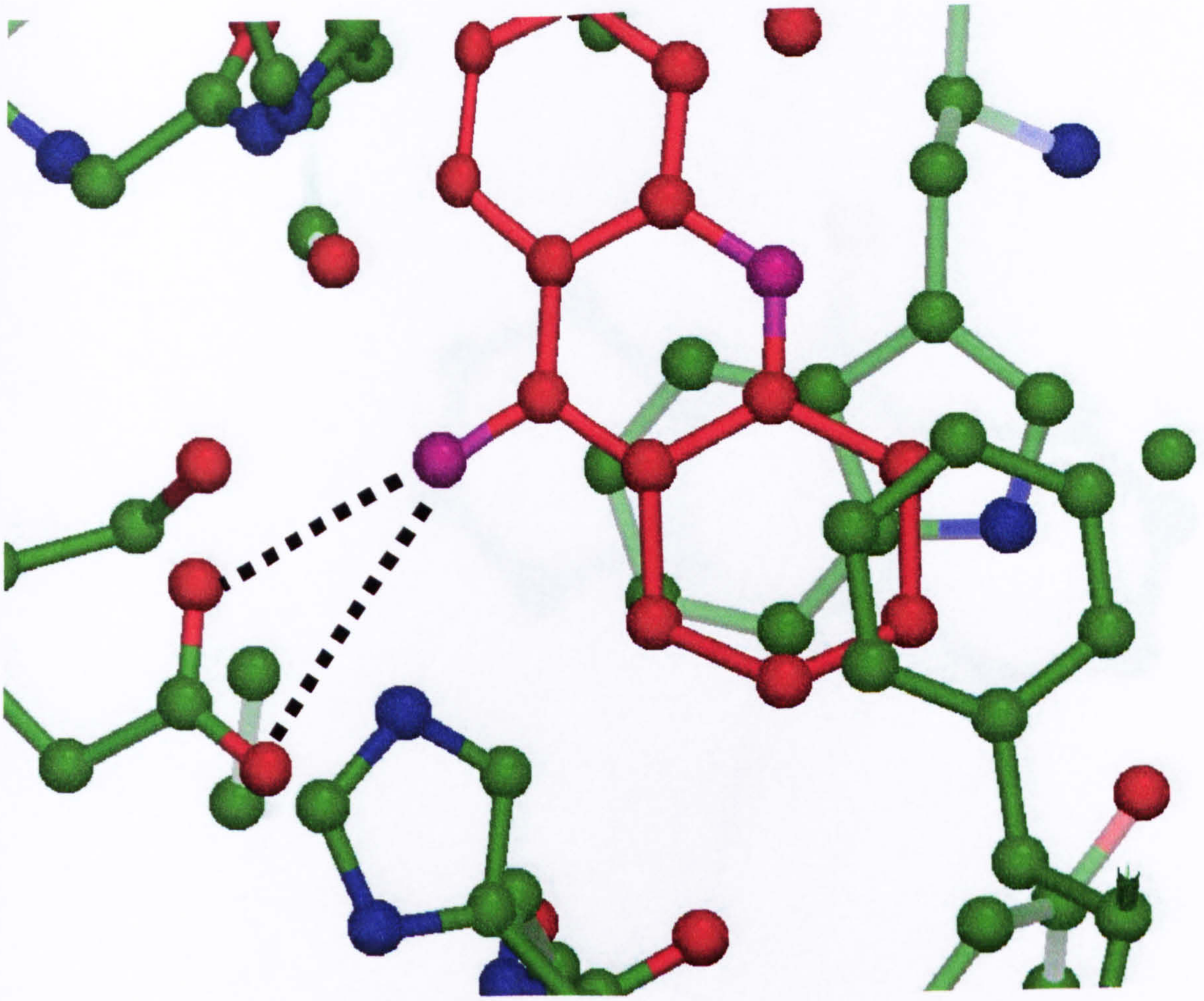


Figure 6-18 A pose of a Pareto solution obtained when docking 1acj. The pose (in red) is more influenced by electrostatic and hydrogen bond interactions, and this is apparent by the two hydrogen bonds made by a nitrogen in the ligand and glutamic acid in the protein binding site.

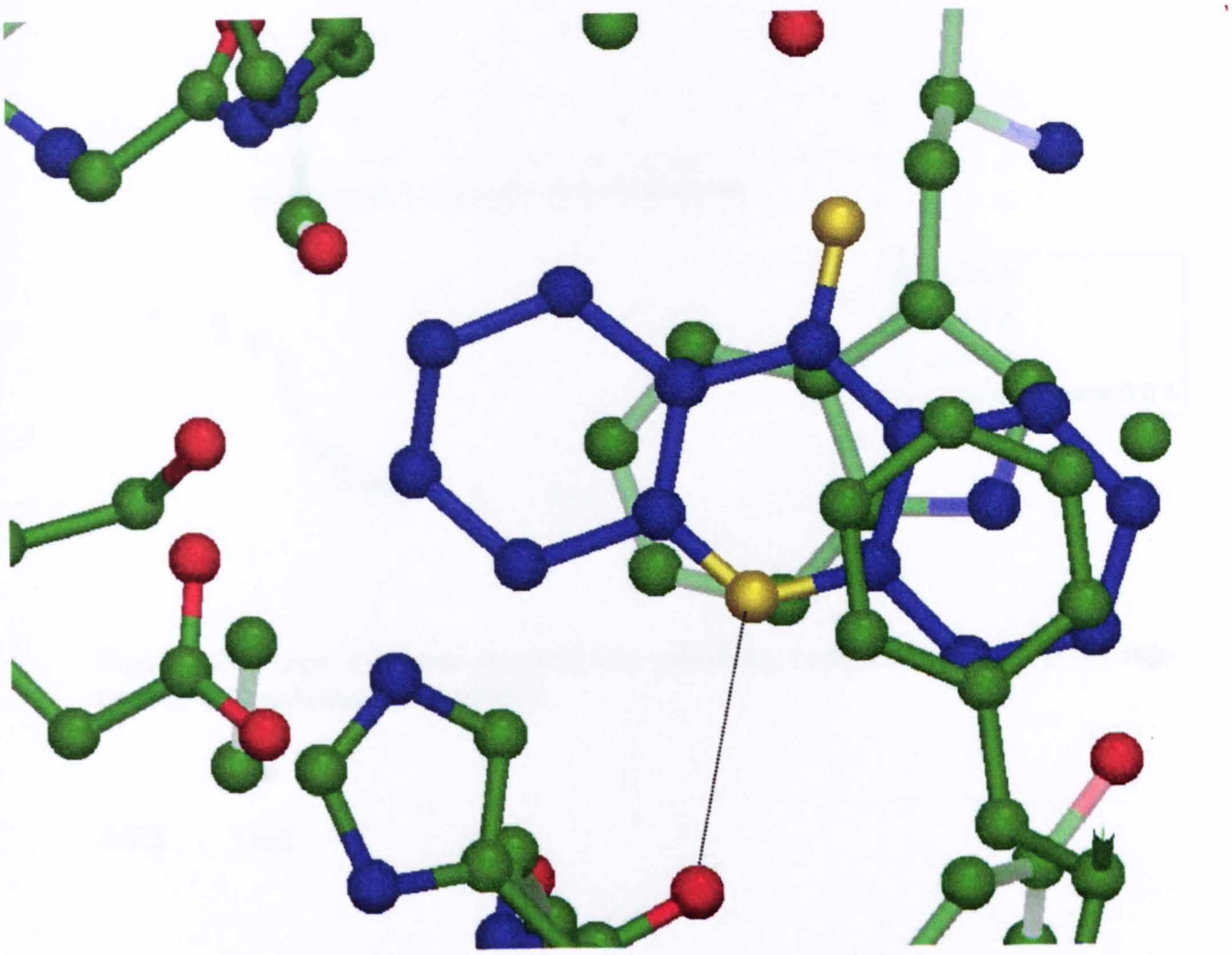


Figure 6-19 A vdw influenced pose obtained when docking 1acj. The pose (in blue) is making vdw interactions predominantly, though the nitrogen in the ligand may possibly be forming a hydrogen bond with a histidine in the protein binding site. The length of this bond pictured is 3.4 Å, which is close to the threshold for hydrogen bonds in GRID (3.5 Å).

6.6.2 1ack

Docking 1ack with the NSGA-II resulted in three clusters, two of which have rmsds below 2.0 Å (Figure 6.20). The 0.8 Å solutions (which are the solutions with the lowest rmsd) have electrostatic and hydrogen bond interactions ranging from ~-6 to ~-10 kcal/mol, and the vdw interactions for this cluster vary from ~-10 to ~-22 kcal/mol. These solutions are therefore more influenced by vdw interactions than by electrostatic and hydrogen bond interactions, though the latter are more negative for this group of solutions than for the correct solutions obtained when docking 1acj. The top-ranked Q-fit solution has a low rmsd of 0.5 Å and is among the correct Pareto solutions.

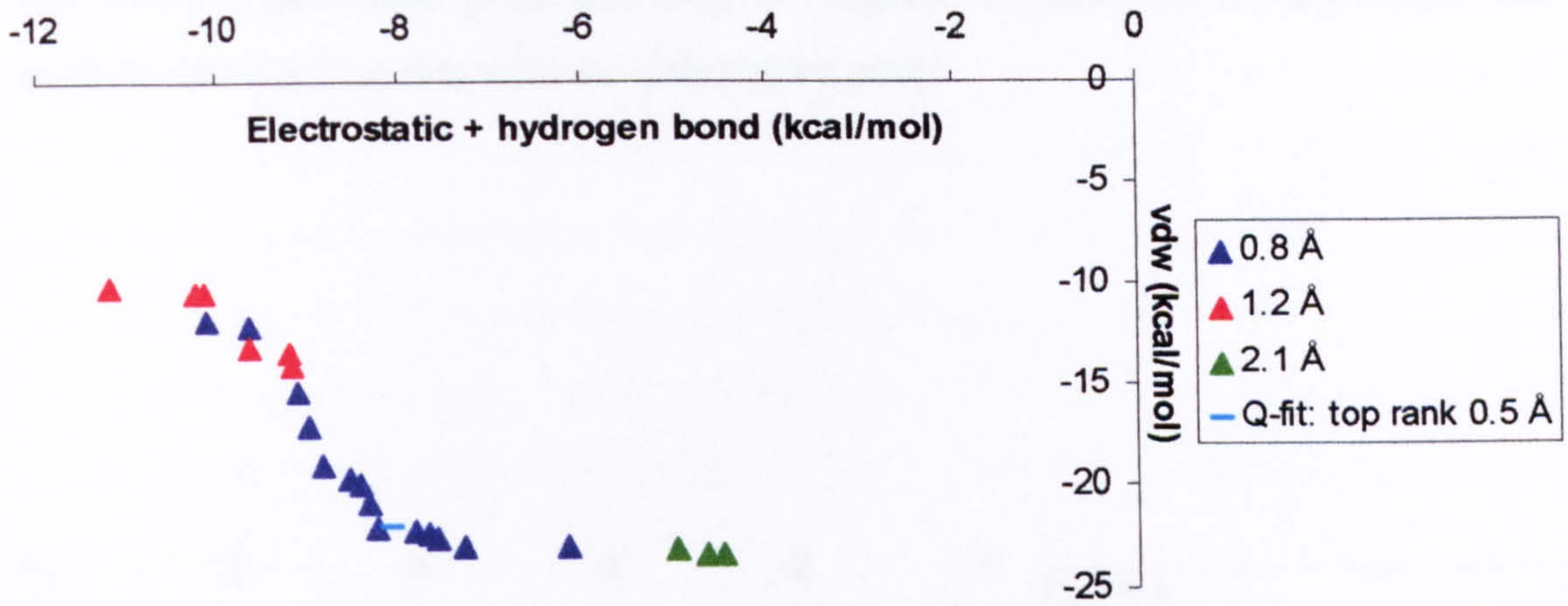


Figure 6-20 Pareto solutions obtained when docking 1ack. The position of the top-ranked Q-fit solution is also shown.

6.6.3 1baf

Nine clusters were obtained when 1baf was docked using the NSGA-II (Figure 6.21). One of these has a good rmsd of 0.6 Å. The position of this cluster is on the right edge of the Pareto front and, in terms of the objectives, is strongly influenced by vdw interactions, which are in the region of -24 kcal/mol. The electrostatic and hydrogen bond interactions, on the other hand, are higher, and have an approximate value of -2.5 kcal/mol. The rest of the Pareto solutions are spread on the Pareto front with increasing electrostatic and hydrogen bond energies and decreasing vdw energies. The top-ranked Q-fit solution has a high rmsd of 4.2 Å and is adjacent to the correct Pareto cluster but with a slightly lower electrostatic and hydrogen bond energy. The lowest rmsd obtained for a solution by Q-fit is 2.1 Å for the solution ranked 11th. As its position indicates, this solution is also dominated by the correct Pareto solutions. The NSGA-II was able obtain solutions with low rmsds of approximately 0.6 Å, and is therefore more successful at docking this case than Q-fit, whose lowest energy solution has a high rmsd (4.2 Å) and whose lowest rmsd solution has a higher energy. The shape of the Pareto front differs from the overall shapes of the Pareto fronts obtained for the other complexes discussed so far. These have the shape of one smooth curve whereas the Pareto front from 1baf is characterised by two smaller

curves which are close to each other. One reason for this may be that each curve represents a particular pose and that the ligand, tetramethyl dinitophenyl, can also dock to this binding site with an alternative pose.

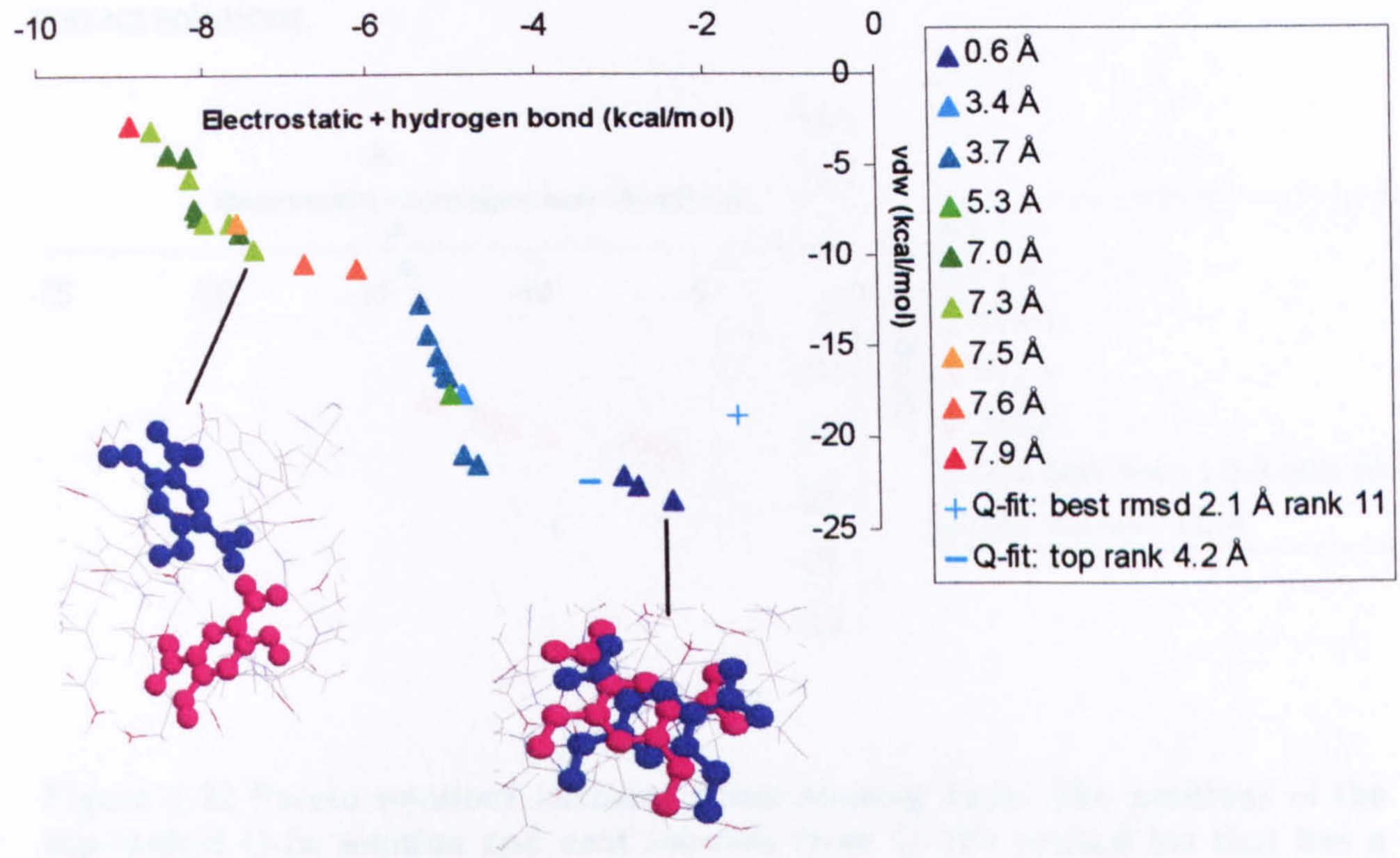


Figure 6-21 Pareto solutions obtained when docking 1baf. The positions of the top-ranked Q-fit solution and next solution from Q-fit's ranked list that has a good rmsd are also shown.

6.6.4 1hdc

With 1hdc, the NSGA-II obtained five clusters, all of which have rmsds which are higher than 2.0 Å (Figure 6.22). The lowest rmsd obtained for a Pareto solution is 9.4 Å, and this cluster has positive vdw interactions and electrostatic and hydrogen bond energies of approximately -15 kcal/mol. The top-ranked Q-fit solution also has a high rmsd of 9.0 Å. This solution has lower vdw energies than the Pareto solutions, and

higher electrostatic and hydrogen bond energies. The lowest rmsd obtained by a Q-fit solution is 1.5 Å, which is at rank 9. This solution has lower vdw energy than all the Pareto solutions and therefore dominates some of these. This implies that the NSGA-II did not succeed in optimising the Pareto solutions, and which has resulted in premature convergence. The low rmsd of the best Q-fit solution implies that its position is where the Pareto front should converge to. If the NSGA-II fails to obtain a correct solution among its Pareto set, as illustrated by this case, then it is not possible to determine which of the objectives, if any, is the most influential in obtaining the correct solutions.

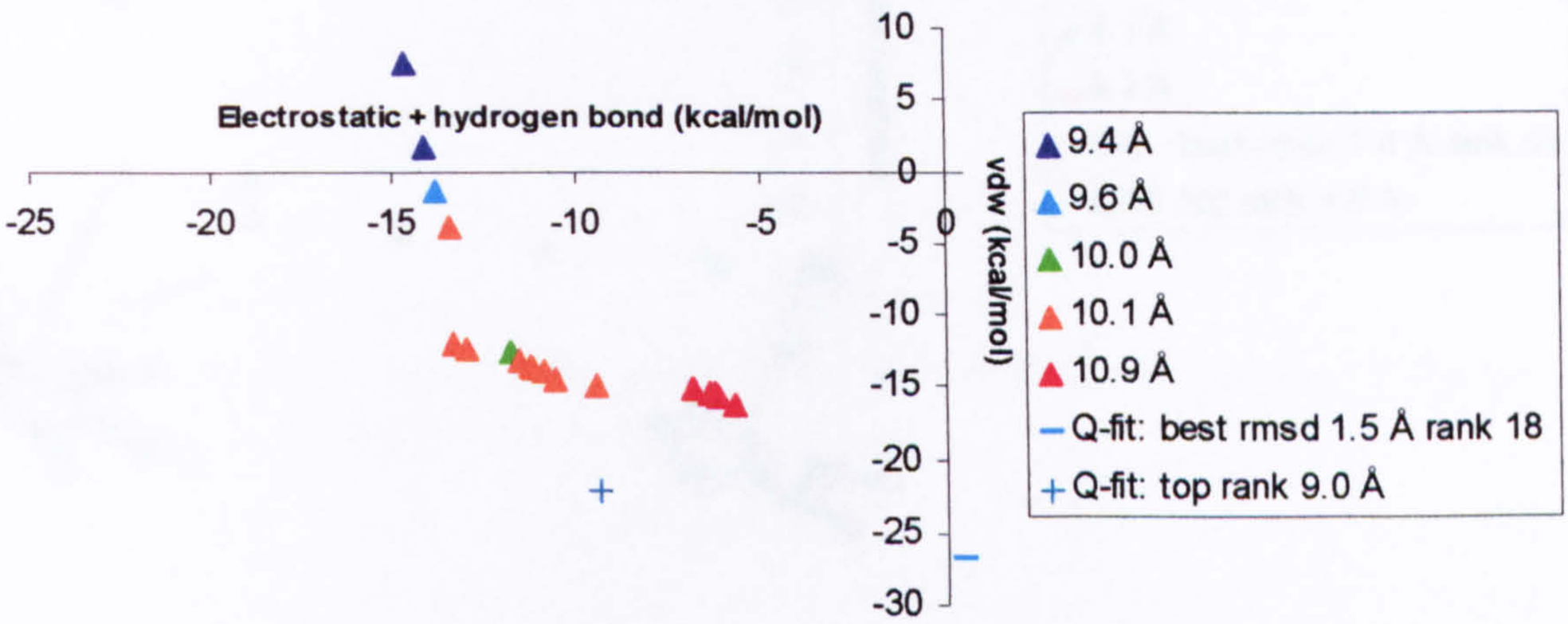


Figure 6-22 Pareto solutions obtained when docking 1hdc. The positions of the top-ranked Q-fit solution and next solution from Q-fit's ranked list that has a good rmsd are also shown.

6.6.5 1mup

One out of four clusters obtained when docking 1mup has a good rmsd of 0.4 Å (Figure 6.23). This cluster has a vdw energy of approximately -14 kcal/mol and electrostatic and hydrogen bond energies of nearly zero. The values of these objectives clearly point to the fact that the vdw energies are having the strongest influence on optimising the poses to obtain good solutions. With Q-fit, the lowest rmsd obtained has an rmsd of 1.4 Å and is at rank 48. This solution is among the correct Pareto solutions. The top-ranked Q-fit solution has an rmsd of 4.6 Å. This solution has lower vdw energies than the correct Pareto solutions and dominates

them. This suggests that the Pareto front has not fully converged and it just so happens that at its prematurely converged point, correct solutions exist. The fact that the Pareto front has not fully converged to the true Pareto front and that the top-ranked Q-fit solution has a high rmsd signify that this complex is difficult to dock.

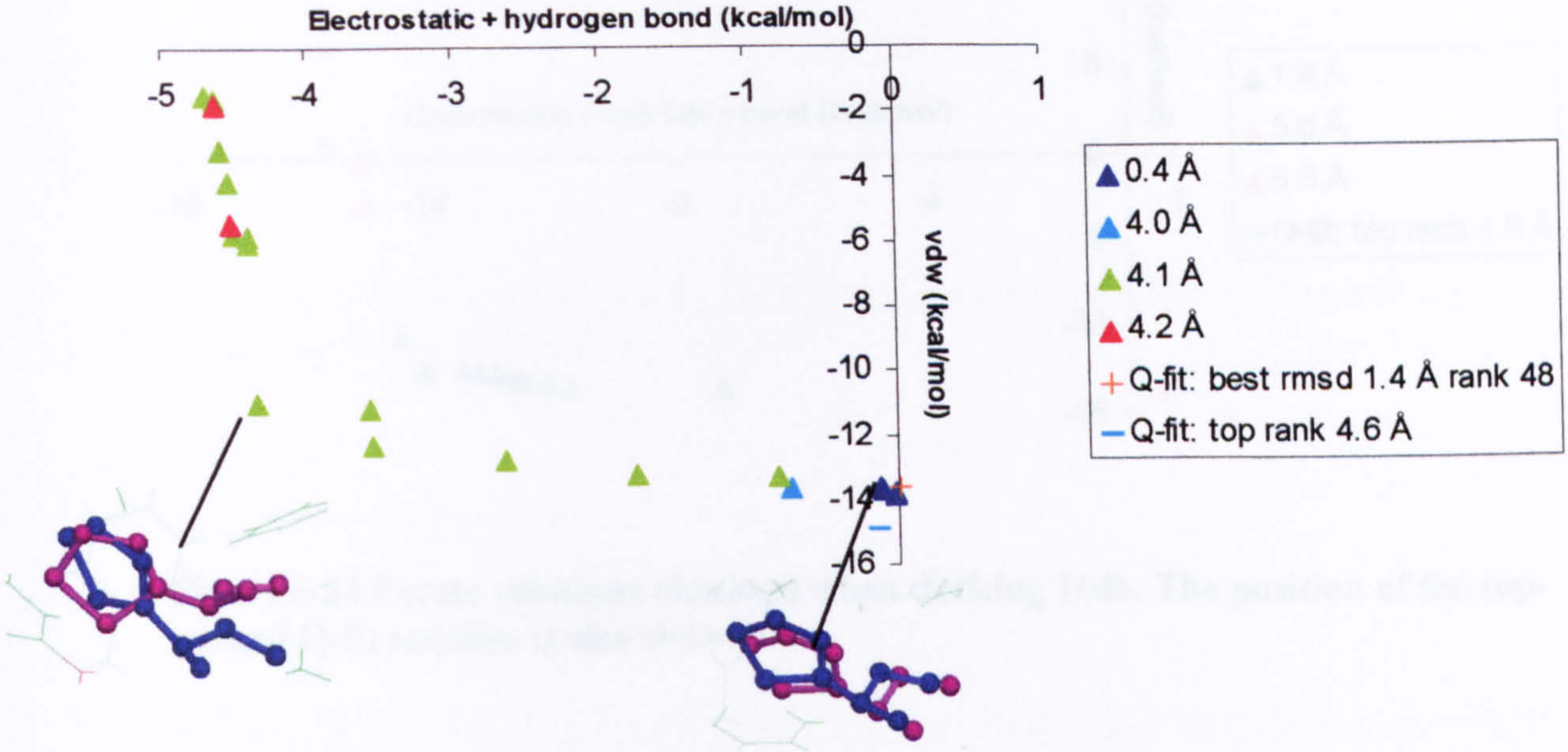


Figure 6-23 Pareto solutions obtained when docking 1mup. The positions of the top-ranked Q-fit solution and next solution from Q-fit's ranked list that has a good rmsd are also shown.

6.6.6 1tdb, 6rsa and 2ak3

The NSGA-II docked 1tdb, 6rsa and 2ak3 successfully, obtaining clusters with good rmsds for all three. With 1tdb, a correct cluster was produced with an rmsd of 1.8 Å, and both objectives appear to be exerting relatively equal influences on these solutions (Figure 6.24). The correct Q-fit solution has an rmsd of 1.5 Å and is among the correct Pareto solutions. With 6rsa one cluster was produced by the NSGA-II, which has an rmsd of 1.0 Å (Figure 6.25). The top-ranked Q-fit solution has an rmsd of 1.4 Å and is dominated by some of the correct Pareto solutions. These solutions are, overall, more influenced by electrostatic and hydrogen bond interactions. Two clusters were produced when 2ak3 was docked, one of which has a good rmsd of 0.8

Å (Figure 6.26). The top-ranked Q-fit solution, as with 6rsa and 1tdb, is also correct and has an rmsd of 1.0 Å.

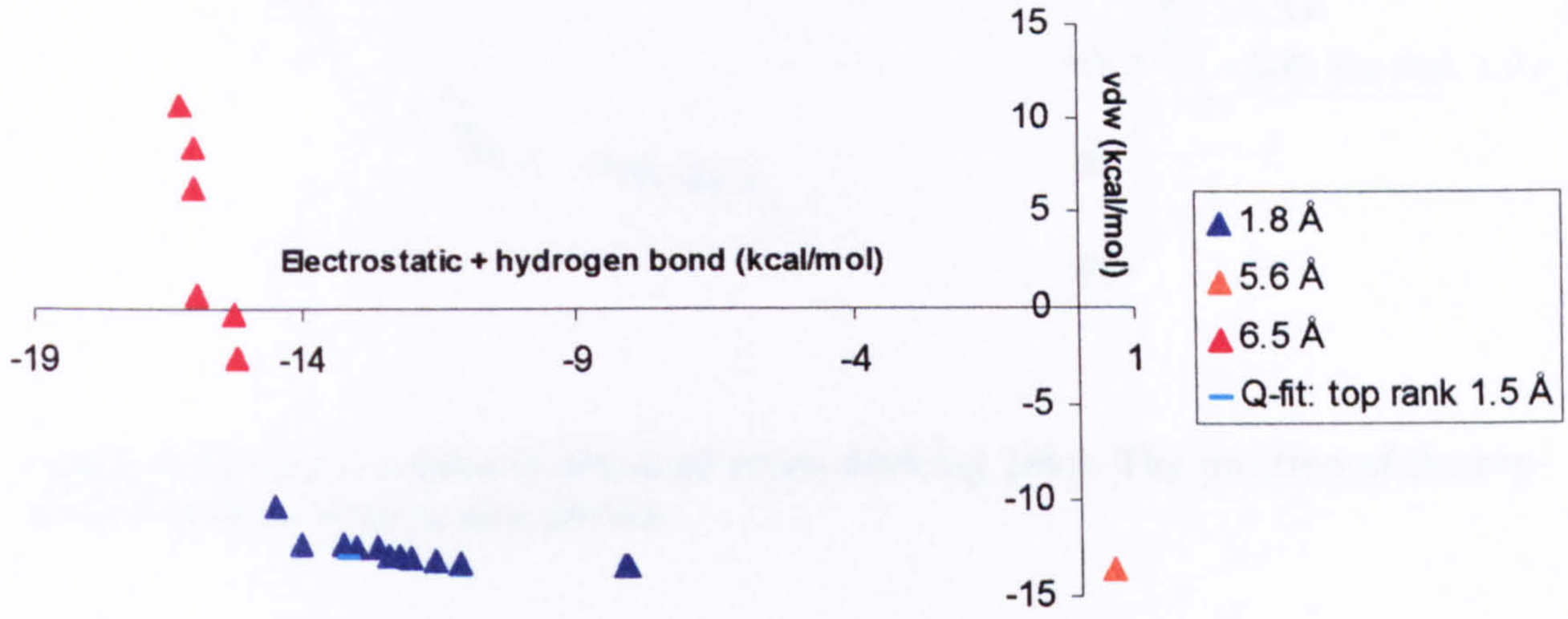


Figure 6-24 Pareto solutions obtained when docking 1tdb. The position of the top-ranked Q-fit solution is also shown.

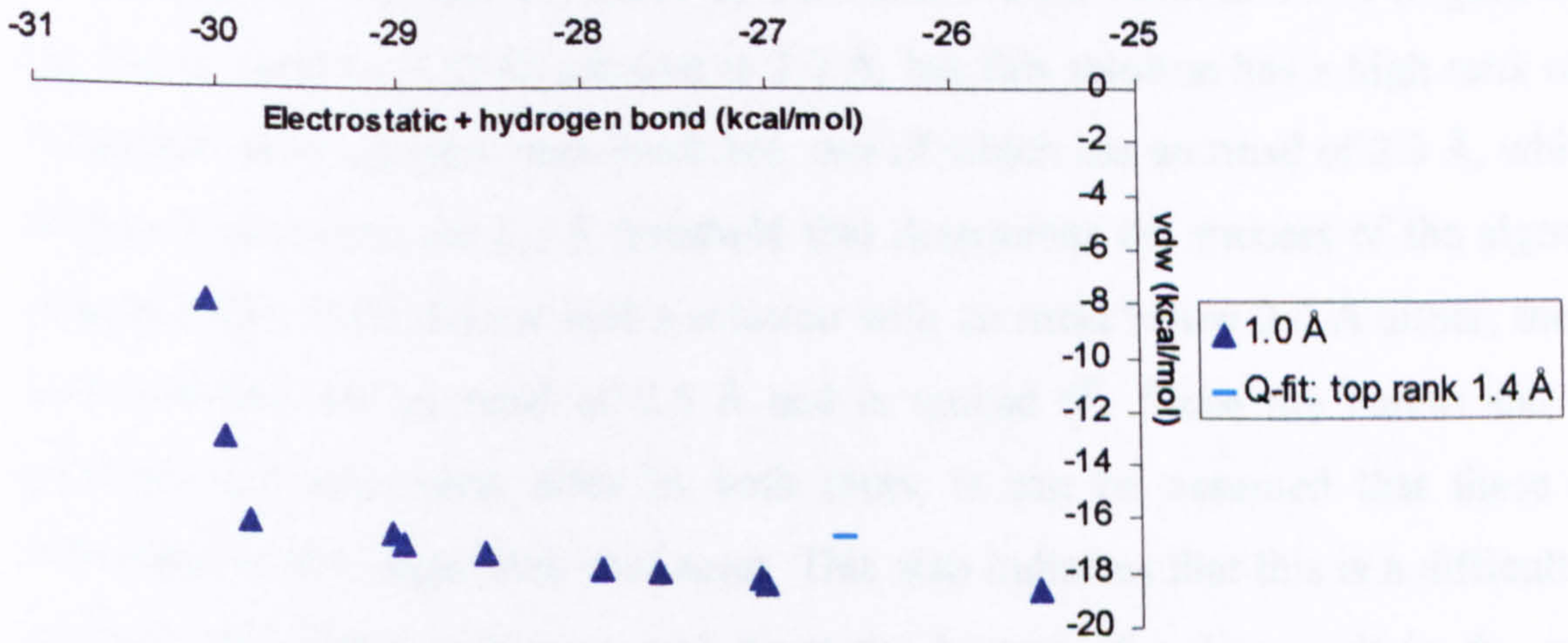


Figure 6-25 Pareto solutions obtained when docking 6rsa. The position of the top-ranked Q-fit solution is also shown.

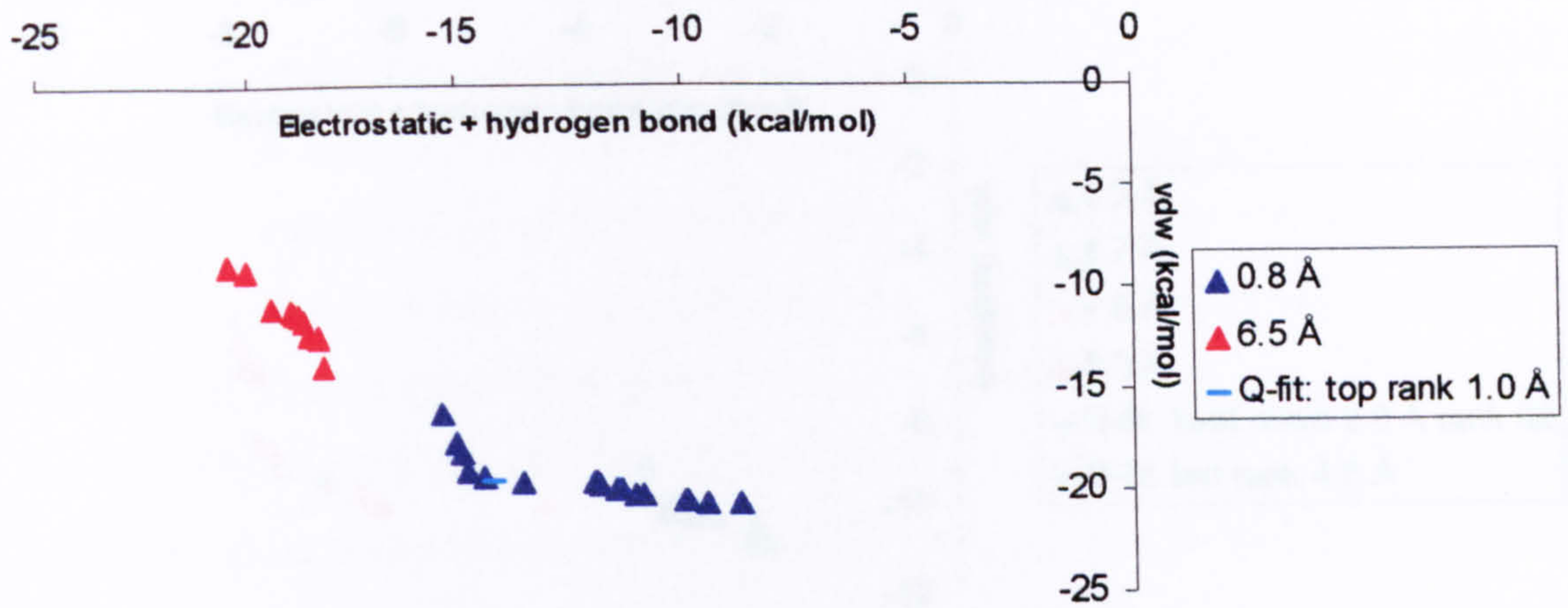


Figure 6-26 Pareto solutions obtained when docking 2ak3. The position of the top-ranked Q-fit solution is also shown.

6.6.7 2mth and 4fab

Neither the NSGA-II nor Q-fit were able to dock 2mth and 4fab successfully. The lowest rmsd for a cluster produced by the NSGA-II for 2mth is 4.3 Å (Figure 6.27); the lowest rmsd for a Q-fit solution is 2.0 Å, but this solution has a high rank of 92. With 4fab, three clusters were produced, one of which has an rmsd of 2.3 Å, which is slightly higher than the 2.0 Å threshold that determines the success of the algorithm (Figure 6.28). Q-fit did not find a solution with an rmsd below 2.0 Å either; the best Q-fit solution has an rmsd of 2.5 Å and is ranked 6th. Since the Pareto and Q-fit solutions are near each other in both plots, it can be assumed that these have converged to the same local minimum. This also indicates that this is a difficult case to dock, the global minimum may be at the bottom of a deep well in the energy landscape which both algorithms failed to reach. As was explained when 1hdc was discussed it is not possible to infer the influence of the individual objectives on finding a correct pose in objective space if the Pareto set does not contain correct solutions.

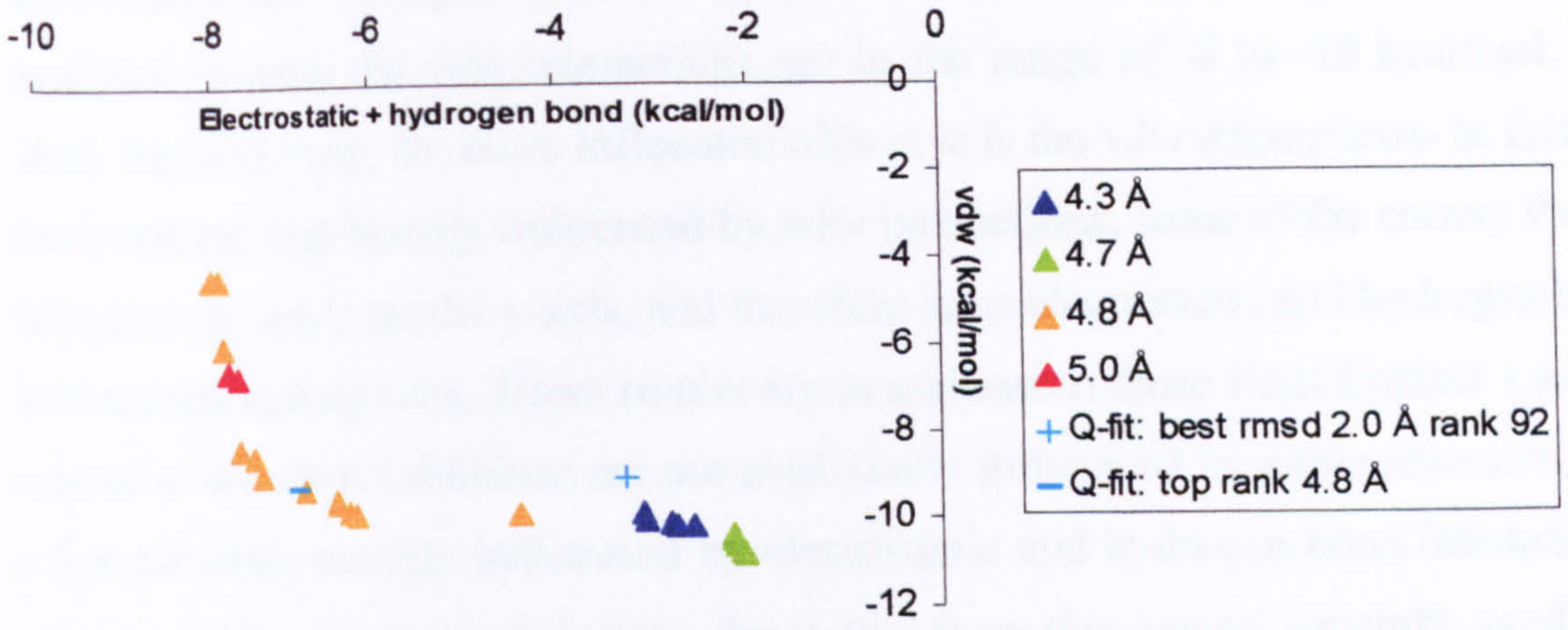


Figure 6-27 Pareto solutions obtained when docking 2mth. The positions of the top-ranked Q-fit solution and next solution from Q-fit's ranked list that has a good rmsd are also shown.

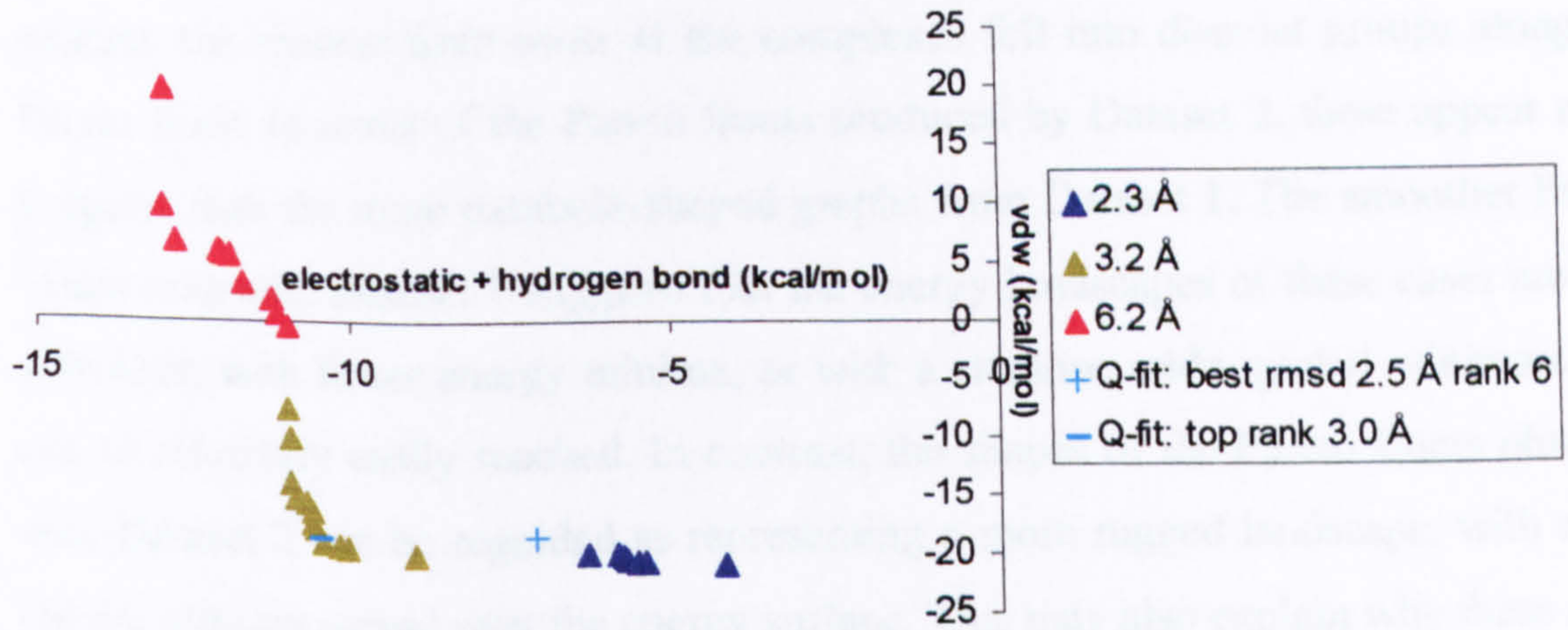


Figure 6-28 Pareto solutions obtained when docking 4fab. The positions of the top-ranked Q-fit solution and next solution from Q-fit's ranked list that has a good rmsd are also shown.

6.6.8 Summary of Dataset 2 results

The NSGA-II was able to obtain correct solutions among the Pareto set in seven out of ten complexes of Dataset 2, which is the more problematic of the two datasets. The positions of the correct solutions in objective space vary between the different complexes. For 1tdb and 2ak3, the Pareto solutions appear to be influenced relatively equally by both objectives. The Pareto solutions for 6rsa are more influenced by

electrostatic and hydrogen bond energies, with values in the range of -25 to -30 kcal/mol, whereas the vdw interactions are in the range of -8 to -18 kcal/mol. For 1baf, 1acj and 1mup, the more influential objective is the vdw interactions- in fact the latter two are very heavily influenced by vdw interactions; some of the correct Pareto solutions are nearly on the y-axis, and therefore have electrostatic and hydrogen bond interactions nearing zero. These results are in contrast to those from Dataset 1 where several of the correct solutions are not particularly influenced by either objective, and a few are more strongly influenced by electrostatic and hydrogen bond interactions. In terms of the numbers of clusters, the results from this dataset generally produced more clusters per complex compared with Dataset 1. This shows that, with Dataset 1, there are fewer orientations that a ligand can take to obtain different balances of the two objectives and vice versa for Dataset 2. The most clusters obtained are with 1baf (9 clusters) and the smallest number of clusters was observed with 6rsa (1 cluster). In general, the clusters from most of the complexes fell into discreet groups along the Pareto front. In terms of the Pareto fronts produced by Dataset 2, these appear more irregular than the more parabola-shaped graphs from Dataset 1. The smoother Pareto fronts seen with Dataset 1 suggests that the energy landscapes of these cases are also smoother, with fewer energy minima, or with a shallow, wide, global minimum that can be relatively easily reached. In contrast, the shapes of the Pareto fronts obtained with Dataset 2 can be regarded as representing a more rugged landscape, with a few similar minima spread over the energy surface. This may also explain why these cases may be considered as problematic, since a rugged landscape provides more opportunities for a search to get stuck in a local minimum.

As was discussed earlier, observing the positions of the Q-fit solutions gives an indication of the performance of the NSGA-II, and allows for the comparison of single versus multiobjective optimisation. For 1acj, 1ack, 1tdb, 6rsa and 2ak3, the top-ranked, or high-ranked lowest rmsd Q-fit solutions are all among the correct Pareto solutions. Since the solutions from both algorithms have low rmsds and are close to each other in terms of energies, it can be assumed that the true Pareto front has been reached by the NSGA-II. For 1baf, the NSGA-II produced better results than Q-fit; the Pareto set contains solutions with low rmsds and these have lower energies than the lowest rmsd Q-fit solution. Also the top-ranked Q-fit solution obtained when docking 1baf has an rmsd that is above 2.0 Å.

Three complexes which the NSGA-II could not dock are 1hdc, 2mth and 4fab. Q-fit did not obtain good solutions at high ranks for these complexes either, though the 18th ranked solution from Q-fit's ranked list of solutions for 1hdc has an rmsd of 1.5 Å. By observing this solution's position, it can be seen that it is influenced strongly by vdw interactions. The Pareto front does not extend as far as this solution; if it did, it may also have found this correct solution. For both 2mth and 4fab, the highest ranked Q-fit solutions have high rmsds, and the Q-fit solutions with the lowest rmsds have high energies. As the NSGA-II was also not capable of finding solutions with rmsds below 2.0 Å within its Pareto set, then it can be inferred that these are difficult complexes to dock, which have global minima that are difficult to find. The ligands in Dataset 2 in general appear to be more strongly influenced by different interaction energy types than the ligands in Dataset 1. Vdw interactions play a more predominant role than the electrostatic and hydrogen bond interactions, and as illustrated by 1baf and 1acj. Overall, Q-fit performed better with Dataset 1 (where it docked all 10 complexes) than Dataset 2 (where it failed to find good solutions at high ranks for 5 complexes). Likewise the NSGA-II was more successful with Dataset 1, where it docked 8 complexes successfully relative to 7 in Dataset 2. The less favourable results obtained with Dataset 2 are not surprising as the complexes in this dataset are recognised for their more problematic nature.

These results have demonstrated that obtaining correct poses is not solely dependent on the total energy, but on the balance of the interaction energies that a particular solution makes within a protein binding site. For example looking at 2ak3 (Figure 6.26), a vdw interaction value of -20 kcal/mol and electrostatic and hydrogen bond energy value of -10 kcal/mol gives a total energy of -30 kcal/mol, and the resultant ligand structure has an rmsd of 0.8 Å. Whereas if these values were inverted, the total energy would still be -30 kcal/mol, but the ligand would have an rmsd of 6.5 Å. Therefore a single objective algorithm, guided by the total energy, would not be able to differentiate between both scenarios, and the algorithm has an equal probability of finding the high-rmsd solution as the low-rmsd solution. This clearly indicates the importance in obtaining the right balance of interaction energy types in order to find solutions with low rmsds. The distinction which the NSGA-II has over single

objective optimisation algorithms is that it can differentiate between these two scenarios.

These results have also shown that that the energy terms have varying influences between different complexes. With single objective optimisation it is not possible to infer which energy term has the strongest effect in optimising the search, whereas the importance of a particular energy type can quickly be inferred through multiobjective optimisation by observing the positions of the correct Pareto solutions in objective space. This extra information which multiobjective optimisation provides highlights the potential for this technique, and which will be further explored in the following chapters.

Chapter summary

In this chapter, the results obtained when testing the NSGA-II on Datasets 1 and 2 have been described. The NSGA-II obtained correct solutions within the Pareto set for eight out of the ten complexes in Dataset 1, and for seven out of the ten complexes of Dataset 2. The results have demonstrated that for some cases a correct balance of the individual interaction energy terms needs to be achieved to find correct solutions. The balance of energy terms may also differ between different complexes.

7 Algorithmic Enhancements to the NSGA-II

The results from the previous chapter show that the NSGA-II performs satisfactorily in obtaining "correct" solutions within the Pareto solution set. Most of the failures which were observed occurred with the complexes in the problematic data set. The one exception is 4dfr, which is part of the Dataset 1, and which the NSGA-II could not dock. 4dfr (dihydrofolate reductase and methotrexate) is regarded as the "classic" test complex. It has been used as a test case in some of the earliest docking programs (Oshiro *et al.*, 1995, DesJarlais *et al.*, 1986). This is mainly because it is a therapeutically important target and has been involved in several inhibitor design studies. The solutions on the Pareto front obtained when docking methotrexate into dihydrofolate reductase are approximately 8 Å from the crystal structure. To be able to understand the reason for this failure, an initial population was created that contained several copies of the crystal structure. When the NSGA-II was run on this population, good solutions were obtained. These also had lower energies than solutions obtained from a random initial population. This is an indication that solutions that have low energies and good rmsds from the crystal structure can exist, but the algorithm cannot find them.

To explore the capabilities of the algorithm, and to improve some of the results obtained with the two datasets, three different strategies were implemented. These are:

1- controlled elitism
2- reduction of E_{max} to smooth the energy landscape
3- refined local search using simplex minimisation with a Lamarckian element

7.1 Controlled Elitism

As was described in Chapter 6, the NSGA-II was found to be very effective at converging to the Pareto front, by ensuring no good solution is ever lost from the

population. However, it has been pointed out that with some functions, especially where there are several non-dominated solutions in the population, the NSGA-II may not be as effective (Jianjun Hu *et al.*, 2003; Deb and Goel, 2001). By only taking the best solutions onto the next generation, which may occupy only 2-3 ranks, there is a risk of flooding the population with good chromosomes that are similar to each other, resulting in the population's premature convergence to a local minimum. Intentionally excluding the worst chromosomes may be intuitive, but this also reduces the amount of diversity in the population which will lead to premature convergence. This has led to the development of NSGA-II with controlled elitism (Deb and Goel, 2001). As the name implies, controlled elitism allows the level of elitism to be "controlled" by a certain parameter. In the controlled NSGA-II the number of chromosomes that are selected from a given rank to pass to the next generation becomes adaptive and is maintained at a specified distribution. This may be any type of distribution- in this case a geometric distribution was used, though an arithmetic distribution was also experimented with.

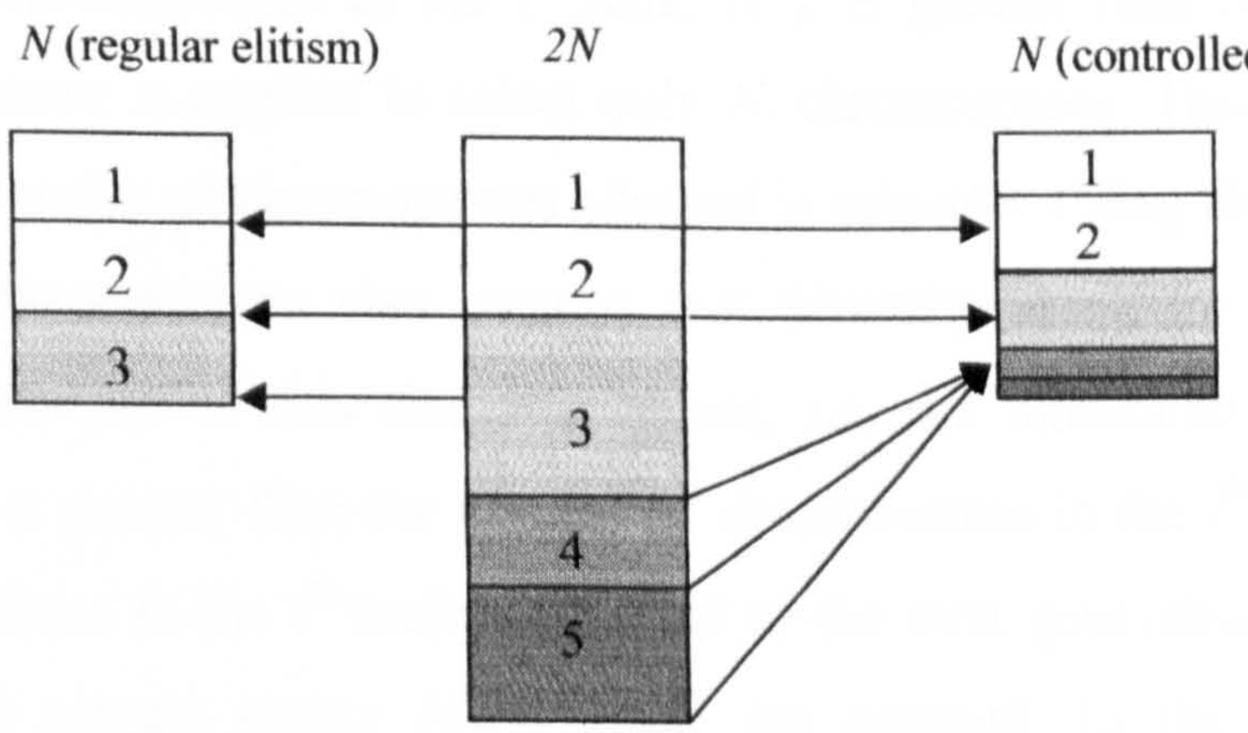


Figure 7-1 Comparing the reduction of population of size 2N down to N in elitism and in controlled elitism. The middle rectangle represents the population 2N, and the separate compartments show a hypothetical distribution of solutions across five ranks. The rectangle on the left shows the ranks of the chromosomes that make up N for the case of non-controlled elitism. The rectangle on the right shows the ranks of chromosomes when N is produced by controlled elitism.

The geometric distribution is shown in the following equation:

$$N_i = rN_{i-1} \qquad \text{where} \quad r < 1 \qquad \text{Equation 7.1}$$

Where N_i is the maximum number of allowed individuals in the i^{th} rank and r is the reduction rate. r is a user-defined parameter, but it also allows the procedure to behave adaptively. As with the standard NSGA-II, in controlled NSGA-II the parent population (N) is combined with the offspring population (N), and the combined population ($2N$) is Pareto ranked. The maximum number of chromosomes (N_i) from each rank that could potentially be passed to the next generation is:

$$N_i = N\frac{1-r}{1-r^K}r^{i-1} \qquad \text{Equation 7.2}$$

where K is the total number of ranks in the combined population ($i = 1, 2, \ldots, K$). Because $r < 1$, the highest rank will have the largest number of allowed chromosomes and N_i will become exponentially reduced with increasing rank numbers.

The situation may arise where there are more solutions in the i^{th} rank than N_i i.e. the number of chromosomes in the i^{th} rank, N'_i, is greater than N_i. In such cases the niching operator is applied to select only N_i chromosomes. This will ensure that the maximum number of chromosomes allowed is selected. Using the niching operator to select the chromosomes also ensures that diversity along the rank is maintained. When the reverse of this situation occurs, i.e. the maximum number of allowed individuals is *greater* than the number of chromosomes in the i^{th} rank ($N_i > N'_i$), then all the solutions in the i^{th} rank are passed to the next generation and the number of slots which remain empty $N_i - N'_i = \rho_i$ are counted. In the following rank, the maximum allowed individuals is increased- so that $N_{i+1} = N_{i+1} + \rho_i$. This is also compared to the number of solutions in the $i + 1$ rank in the same manner as described for the previous rank, and the process continues, until N solutions have been selected. In this way a distribution of solutions is selected from the majority of all ranks, and the resulting population N will retain the diversity of $2N$.

The situation may arise where, after a pass has been made over all the ranks, a few slots remain empty. This is most likely to occur with a large r. In such situations another pass is made with the chromosomes that have been left out, in

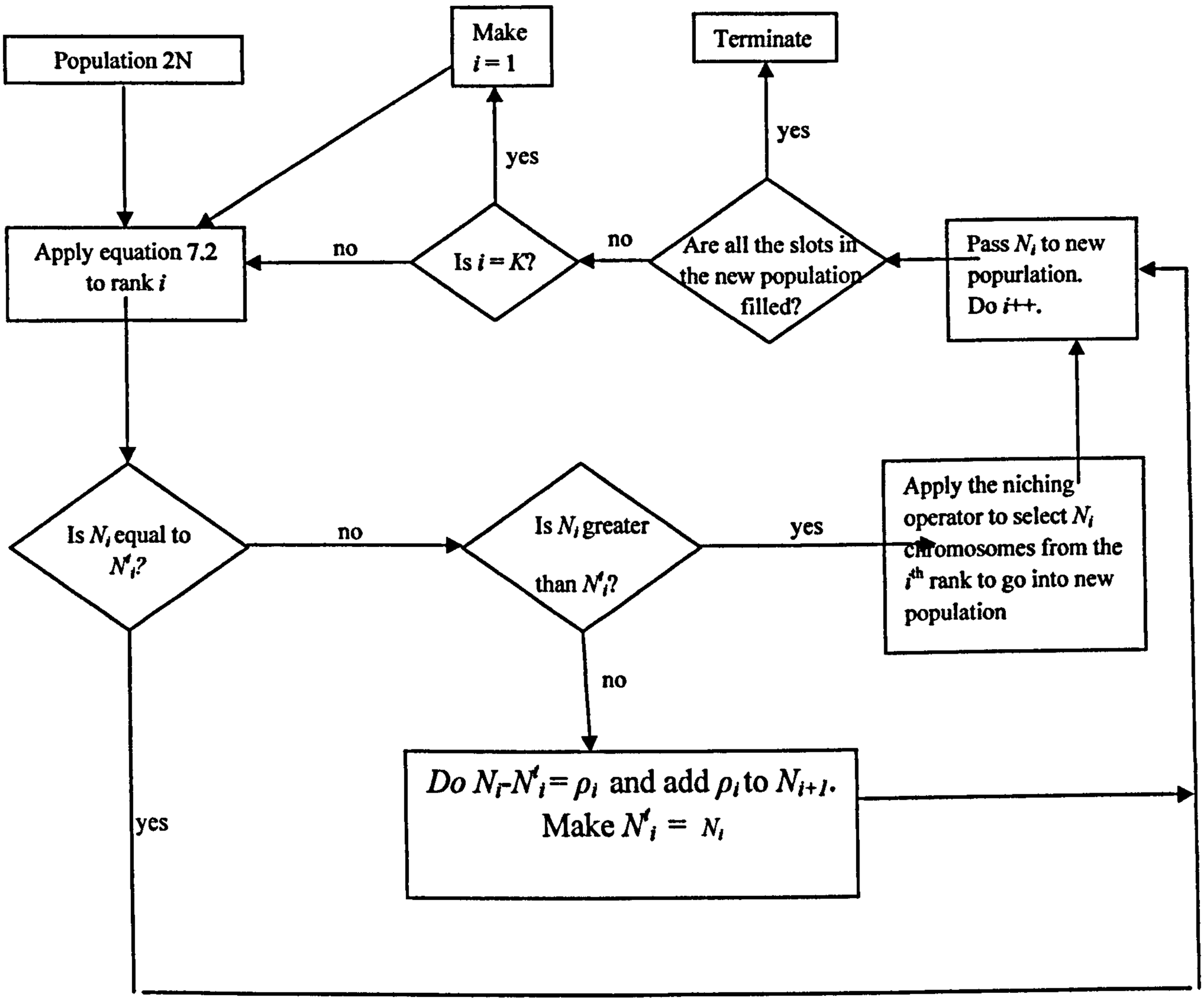


Figure 7-2 Schematic of controlled elitism feature of NSGA-II

the same manner as described previously. This procedure is repeated until all slots of N are occupied.

The NSGA-II was modified so that, at the end of a generation, the technique described above is used to select the chromosomes from the combined *2N* population, to form population *N* that is passed on the next generation. The reduction rate, or *r*, used is 0.1. This version was used to dock the more problematic complexes, such as 4dfr.

The effectiveness of the controlled NSGA-II was tested by redocking the ligands from Datasets 1 and 2 back into their binding sites. The new modification did not cause a profound change to the results discussed in the previous chapter. It was, however, interesting to observe if there was any improvement in docking methotrexate into 4dfr. As was discussed in the previous section, 4dfr is one of the more significant test cases and an algorithm's ability to dock the molecule is a good indication of the algorithm's performance. The Pareto solutions in objective space obtained when docking methotrexate back into the dihydrofolate reductase active site are shown in Figure 7.3. As the figure indicates, none of the Pareto solutions have good rmsd values relative to the crystal structure, which indicates that a low diversity in the population may not be the reason why the algorithm cannot dock 4dfr correctly. However, comparing this plot with that generated from the initial NSGA-II (Figure 6.16), it can be seen that the controlled NSGA-II produces a Pareto front that is slightly more advanced in terms of the electrostatics and hydrogen bond energies (most favourable electrostatics and hydrogen bond energy obtained is approximately -17 kcal/mol; this is relative to ~ -15 kcal/mol obtained with the initial NSGA-II). The advancement of the Pareto front generated by the controlled NSGA-II on the y-axis (the vdw energies) is similar to that of the initial NSGA-II. The two plots have the same number of clusters, but the clusters from the controlled NSGA-II are more diverse in terms of orientations. The rmsds of these clusters range from 8.8 Å to 16.5 Å. The rmsd represented by red triangles (16.5 Å) is high because the vdw energies for this cluster are positive, which means that these solutions may be physically clashing with atoms of the binding site. All the clusters from the initial NSGA-II are orientationally similar, with rmsds of approximately 8 Å from the crystal structure. The Pareto solutions from the controlled NSGA-II have a wider range in terms of the vdw interactions than the initial NSGA-II (range from -15.8 to 15.2 kcal/mol versus -15.3 to -8.2 kcal/mol), and slightly less varied in terms of the electrostatic interactions (-19.01 to -13.2 kcal/mol versus -15.37 to -8.4 kcal/mol).

This result indicates that using controlled elitism can slightly improve the performance of the algorithm when run on 4dfr, in terms of the optimization of the

electrostatics and hydrogen bond energies. Also the diversity in orientation of the Pareto solutions is higher with the controlled NSGA-II than the initial NSGA-II. Nevertheless the controlled NSGA-II was not able to successfully dock any of the failures from datasets 1 and 2. This led to further developments in the algorithm, as explained below.

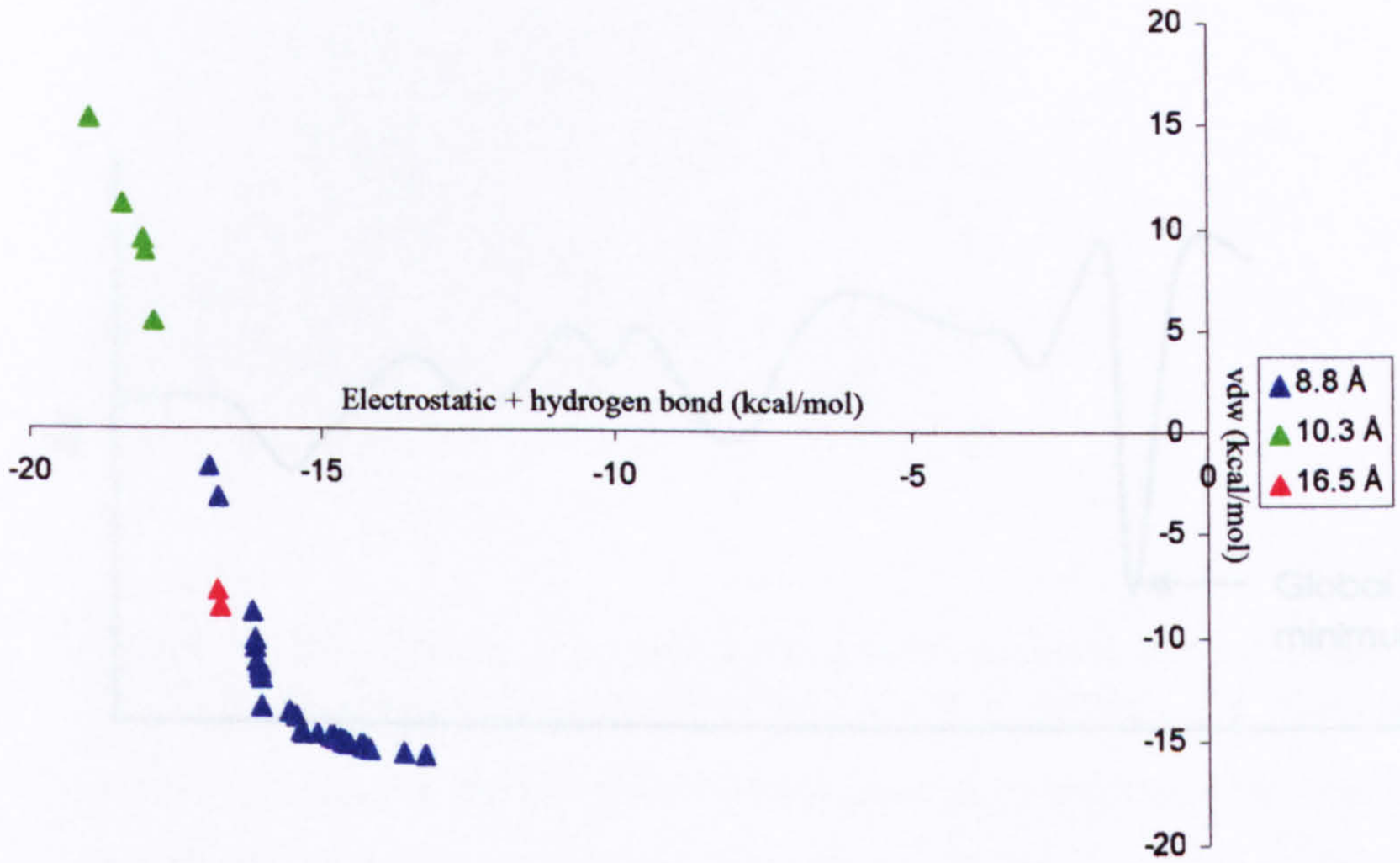


Figure 7-3 Pareto solutions obtained when docking 4dfr using controlled NSGA-II.

7.2 A deep energy well- Reducing E_{max}

An issue that may arise is the ability of the algorithm to go down narrow, steep, energy wells. Given an energy surface with several shallow local minima and a global minimum that is at the bottom of a narrow and steep well, there is a possibility for the algorithm to only explore the shallower, more accessible local minima, without reaching the global minimum (Figure 7.4).

This scenario was observed with 1hdc. As can be seen Figure 7.5, the initial population is evenly distributed within the grid box. The population then very rapidly

congregates to a shallow cleft – away from the binding site. One way for the population to emerge from this situation is through a random event, i.e. for the mutation operator to generate a change in a gene which is substantial enough to move a chromosome to the more energetically favourable and correct cleft. However, if such large changes were allowed this would make it very difficult to ensure that the algorithm is successful at reaching the true Pareto front each time.

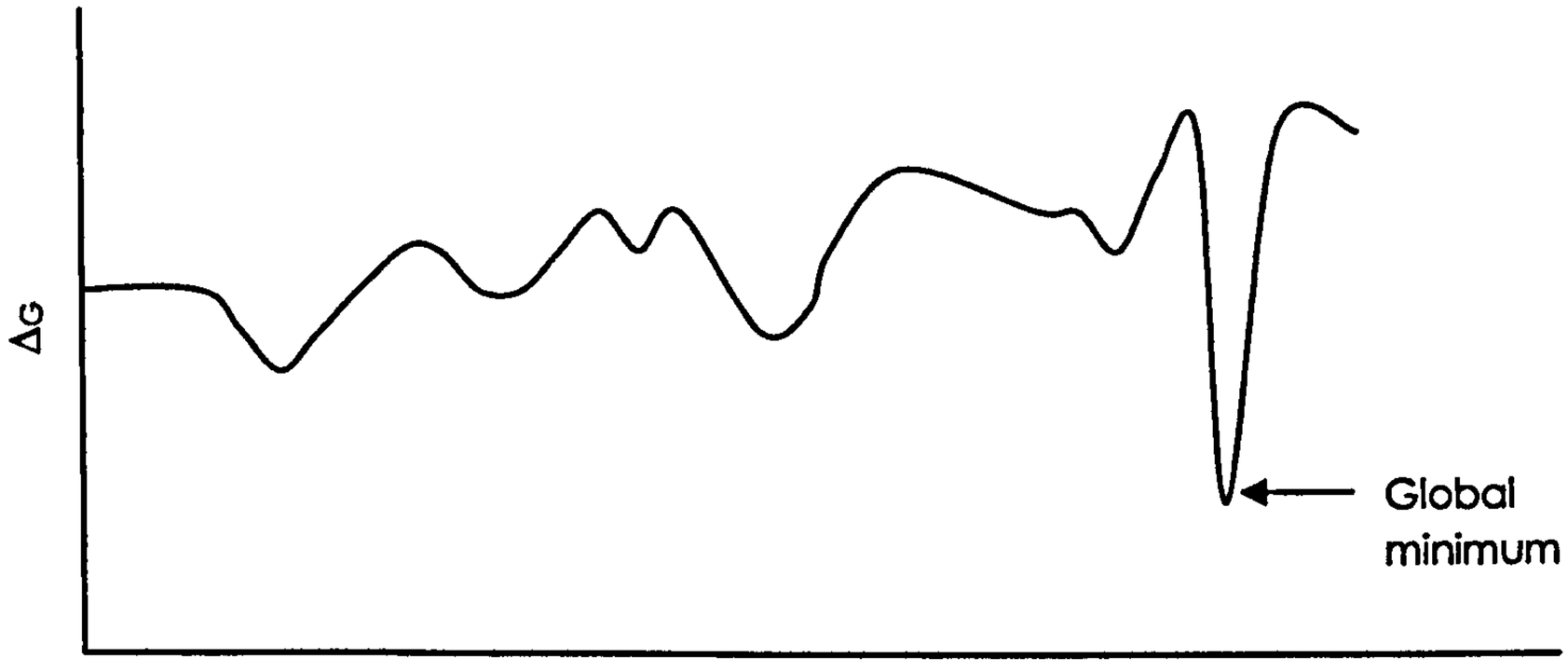


Figure 7-4 Hypothetical energy surface an illustration of an energy landscape that may prove challenging to the NSGA-II. The landscape consists of several local minima that are more easily accessible than the global minimum, which is down a steep well. Note that for clarity a single objective plot has been used here. A multiobjective plot consists of several dimensions.

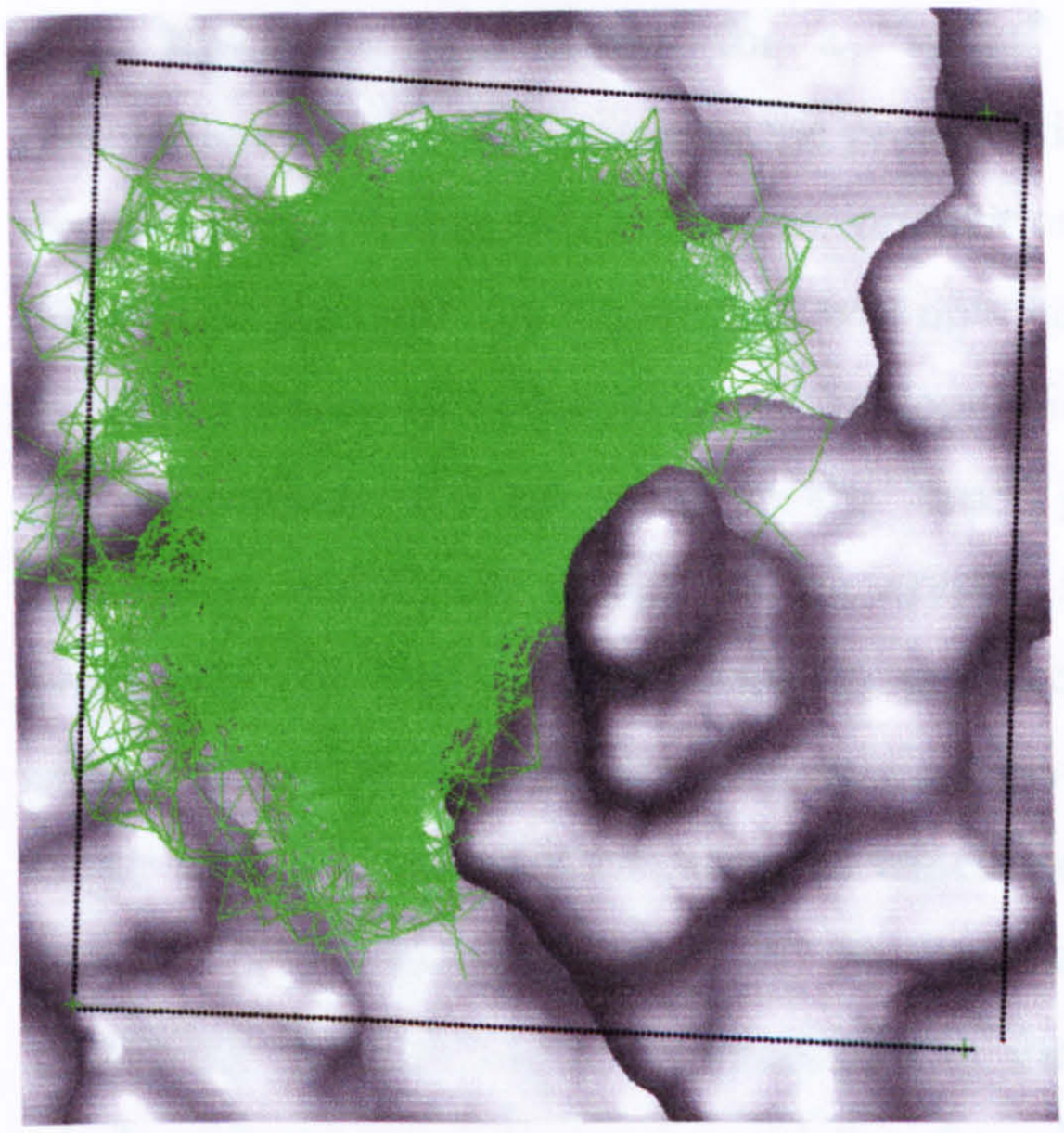


Figure 7-5 Distribution of initial population of NSGA-II when docking 1hdc

By studying these figures more closely, and observing the binding site, it may be presumed that, for the genes of a chromosome to translate the ligand from the corner of the box, to the binding cleft, they will need to be very close to the "true" genes, i.e. those genes that would translate the ligand to the crystal structure's orientation. Because the binding site of the protein is tight, any chromosome that produces a ligand orientation relatively similar to the crystal structure may result in orientations that sterically clash with the protein- and hence have high van der Waals energies. These are either rejected for exceeding the bumps threshold or are eliminated early from the population due to their low Pareto rank which puts them at a disadvantage for selection by the reproduction operator.

The most effective way of dealing with this issue is to soften the energy landscape. Energy landscape softening is a feature of the docking tool GOLD (Jones *et al.*, 1997) where a 4-8 Lennard-Jones potential is used rather than a 6-12 and the choice of parameters ensures that its minimum equals that of a 6-12 potential. In our case, reducing the energy threshold E_{max} will have a similar effect. E_{max} is the maximum energy allowed for any grid point. All points that have energies higher than E_{max} are levelled to the E_{max} value. Therefore points on the grid map with high energies, and also points which are very close to the protein atoms, will have the energy value of

E_{max}. Reducing E_{max} allows for a "softer" docking approach- therefore any orientations that are very close to the protein will have, with a lowered E_{max}, a lower overall energy. Using this technique, orientations which would previously have had higher energies with the higher E_{max} will now have lower energies. This would allow for the existence of orientations in the population that would normally be excluded, i.e. orientations that are close to the "true" genes. Having chromosomes in the population which are nearer to the crystal structure's orientation increases the probability for these chromosomes to produce good solutions - through mutation and/or crossover. Therefore reducing E_{max} gives the algorithm scope to explore a larger variety of orientations, some of which are more likely to be near to the crystal structure orientation. Because of the "softness" of this approach, there is the possibility that the fittest chromosomes in the population have orientations which clash with the protein. To prevent this from happening, the E_{max} value is, after a certain number of generations, returned to the default value of 5.0 kcal/mol and the entire population is rescored. This will increase the energies of the clashing orientations, giving them higher Pareto ranks and decreasing their ability of surviving into future generations, which will eventually weed them out of the population. By this point chromosomes closer to the crystal structure and not clashing with the protein would already have been created. These are not removed when E_{max} is increased, and their good Pareto ranks will enable them to survive in the population, and to produce offspring which are good solutions.

Liggrid allows the user to change the default value of E_{max} (5.0 kcal/mol), so one way of implementing this is to generate two sets of probe map files, one with a low E_{max} and one with the default E_{max}. This would reduce the flexibility of the algorithm because experimenting with different E_{max} values would mean that liggrid would need to be run every time to generate new probe map files. Another option is to read in the probe map files as produced by liggrid using the default E_{max} value, and modifying these values intrinsically, so that they correspond to a lowered E_{max} value. This would also provide the option to change E_{max} several times during the run, without having had to generate different probe map files previously. Hence this is the method that was implemented.

7.3 Reducing E_{max}

Because the second option described above provides the algorithm with flexibility, it was selected for implementation. The probe files (created with the default E_{max} of 5.0 kcal/mol) are read in as normal, and their energy values are then adjusted to a lower E_{max} value. The reduced E_{max} that has been selected is -2.0 kcal/mol. This threshold was selected because it was found to be the most effective value that avoided the generation of chromosomes with excessively artificial low energies. Lower E_{max} values produce too many such chromosomes the majority of which are quickly selected out of the population when E_{max} is reverted to 5.0 kcal/mol. Equally a higher value of E_{max} did not result in the desired effect of "softening" the energy landscape. Therefore all energy values that are 5.0 kcal/mol are converted to -2.0 kcal/mol. To maintain the energy gradient, values between 5.0 and -3.0 kcal/mol are scaled as follows:

Figure 7-6 (a) represents energy values where the value of Emax is 5.0 kcal/mol; (b) shows how these are altered when E_{max} is reduced to -2.0 kcal/mol.

The adjustments of the energy values are calculated using the following equations:

if $E_{old} < -3.0$;

$$E_{new} = E_{old} \qquad \text{Equation 7.3}$$

if $E_{old} > -3.0$;

$$E_{new} = -3 + 1/8\,(E_{old} + 3) \qquad \text{Equation 7.4}$$

Where E_{old} is the energy value as input from a probe map file, and E_{new} is the energy adjusted based on the scale shown in Figure 7.6. This adjustment ensures the maintenance of the energy gradient.

7.3.1 Effect of changing E_{max}

To view the effect of E_{max} manipulation, distributions of the NSGA-II populations were observed within the protein binding site relative to the crystal structure. Figure 7.7 shows the population at 500 generations when docking 1hdc, and where the E_{max} value is set to its default (5.0 kcal/mol). The figure clearly indicates that the population is concentrated at a different part of the binding site, meaning that the algorithm is not exploring the correct area of the search space. Only a chance event (which may be introduced by the mutation operator), may move a chromosome from that location closer towards the crystal structure. Figure 7.8 shows the distribution of the population with the introduction E_{max} manipulation, as well as the crystal structure. The population is now located much closer to the crystal structure's position. The genes of the chromosomes of this population are closer in representation to the crystal structure's position, therefore the genetic operators are more likely to yield solutions with orientations close to the crystal structure. This implementation of the algorithm, in combination with the local minimisation technique (described below) was tested on Datasets 1 and 2- and the results are shown in section 7.5.

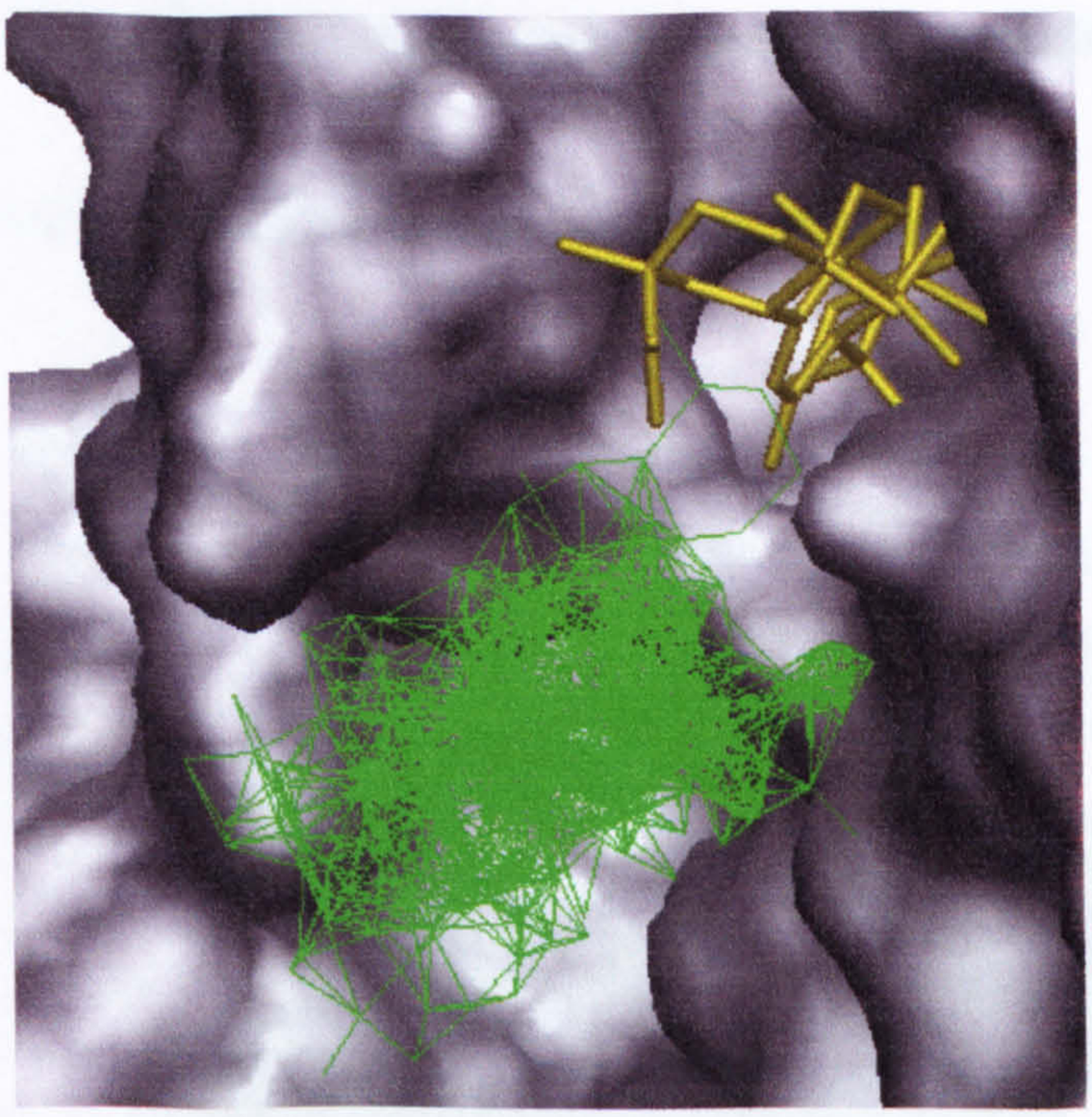


Figure 7-7 Solutions from NSGA-II (green line figures) in binding site of 1hdc at 500 generations when E_{max} is set at 5.0 kcal/mol. The solutions are occupying a region away from the crystal structure's (yellow cylinder) location in the binding site.

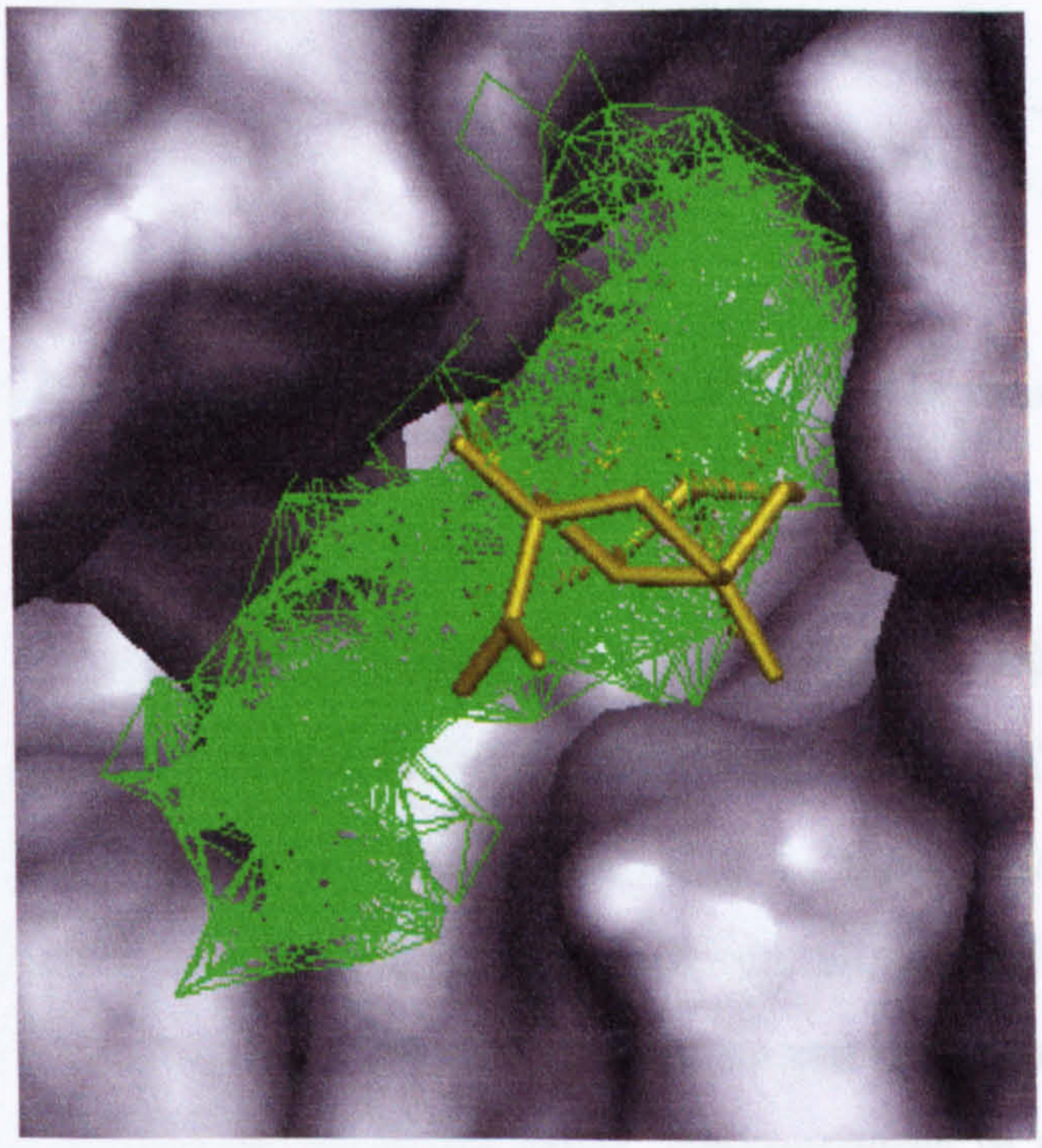


Figure 7-8: Solutions from NSGA-II (green line figures) in binding site of 1hdc at 500 generations when E_{max} is lowered. The solutions are at the location of the crystal structure (yellow cylinder) in the binding site, indicating that the algorithm is, at this stage, exploring the correct region in the search space.

7.4 Downhill simplex minimisation in multidimensions

7.4.1 Distribution of initial population

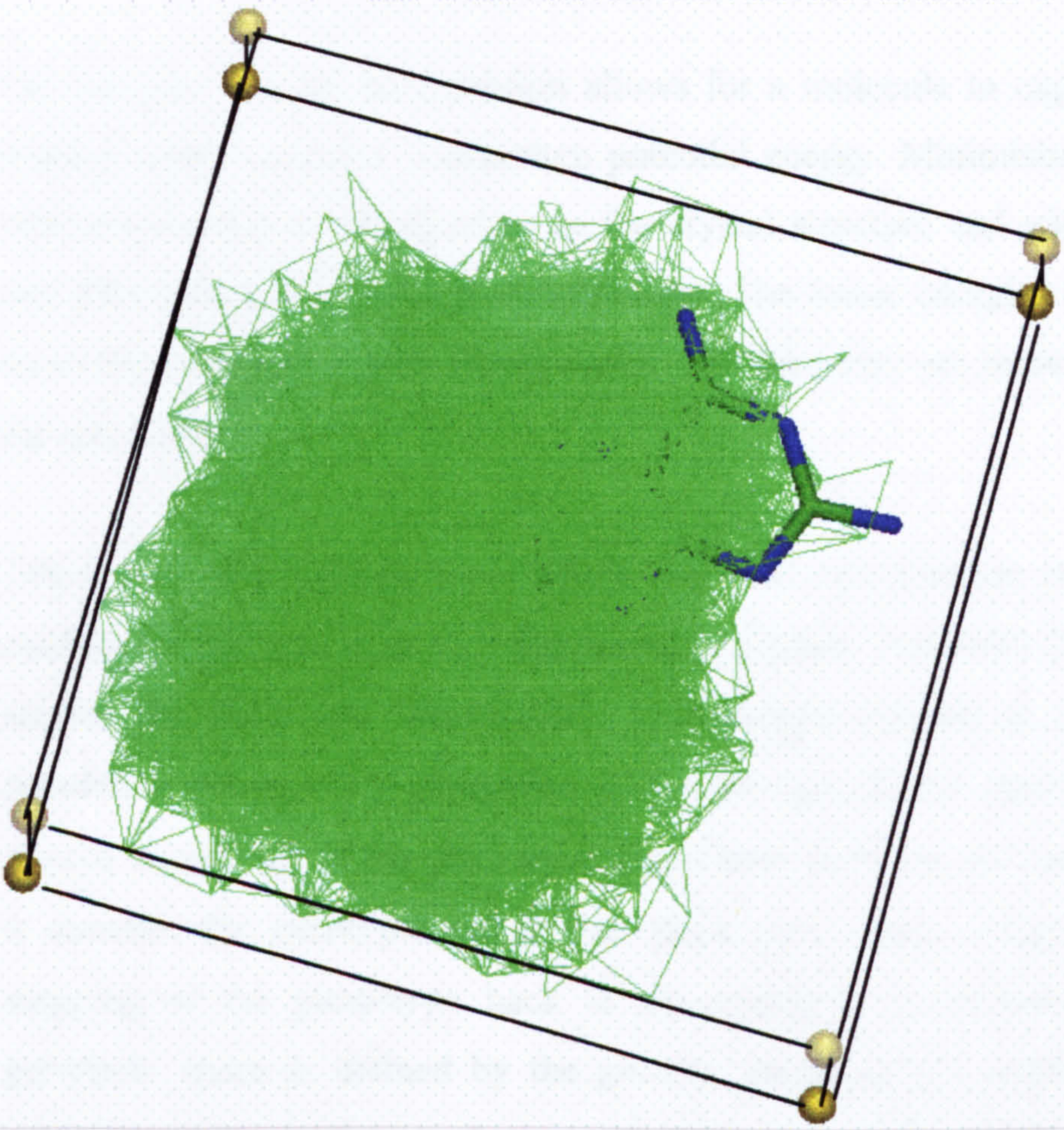


Figure 7-9 Distribution of initial population in relation to the GRID box.

Looking at the initial population in Figure 7.9, it can be seen that this is evenly distributed throughout the GRID box. A few of the chromosomes are similar to the crystal structure in orientation (i.e. they have low rmsd values). After a few generations these chromosomes are eliminated from the population. Despite their relatively good orientations, they do not have good Pareto ranks. This is because they are being outranked by other chromosomes in the population, which are occupying regions that are more energetically favourable. Therefore they are very early on eliminated from the population. The elimination of these orientations results in the loss of valuable information encoded by the chromosomes, which has the potential of producing offspring with good Pareto ranks and rmsds close to the crystal structure.

7.4.2 Downhill simplex minimization

The process of local minimisation allows for a molecule to explore the degrees of freedom until it reaches a minimum potential energy. Minimising the chromosomes with orientations relatively close to the crystal structure and with low Pareto ranks will allow them to reach a local minimum. The lower energies attained would give these chromosomes higher Pareto ranks, therefore they are more likely to survive in the population.

There are a few docking tools which use local minimisation techniques to aid the conformational search, or to refine solutions further. Autodock (Morris *et al.*, 1998) applies the Solis and Wets (1981) local search method to a proportion of the population during every generation. The advantage of this method is that it does not require knowledge of the gradient at a particular point on the energy landscape. Also it searches the genotypic rather than phenotypic space, which avoids the inverse mapping of the phenotype back to the genotype. In protein-ligand docking the genotypic space is defined by the genetic operators, i.e. crossover, roulette wheel selection and mutation, and represented by the genes of the chromosomes. The phenotype of a chromosome is the atomic coordinates of the ligand pose as encoded by its genes, as well as the energy score of that ligand orientation. However, in those cases where an inverse mapping function exists, i.e., one which yields a genotype from a given phenotype, it is possible to finish a local search by replacing the individual with the result of the local search.

The Lamarckian element of Autodock (section 3.4.3.5) ensures that chromosomes resulting from a local search are incorporated back into the population, and passed to the next generation.

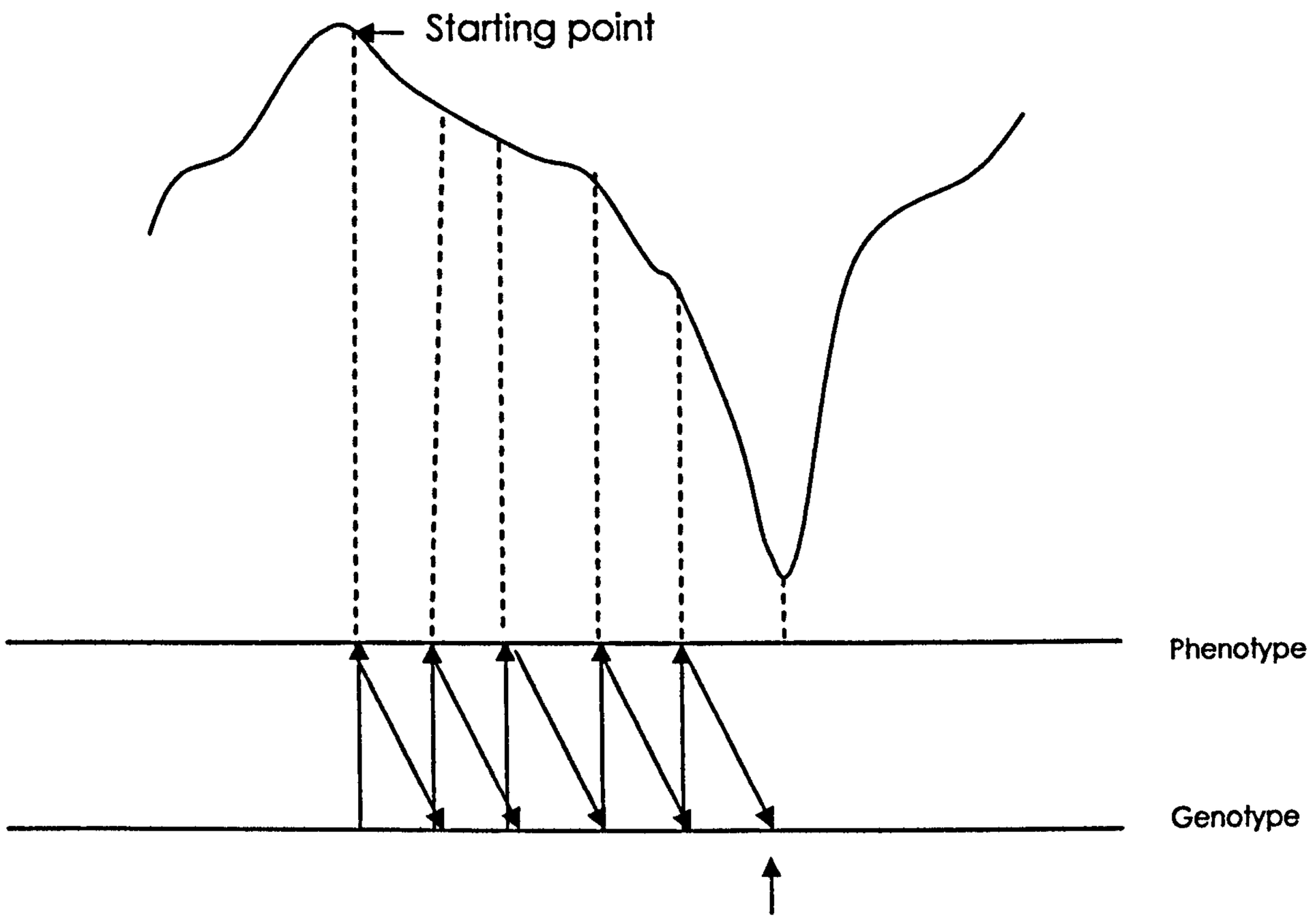


Figure 7-10 Effect of local minimisation as implemented by NSGA-II. The amoeba method (section 7.4.3) uses the genes representing the energy at the starting point to construct the initial simplex. The method then drives the simplex down the slope, taking steps in genotypic space, and assessing the quality of each step by calculating the energy at that point (the phenotype). The resulting energy value determines what direction the simplex should take next. The changes from genotype to phenotype during the minimisation procedure are shown as arrows. The Lamarckian element ensures that the final set of genes at the point of the minimised energy are passed into the next generation (adapted from Morris *et al.*, 1998).

Q-fit (Jackson, 2002) also incorporates a minimisation step in its procedure. It uses the downhill Simplex algorithm of Nealder and Mead (Nelder and Mead, 1965) for rigid-body minimisation of every fragment that is placed in the binding site. This method, like Autodock's Solis and Wets method, also does not require gradient information. It was also found to be fast when compared to a steepest descent method (Jackson *et al.*, 1998), which incidentally also requires gradient information.

7.4.3 Implementation of energy minimisation in the NSGA-II

The minimisation procedure that has been implemented for the NSGA-II combines factors from both Autodock and Q-fit. The minimisation method used is that of Nelder and Mead (1965), and a Lamarckian element ensures that minimised chromosomes are inherited by future generations.

The Nelder and Mead method, like the Solis and Wets method, is able to search genotypic space. This is advantageous to a phenotypic search because the latter would involve mapping the genes of a chromosome to its phenotype (i.e. to its coordinates and energy score), minimising these, and inverse mapping the resulting coordinates into their corresponding genes. A genotypic space search will directly minimise the genes of a chromosome, thus avoiding the need for inverse mapping (a time-consuming step).

The NSGA-II uses the *amoeba* method (Press *et al.*, 1992), to implement the Nelder and Mead minimisation technique. The theory behind this technique is based on the generation of a simplex, a geometric figure represented, in N dimensions by $N+1$ vertices which are all interconnected by lines. N represents the number of degrees of freedom encompassing the problem, which in this case is six (the three translational and three rotational degrees of freedom). The generated simplex then undergoes various transformations (Figure 7.11). A reflection moves the highest point of the simplex (this is the vertex with the highest energy) through the opposite face to a lower point (or energy). Taking bigger steps results in the expansion of the simplex. If the simplex is trying to go through a "tight" spot (or the eye of a needle), it can contract itself in all directions, so that it squeezes itself out. Contracting in one direction helps it to ooze out of a valley. This behaviour is likened to that of an amoeba, which explains the naming of the function.

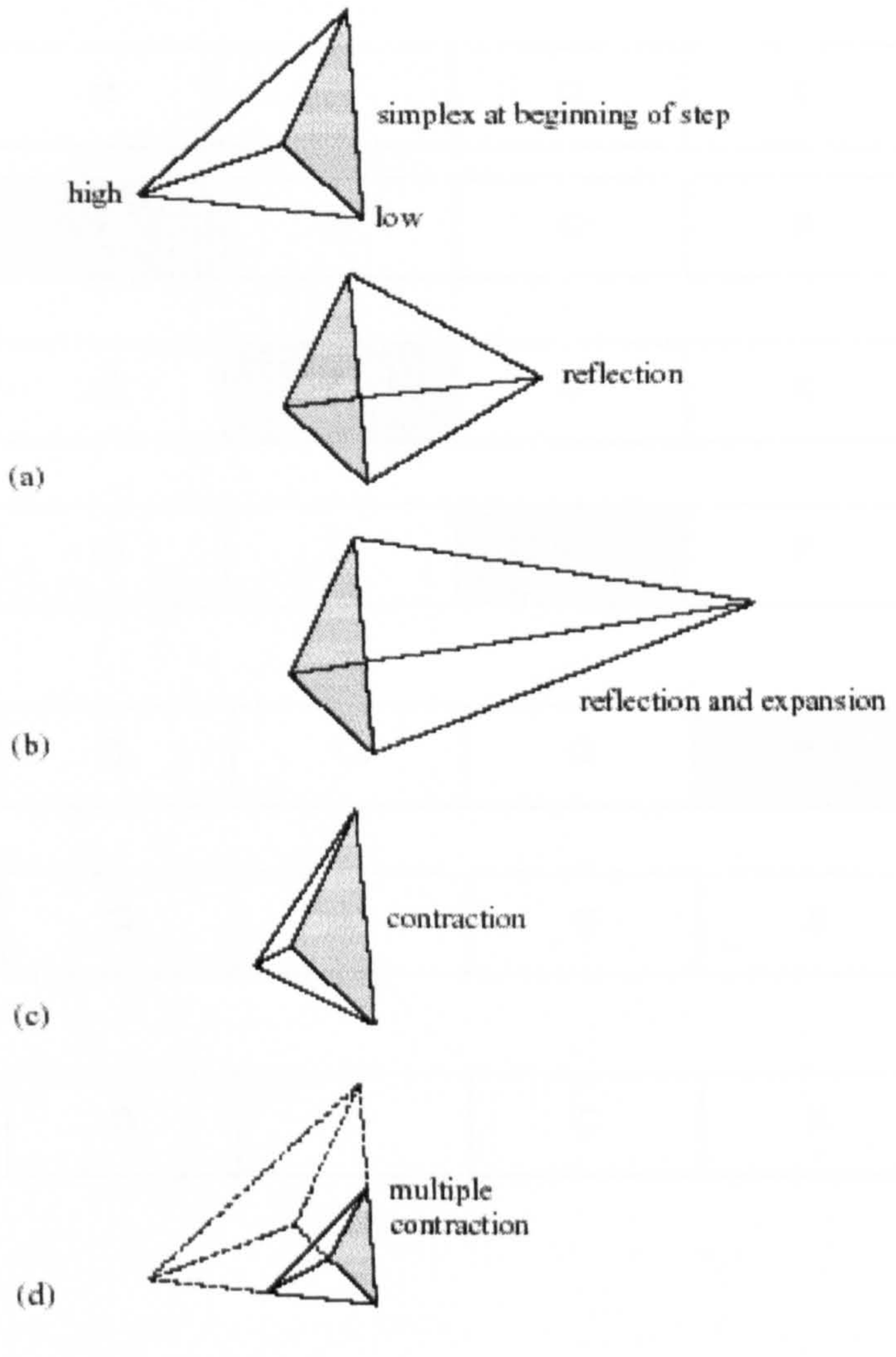


Figure 7-11 Transformations undertaken by the simplex generated by the local minimisation feature (Press *et al.*, 1992)

7.4.4 Simplex generation

The simplex consists, as mentioned above, of seven vertices (total number of degrees of freedom +1). Each vertex or point makes up a possible solution for the problem. The starting vertex is represented by the genes of the chromosome being minimised. The remaining six vertices consist of modified versions of the chromosome at the starting vertex. These are modified by changing a single gene of each by a factor λ. The magnitude of λ is an estimate of the scale of the problem, and different values for λ were be used for different genes. Therefore for changing any of the translation genes, a λ value of 0.4 Å was found most effective (λ_1). For the rotational genes a value of 0.087266 rads (equivalent to 5°) was used (λ_2). This is shown in Figure 7.12.

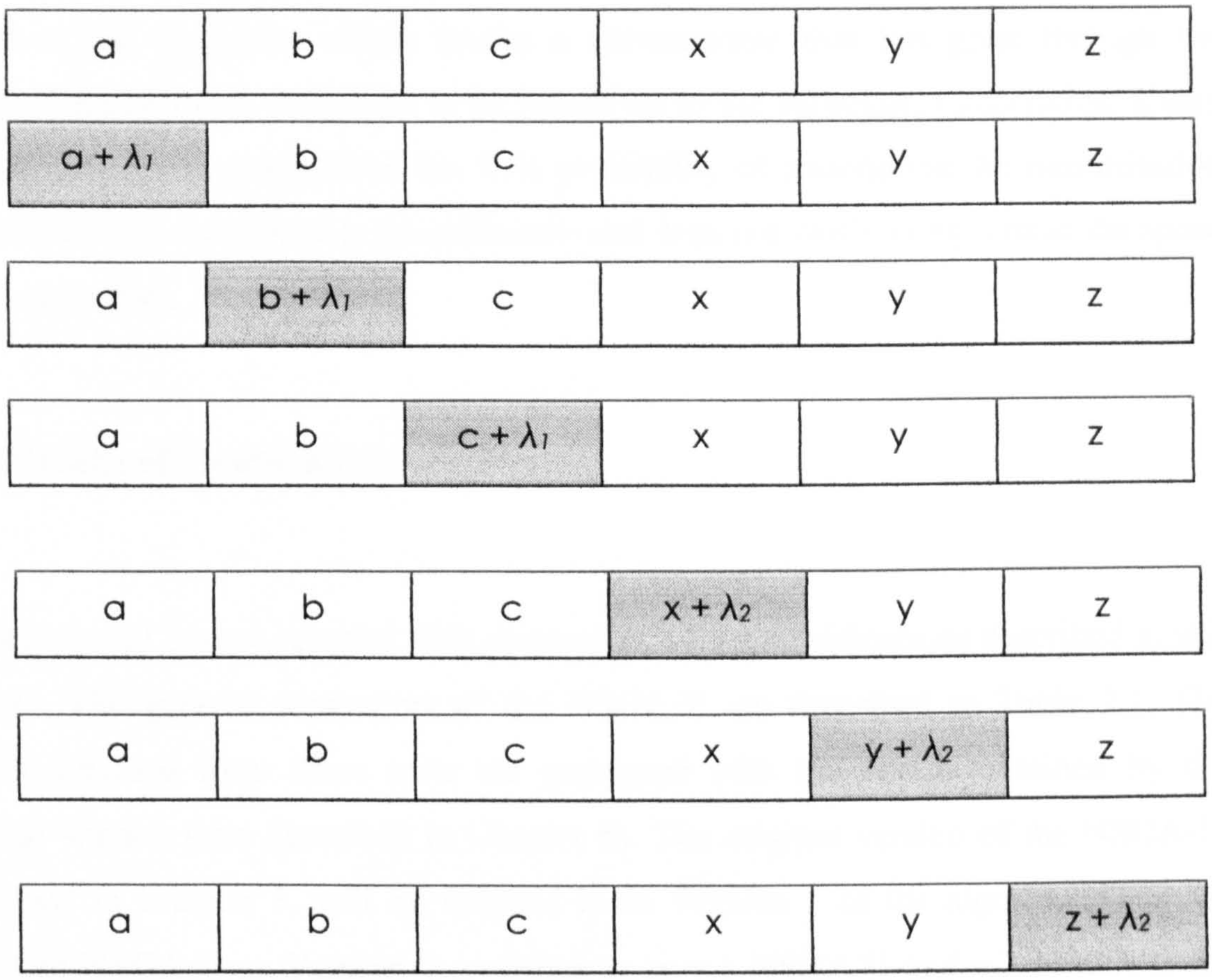


Figure 7-12 Chromosomes representing the seven vertices of the simplex.

The termination criterion is reached when a certain move of the simplex results in a functional distance that is fractionally smaller than a particular threshold. The threshold which has been used here is 0.5.

Because the local minimisation technique attempts to decrease the total energy of a solution using small changes in orientation (the chromosomes of a simplex only have fractional changes in their genes), then it is hoped that solutions in the early generations which are close to the crystal structure- but have relatively high Pareto ranks, will retain their overall good orientations, but will have their energies minimised by the local minimisation technique, increasing their Pareto rank within the population and increasing the probability that they remain in future generations.

The Lamarckian element which has been implemented by the NSGA-II refers to the aspect of the algorithm which allows a chromosome that has gone through the simplex minimisation procedure to be passed on to the subsequent generation. Every chromosome in the population has 10% probability of undergoing the minimisation procedure. This was found to be sufficient- and does not vastly compromise the speed of the algorithm.

7.5 Results of Modifications

Datasets 1 and 2 were retested with some or all of the modifications described in this chapter. The general parameters of the NSGA-II are described in Table 7.1. The results obtained from these tests are compared with the results obtained by the original version (and described in Chapter 6). The original version of the NSGA-II, described in Chapter 5, will be referred to as Version 1 of the algorithm, and the modified NSGA-II as Version 2 (referred to as v.1 NSGA-II and v.2 NSGA-II for brevity). To ease the comparison between the two versions, the Pareto fronts from both are plotted on the same axes. The Pareto solutions from Version 2 are clustered, based on orientation, into groups which are within 1 Å of each other. All of these are shown as different coloured triangles. The lowest rmsd of a solution in a given group is quoted in the legend key. The Pareto solutions from Version 1 are shown in two groups, solutions with rmsds below 2.0 Å (shown as grey circles), and solutions with rmsds higher than 2.0 Å (red circles). The next sections show the results for both datasets.

Generation number	15000
Population size	200
Niche radius	0.5 Å
Mutation parameters	
Mutation rate	30%
Rotation step size	2π rads
Translation step size	2 Å

Table 7.1 Parameters used in modified NSGA-II.

7.5.1 Dataset 1

Dataset 1 is described in Chapter 6, and consists of complexes that are considered to be fairly easy to dock. v.1 NSGA-II successfully docked eight out of the ten complexes, and failed with 4dfr and 2phh. Figures 7.13 to 7.23 show the Pareto solutions obtained when docking the different molecules. The following section compares the results obtained from the two versions.

7.5.1.1 1dbb

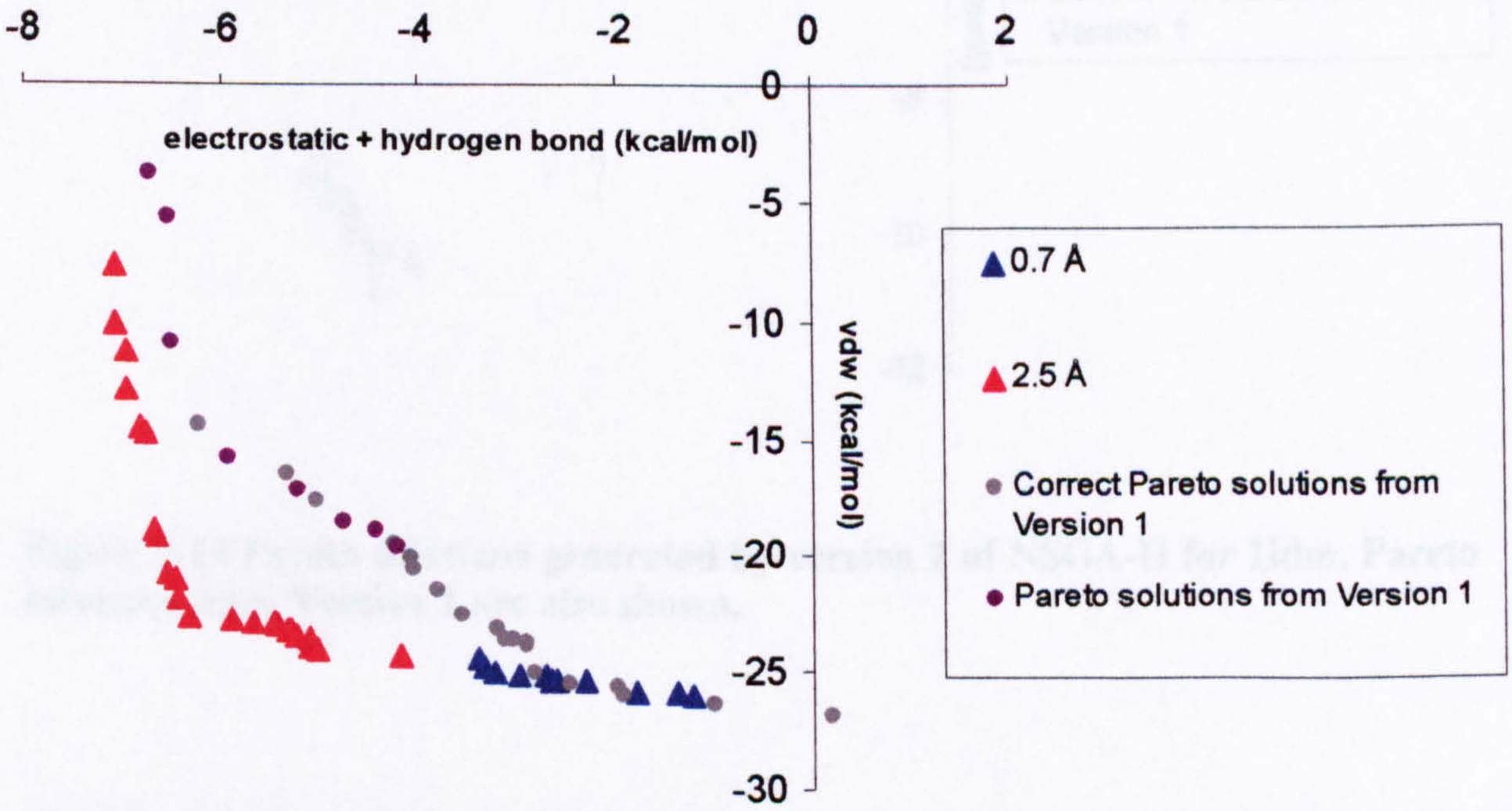


Figure 7-13 Pareto solutions generated by version 2 of NSGA-II for 1dbb. Pareto solutions from Version 1 are also shown.

Both versions of the NSGA-II were successful in docking 1dbb (Figure 7.13). Looking at the position of the correct Pareto solutions from both versions, it can be seen that both have converged to similar points. Overall, the Pareto front from Version 2 is clearly more advanced than the Version 1 Pareto front. Also, comparing the number of clusters generated by both versions, fewer clusters are generated by v.2 NSGA-II. This may be because at the more advanced position of the Version 2 Pareto front, the Pareto solutions are more similar to each other orientationally- resulting in fewer clusters.

7.5.1.2 1ldm, 2gbp and 1stp

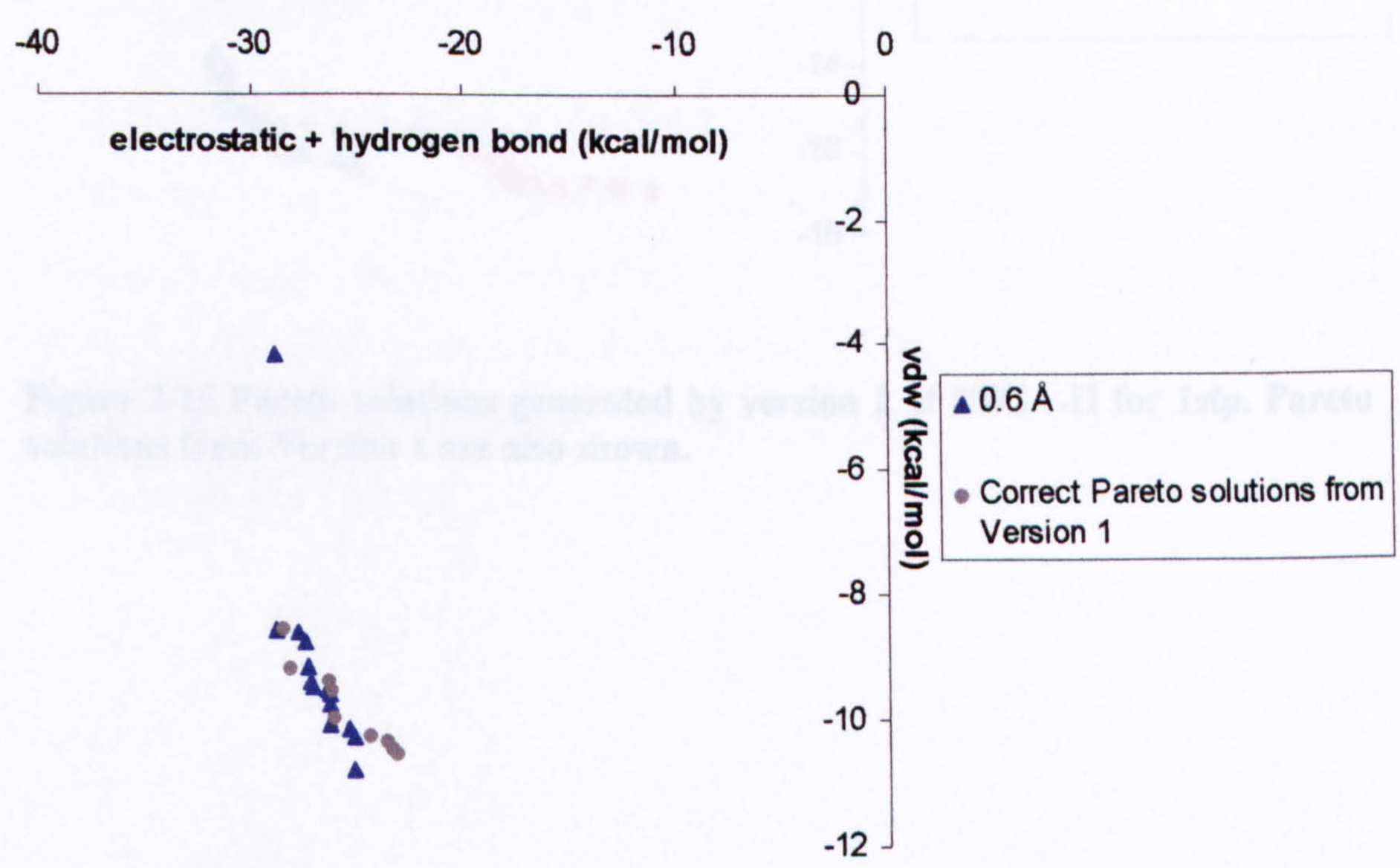


Figure 7-14 Pareto solutions generated by version 2 of NSGA-II for 1ldm. Pareto solutions from Version 1 are also shown.

All of the Pareto solutions of 1ldm from Version 2 have good rmsds of approximately 0.6 Å, which fall into one cluster (Figure 7.14). Some of these solutions are slightly

more advanced than the Version 1 Pareto solutions. The Version 2 solutions also fall into a single cluster, with a higher rmsd of 1.1 Å.

The Pareto solutions of 1stp fall into two clusters, which have approximate rmsds of 0.7 Å and 3.4 Å (Figure 7.15). This Pareto front is sparser than that obtained with NSGA-II Version 1.

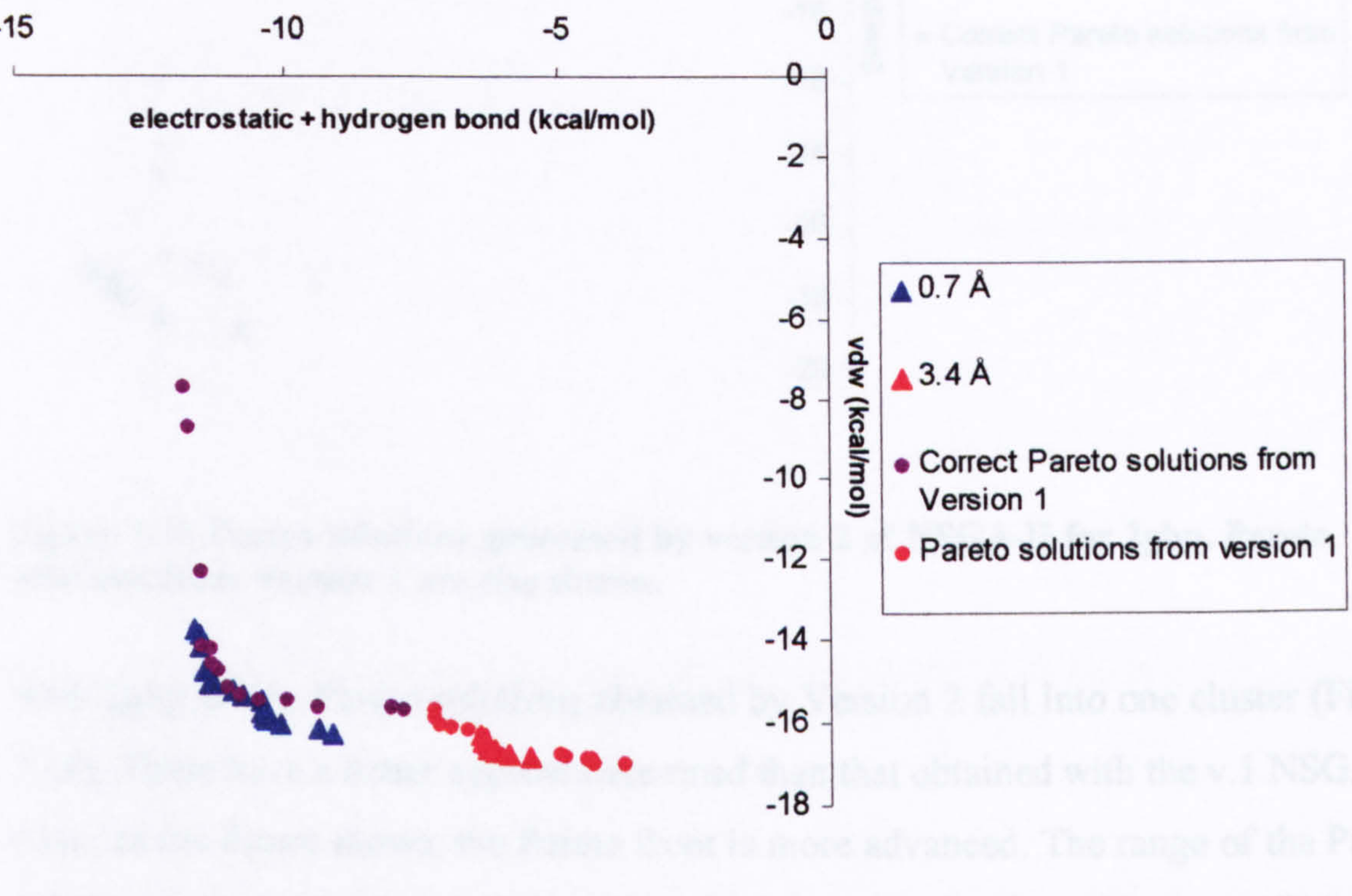


Figure 7-15 Pareto solutions generated by version 2 of NSGA-II for 1stp. Pareto solutions from Version 1 are also shown.

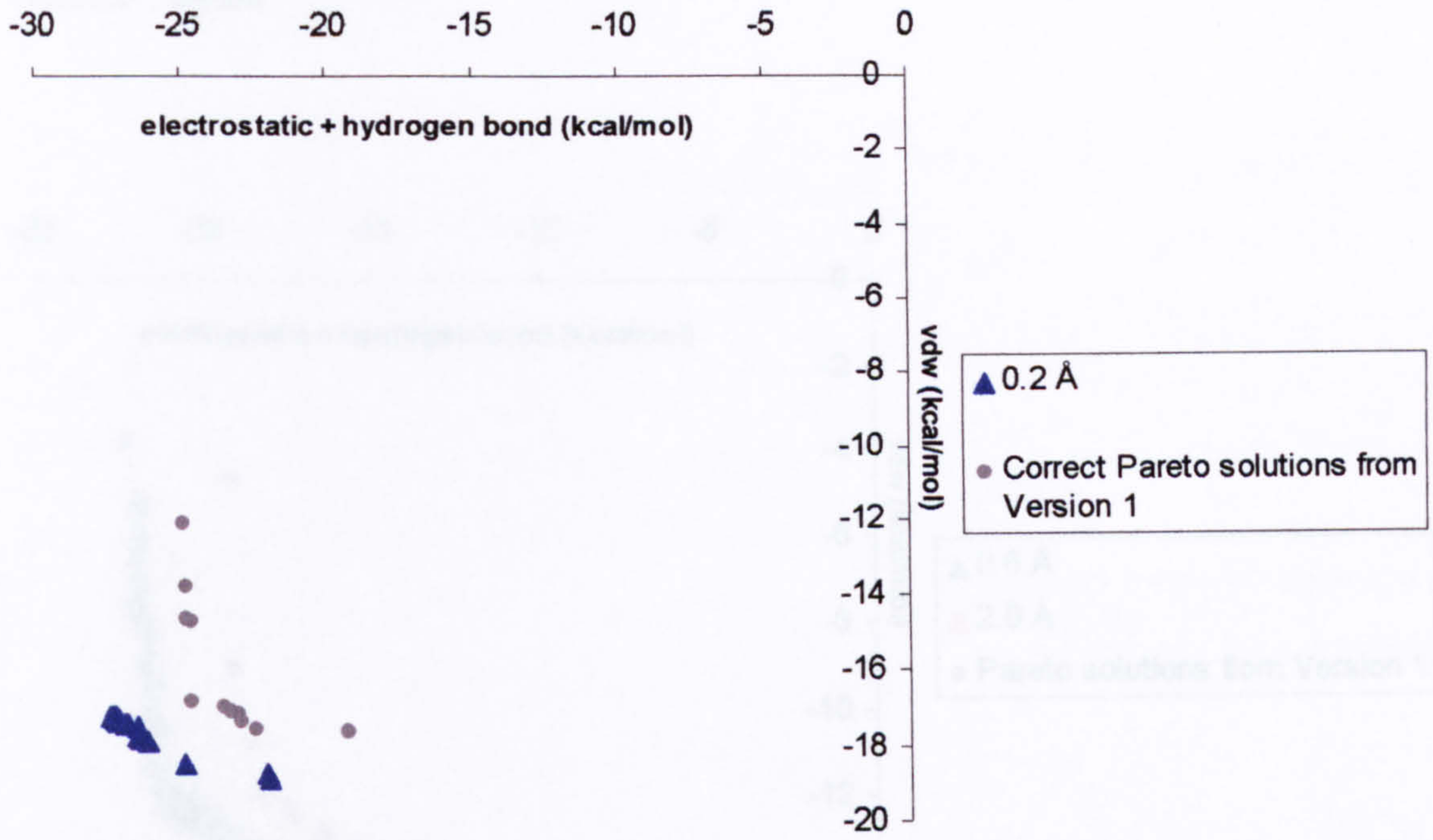


Figure 7-16 Pareto solutions generated by version 2 of NSGA-II for 2gbp. Pareto solutions from Version 1 are also shown.

With 2gbp all the Pareto solutions obtained by Version 2 fall into one cluster (Figure 7.16). These have a better approximate rmsd than that obtained with the v.1 NSGA-II. Also, as the figure shows, the Pareto front is more advanced. The range of the Pareto solutions' vdw energies is much smaller than that obtained from Version 1. This may be because at that particular point in objective space all the Pareto solutions have a smaller vdw energy range.

7.5.1.3 2phh

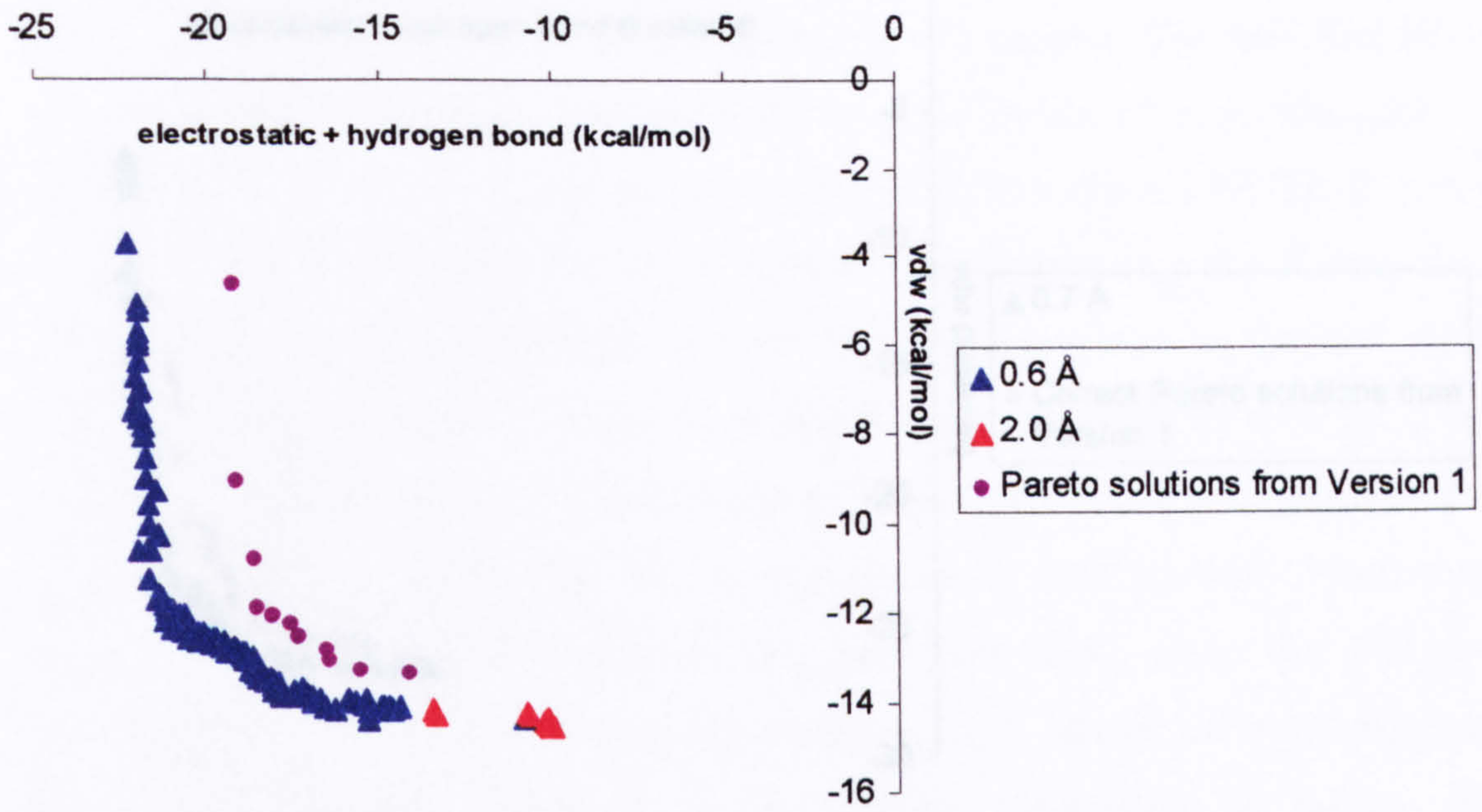


Figure 7-17 Pareto solutions generated by version 2 of NSGA-II for 2phh. Pareto solutions from Version 1 are also shown.

All of the Pareto solutions obtained for 2phh have good rmsds (Figure 7.17). This is in contrast to the result obtained with the initial NSGA-II, where the best solutions had rmsds of ~ 4 Å. The Pareto front has advanced further than that obtained with the v.1 NSGA-II, which implies that v.1 NSGA-II was unsuccessful because the algorithm had not converged to the true Pareto front.

7.5.1.4 3tpi and 4dfr

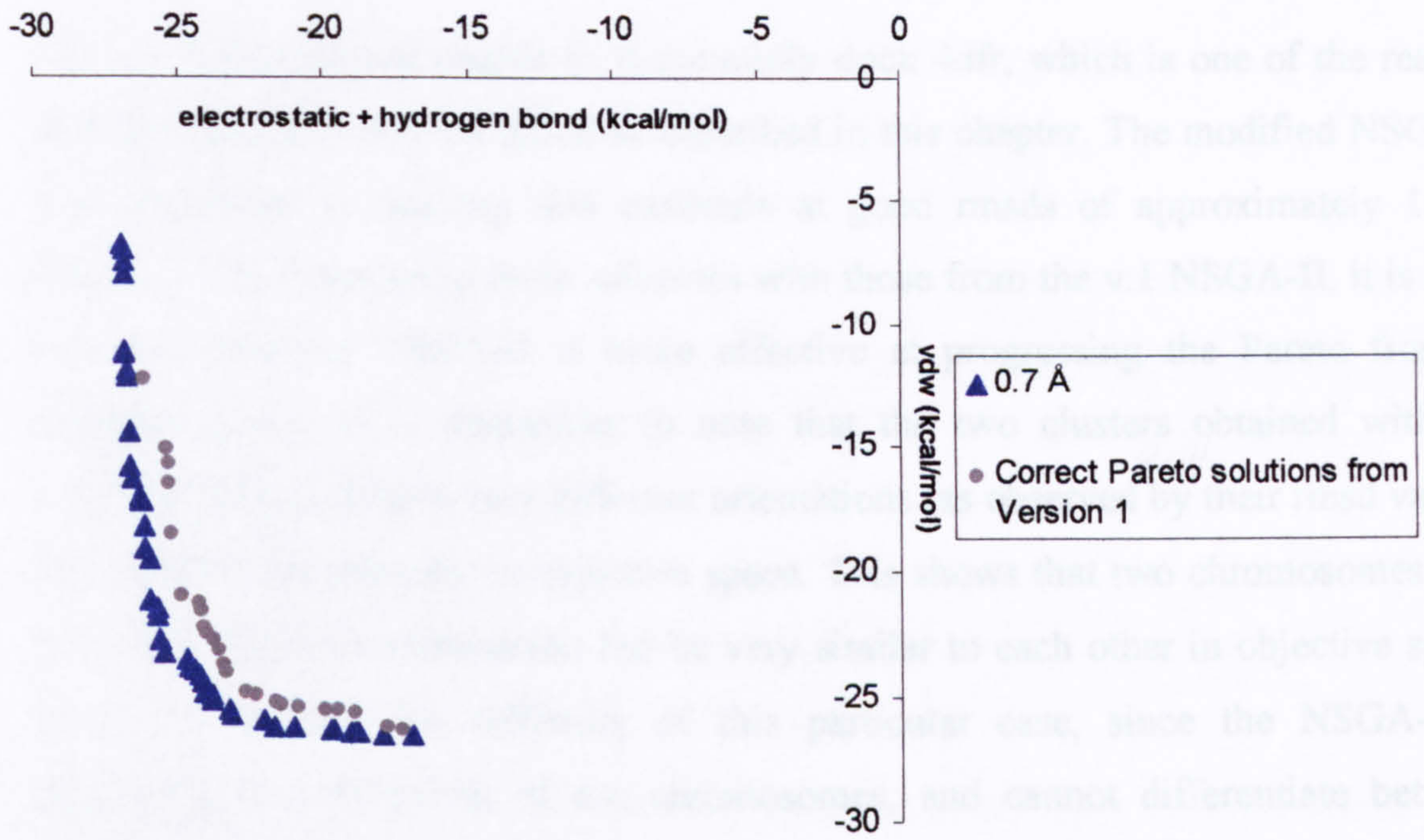


Figure 7-18 Pareto solutions generated by version 2 of NSGA-II for 3tpi. Pareto solutions from Version 1 are also shown.

For 3tpi, the Pareto front generated by the modified NSGA-II is slightly more advanced than that obtained with v.1 NSGA-II (Figre 7.18). Correct solutions were obtained in both cases.

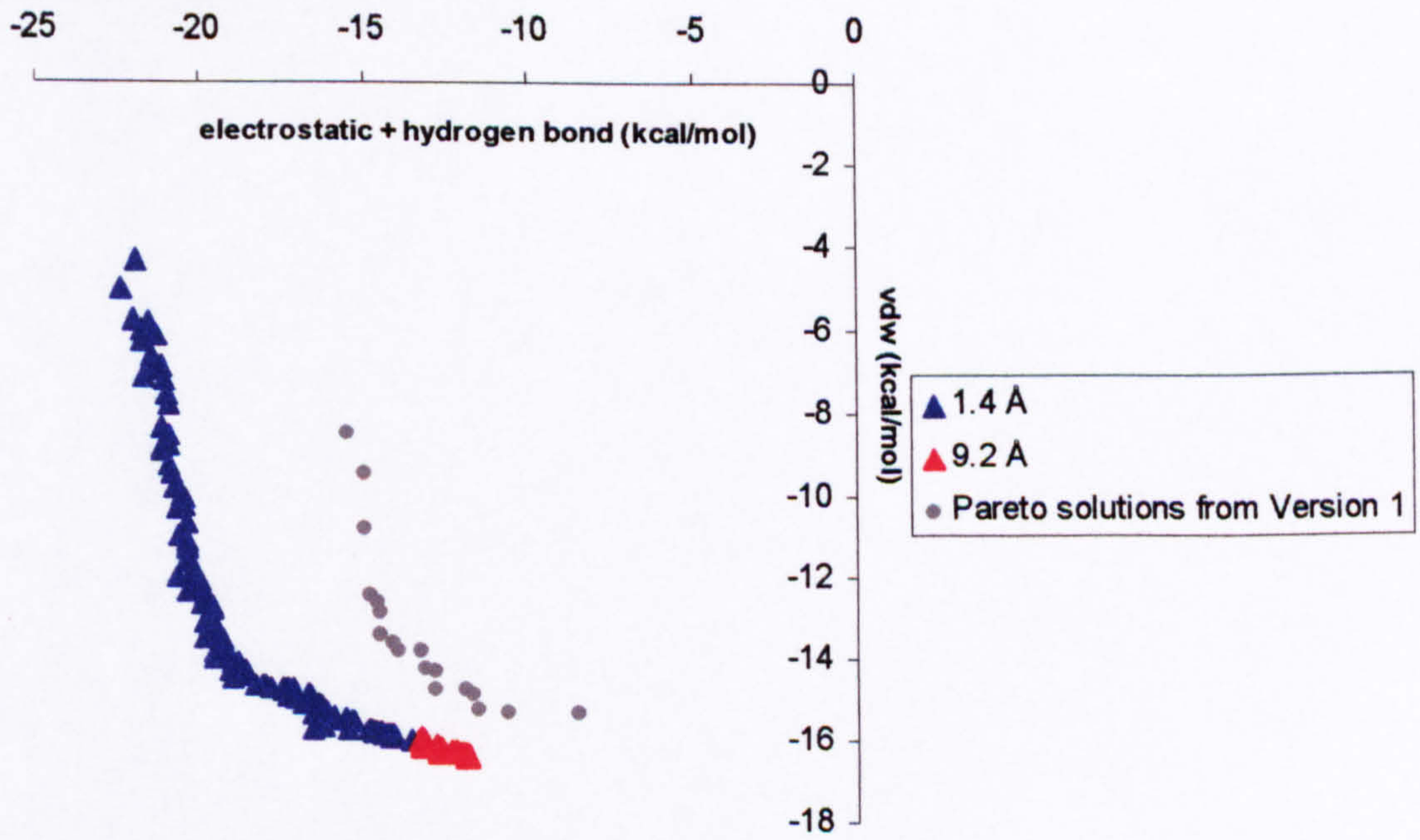


Figure 7-19 Pareto solutions generated by version 2 of NSGA-II for 4dfr. Pareto solutions from Version 1 are also shown.

The v.1 NSGA-II was unable to successfully dock 4dfr, which is one of the reasons why the algorithm was modified as described in this chapter. The modified NSGA-II was successful in docking this molecule at good rmsds of approximately 1.4 Å (Figure 7.19). Comparing these solutions with those from the v.1 NSGA-II, it is clear that the modified NSGA-II is more effective at progressing the Pareto front in objective space. It is interesting to note that the two clusters obtained with the modified NSGA-II have very different orientations (as observed by their rmsd values) and yet they are adjacent in objective space. This shows that two chromosomes may have very different orientations but be very similar to each other in objective space. This demonstrates the difficulty of this particular case, since the NSGA-II is optimising the objectives of the chromosomes, and cannot differentiate between different orientations. Therefore if the "wrong" orientation is present in the population, then it will not be at a substantial disadvantage from a "correct" orientation with similar objectives. This problem can, hypothetically, be circumvented by using decision space niching (section 5.2.4), which applies niching to the orientations of the ligands rather than the objectives. In this way a diverse selection of orientations are maintained within the population, and therefore increase the probability of finding solutions that have both, low interaction energy types and correct orientations.

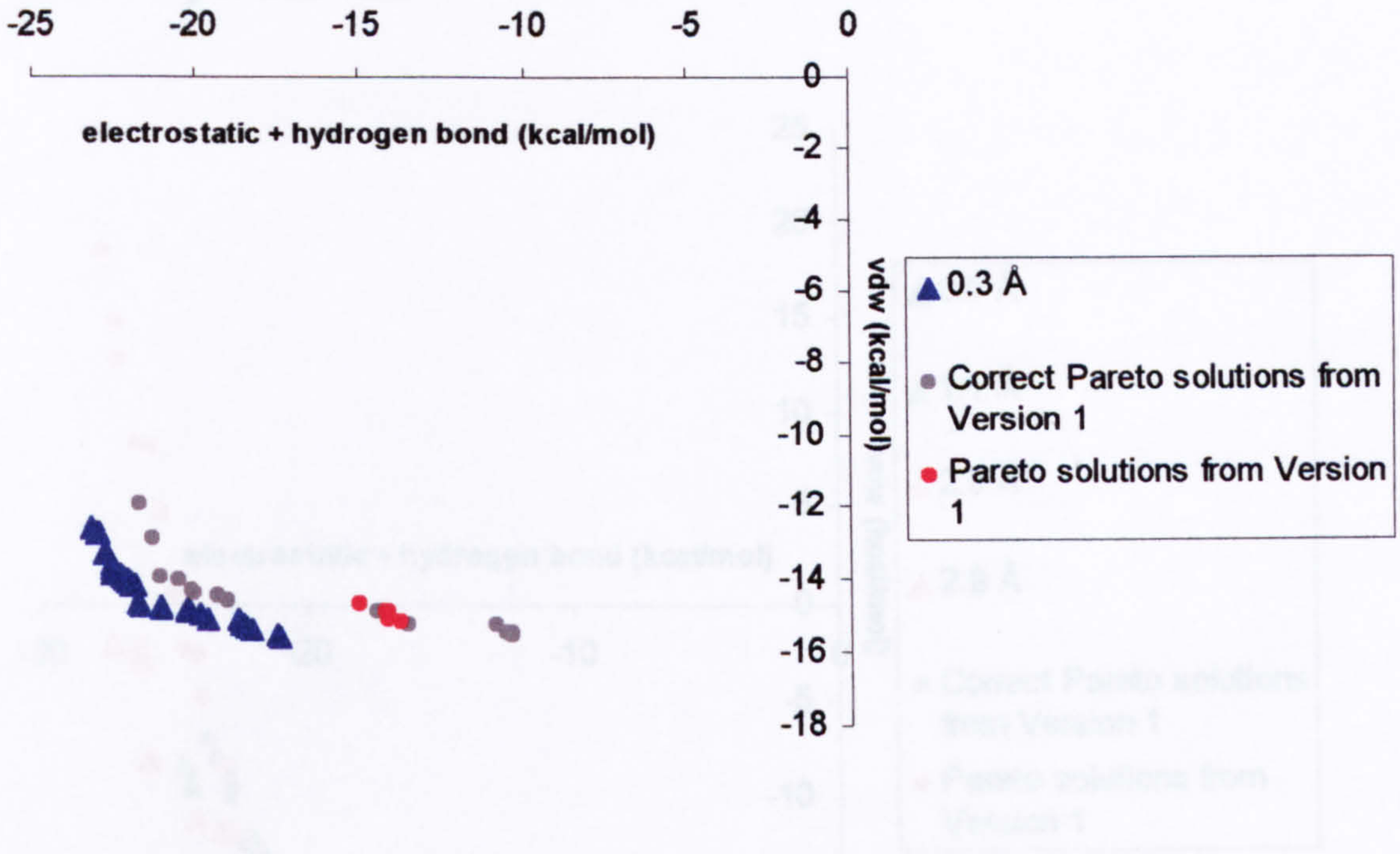


Figure 7-20 Pareto solutions generated by version 2 of NSGA-II for 1abe. Pareto solutions from Version 1 are also shown.

7.5.1.5 1abe

1abe was docked by both versions of the algorithm. The correct clusters in both Pareto fronts have good rmsds of approximately 0.3 Å. The Pareto front from Version 1 has a second cluster of 2.1 Å. This is not in the Pareto front from the second version because its Pareto solutions would dominate these clusters, which is why they would have been excluded from the Pareto set. The Version 2 Pareto front is slightly more advanced than the Pareto front from Version 1.

7.5.1.6 3ptb and 1ulb

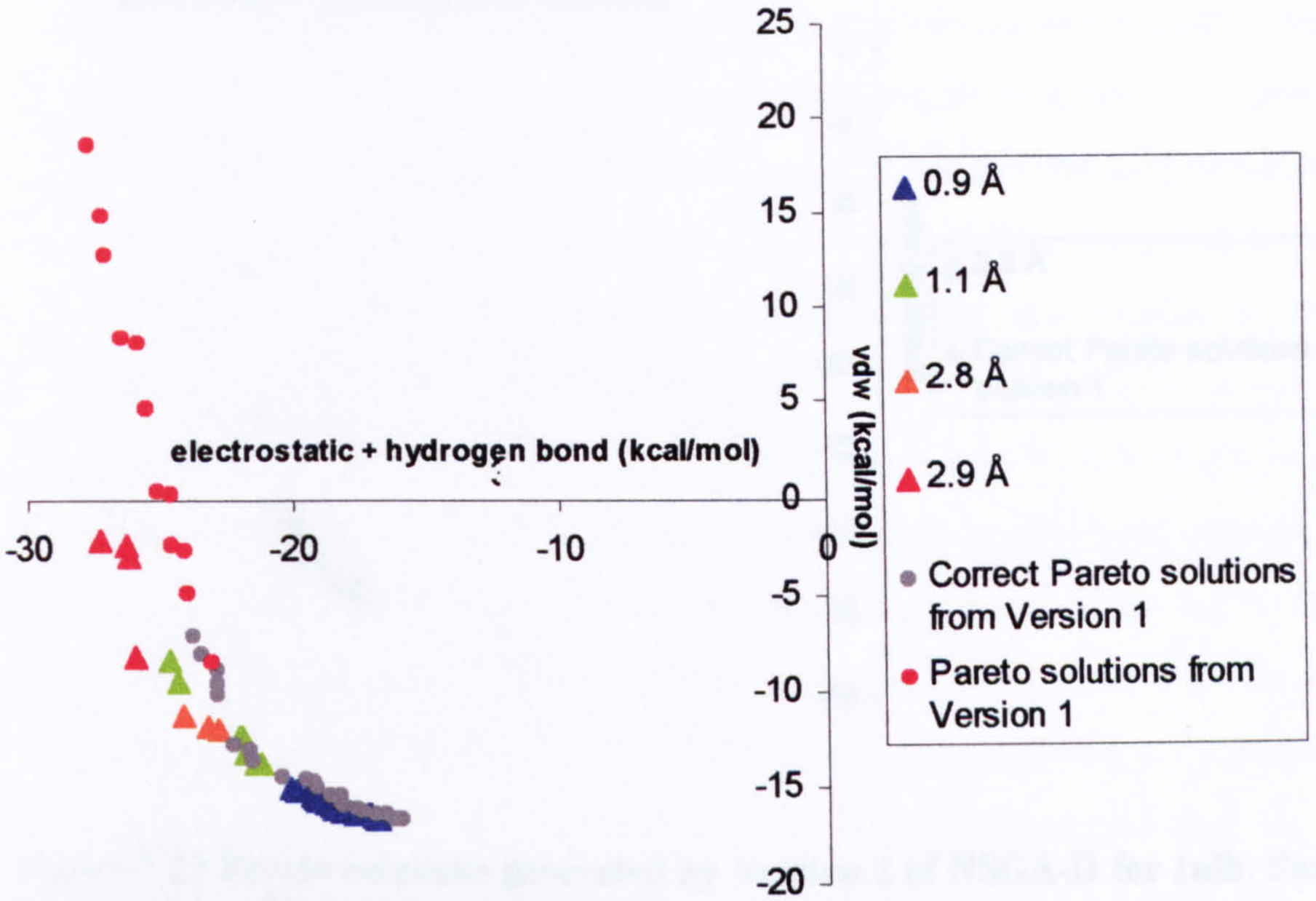


Figure 7-21 Pareto solutions generated by version 2 of NSGA-II for 3ptb. Pareto solutions from Version 1 are also shown.

3ptb was successfully docked by both versions of the NSGA-II, but the Pareto front obtained with v.2 NSGA-II is more advanced along the x axis, where some of the solutions have decreased electrostatic and hydrogen bond energies (Figure 7.21). This front, however, does not extend into the positive vdw energy space. The clusters with good rmsds from both fronts have comparable objective values.

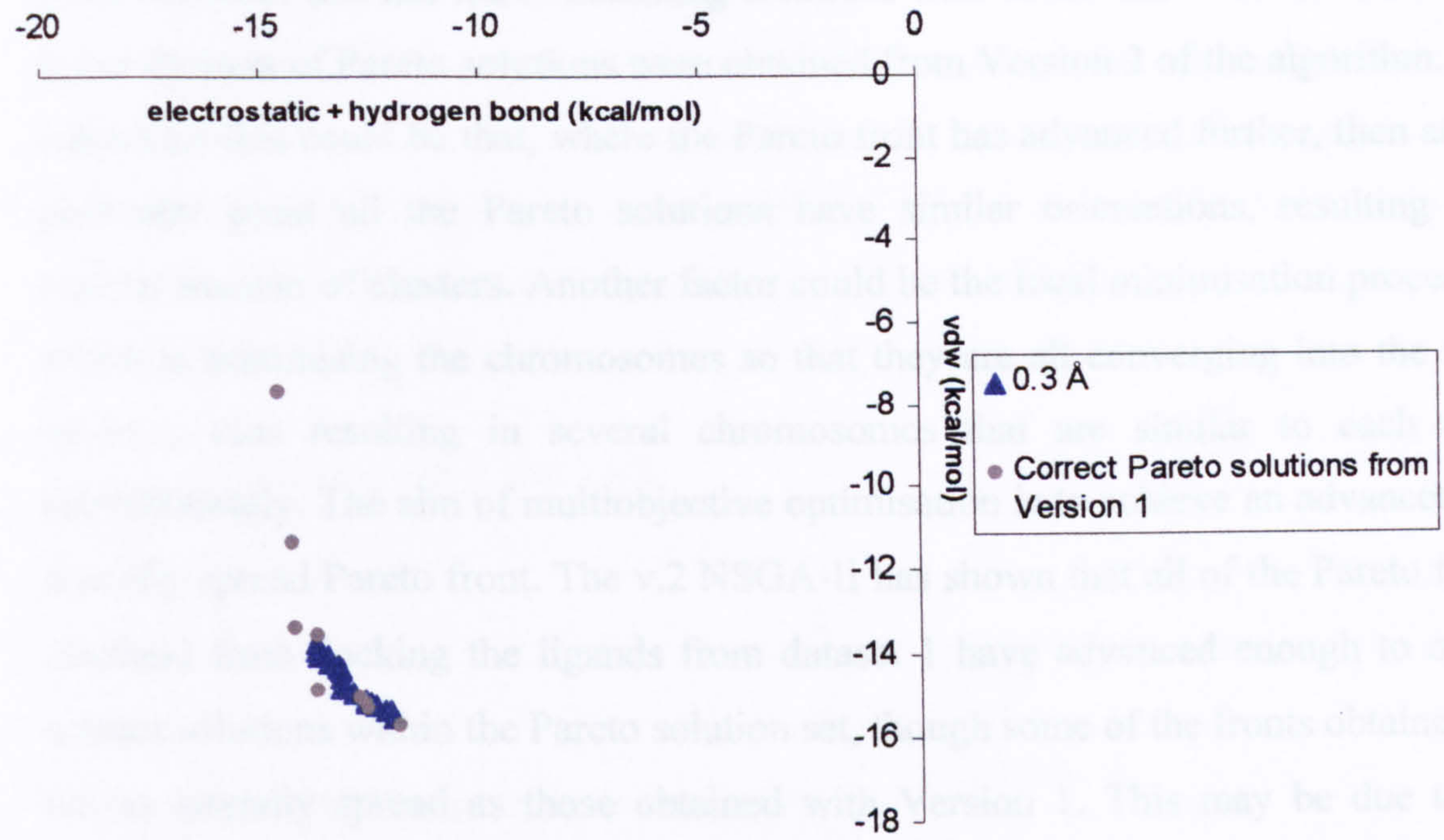


Figure 7-22 Pareto solutions generated by Version 2 of NSGA-II for 1ulb. Pareto solutions from Version 1 are also shown.

1ulb is one of the cases docked successfully by both versions of the NSGA-II. The Pareto front obtained with the v.2 NSGA-II is slightly less advanced along the x-axis: the lowest electrostatic and hydrogen bond energy for a solution is -13.8 kcal/mol. This is relative to -14.5 kcal/mol, the lowest electrostatic and hydrogen bond energy obtained by a Pareto solution from the v.1 NSGA-II (Figure 7.22). The extension of the Pareto front along the y-axis is similar for both algorithms. A single cluster was obtained with Version 2, and all of its solutions are similar to each other in terms of the two objectives. The vdw range of these solutions is also much narrower than those obtained with Version 1.

7.5.1.7 Summary of results obtained from Dataset 1

The NSGA-II v.2 described in this chapter has been successful in docking the entire Dataset 1, including 4dfr, the "model" docking test case. The Pareto fronts obtained have, in general, advanced further than those from Version 1. As demonstrated with

2phh and 4dfr, this has led to obtaining solutions with better rmsds. It was noted that fewer clusters of Pareto solutions were obtained from Version 2 of the algorithm. One reason for this could be that, where the Pareto front has advanced further, then at that particular point all the Pareto solutions have similar orientations, resulting in a smaller number of clusters. Another factor could be the local minimisation procedure, which is minimising the chromosomes so that they are all converging into the same minima, thus resulting in several chromosomes that are similar to each other orientationally. The aim of multiobjective optimisation is to achieve an advanced and laterally spread Pareto front. The v.2 NSGA-II has shown that all of the Pareto fronts obtained from docking the ligands from dataset 1 have advanced enough to obtain correct solutions within the Pareto solution set, though some of the fronts obtained are not as laterally spread as those obtained with Version 1. This may be due to the Lamarckian element of the algorithm- continually minimising a proportion of the population during a run may be producing several chromosomes that are similar to each other, therefore reducing the diversity of the population- which may result in a Pareto front that is less diverse- i.e. that is not as spread out laterally. The effect of the local minimisation/Lamarckian elements may be the faster convergence of the Pareto front (which means fewer generations are needed for running the algorithm) at the expense of a small loss in diversity. Despite the latter point the algorithm was successful in obtaining correct solutions within the Pareto solution set for all complexes of Dataset 1- an improvement from the initial version of the algorithm. By obtaining correct solutions it is possible to observe where these are falling in objective space, and therefore infers which of the objectives, if any, is dominating.

7.5.2 Dataset 2

Dataset 2, one of the two datasets used to validate Q-fit, was described in Chapter 6. This is a more problematic dataset- both Q-fit and the original GOLD docking tool found these complexes difficult to dock- obtaining erroneous results. The v.2 NSGA-II was retested on this dataset, and the results obtained were compared to the results from Version 1. The local minimisation procedure has not been implemented with this dataset because it was found that including this procedure had a detrimental

effect on the results, and failed to dock complexes which were previously docked successfully. The effect of the local minimisation procedure on this dataset was realised through various trial and error experiments, and this led to the decision not to include this procedure, though it is worth noting that Dataset 1 was successfully docked when this feature was implemented. The following section describes Dataset 2 results.

7.5.2.1 1acj, 1ack, 2ak3 and 1tdb

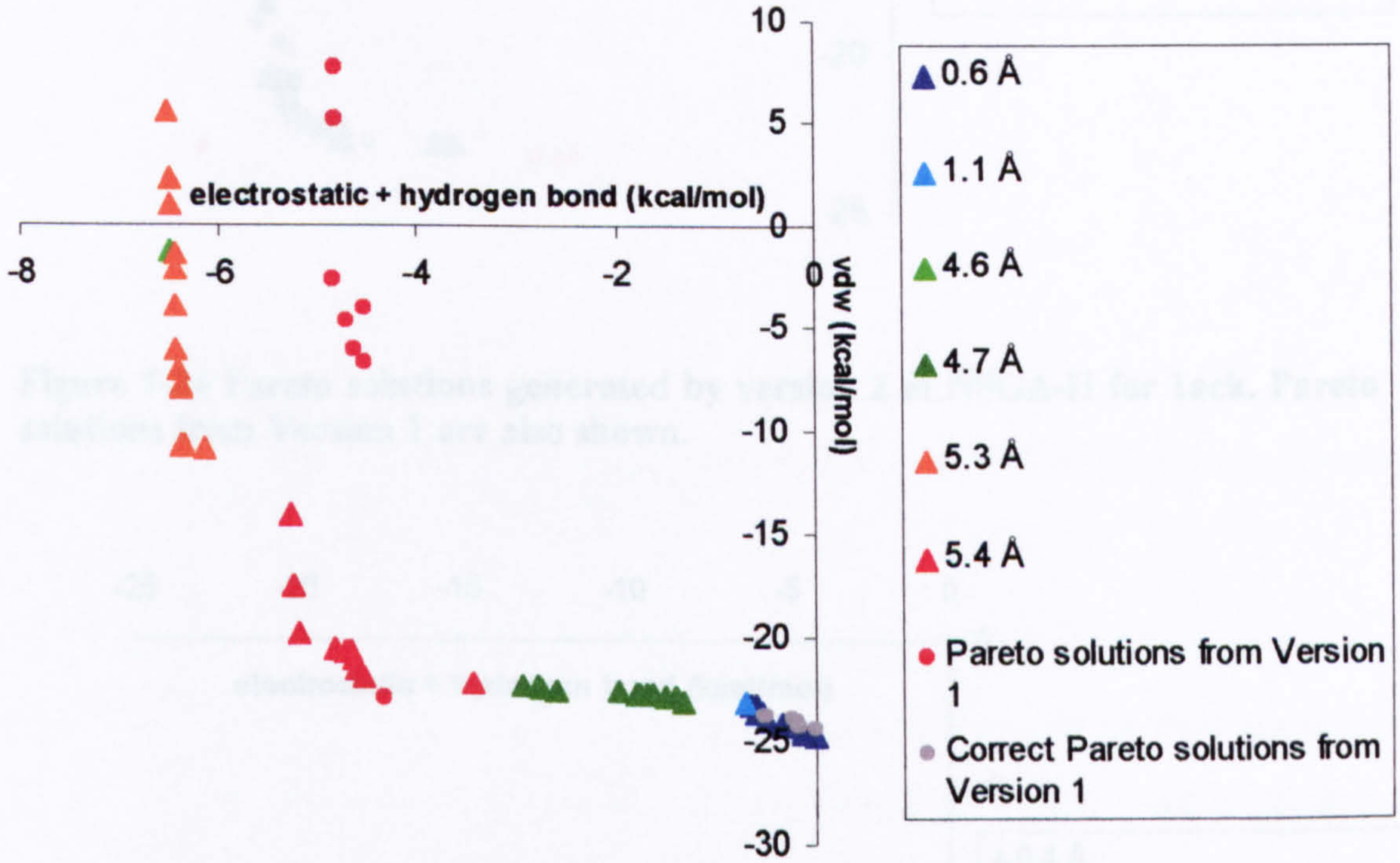


Figure 7-23 Pareto solutions generated by version 2 of NSGA-II for 1acj. Pareto solutions from Version 1 are also shown.

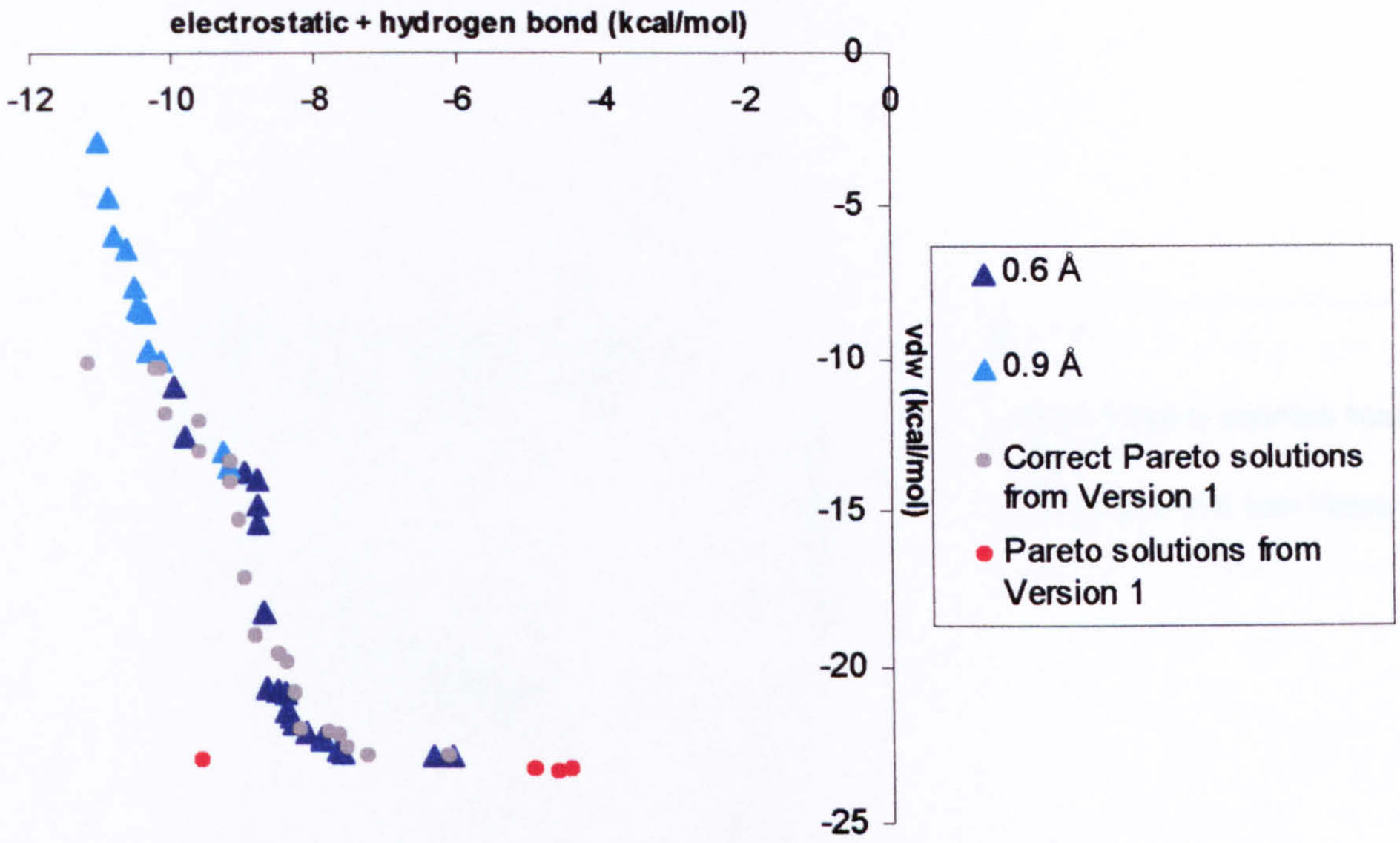


Figure 7-24 Pareto solutions generated by version 2 of NSGA-II for 1ack. Pareto solutions from Version 1 are also shown.

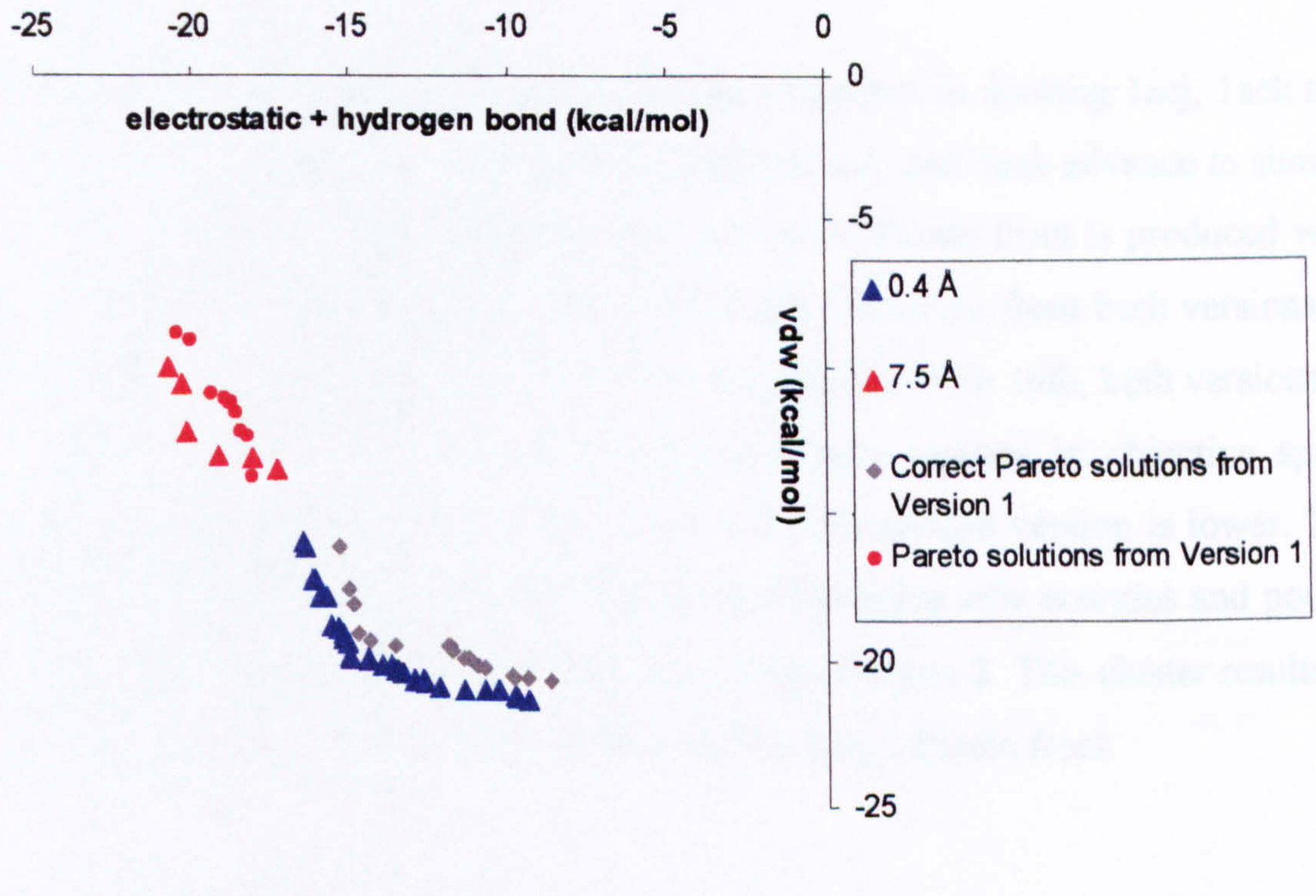


Figure 7-25 Pareto solutions generated by version 2 of NSGA-II for 2ak3. Pareto solutions from Version 1 are also shown.

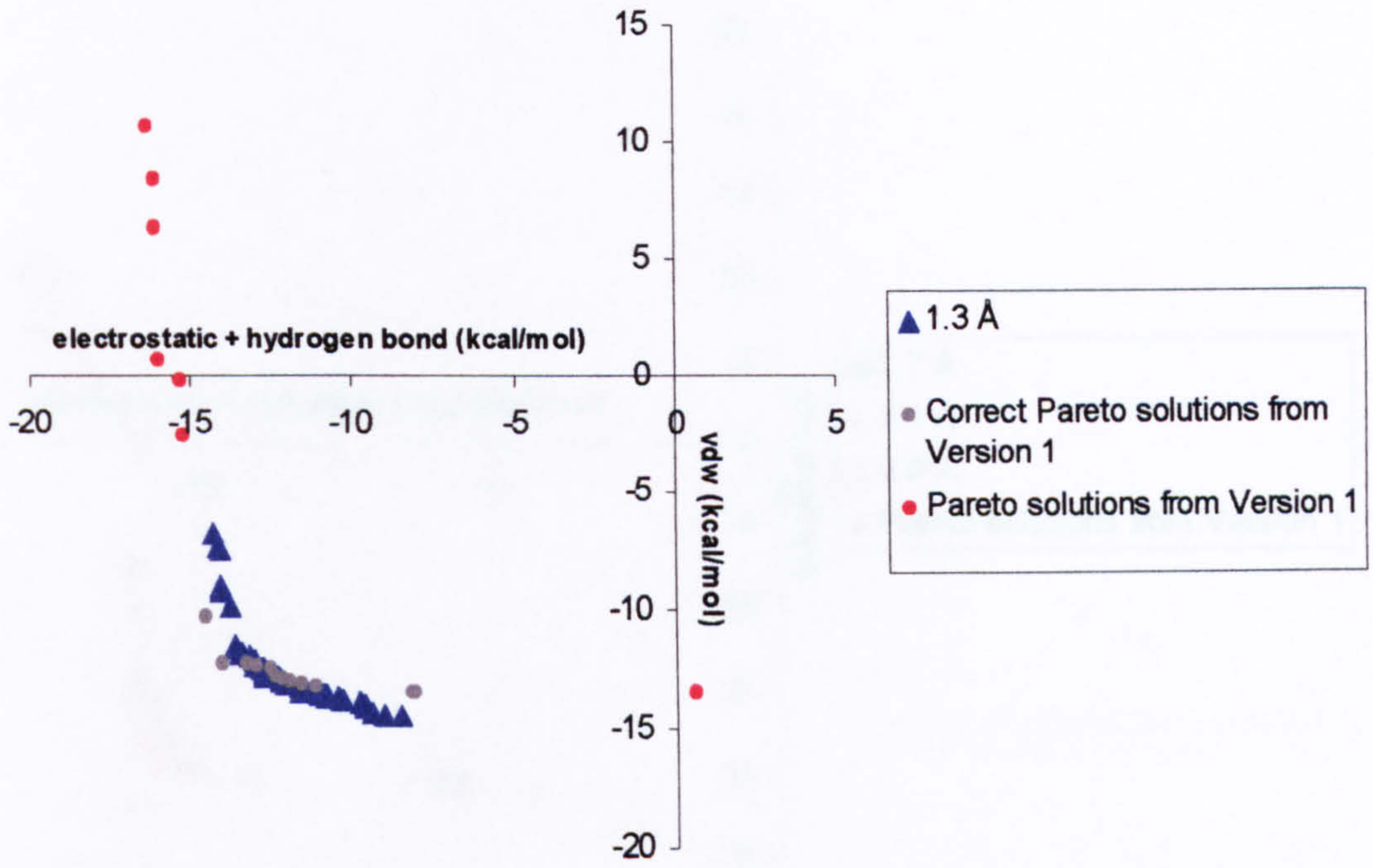


Figure 7-26 Pareto solutions generated by version 2 of NSGA-II for 1tdb. Pareto solutions from Version 1 are also shown.

Version 2 was as successful as Version 1 of the algorithm in docking 1acj, 1ack and 2ak3 (Figures 7.23, 7.24, 7.25). The Pareto fronts of 1acj and 1ack advance to similar positions in objective space- a slightly more advanced Pareto front is produced with Version 2 for 2ak3. The rmsds of the correct Pareto solutions from both versions of the algorithm are similar for all three of these complexes. With 1tdb, both versions of the program produced a cluster that occupies similar regions in objective space (Figure 7.26)- though the rmsd of the cluster from the second version is lower. The v.1 NSGA-II produced a second cluster with more positive vdw energies and poorer rmsds, and which is absent from the Pareto set from Version 2. This cluster results in the Pareto front being further extended than the Version 2 Pareto front.

7.5.2.2 4fab, 1mup and 2mth

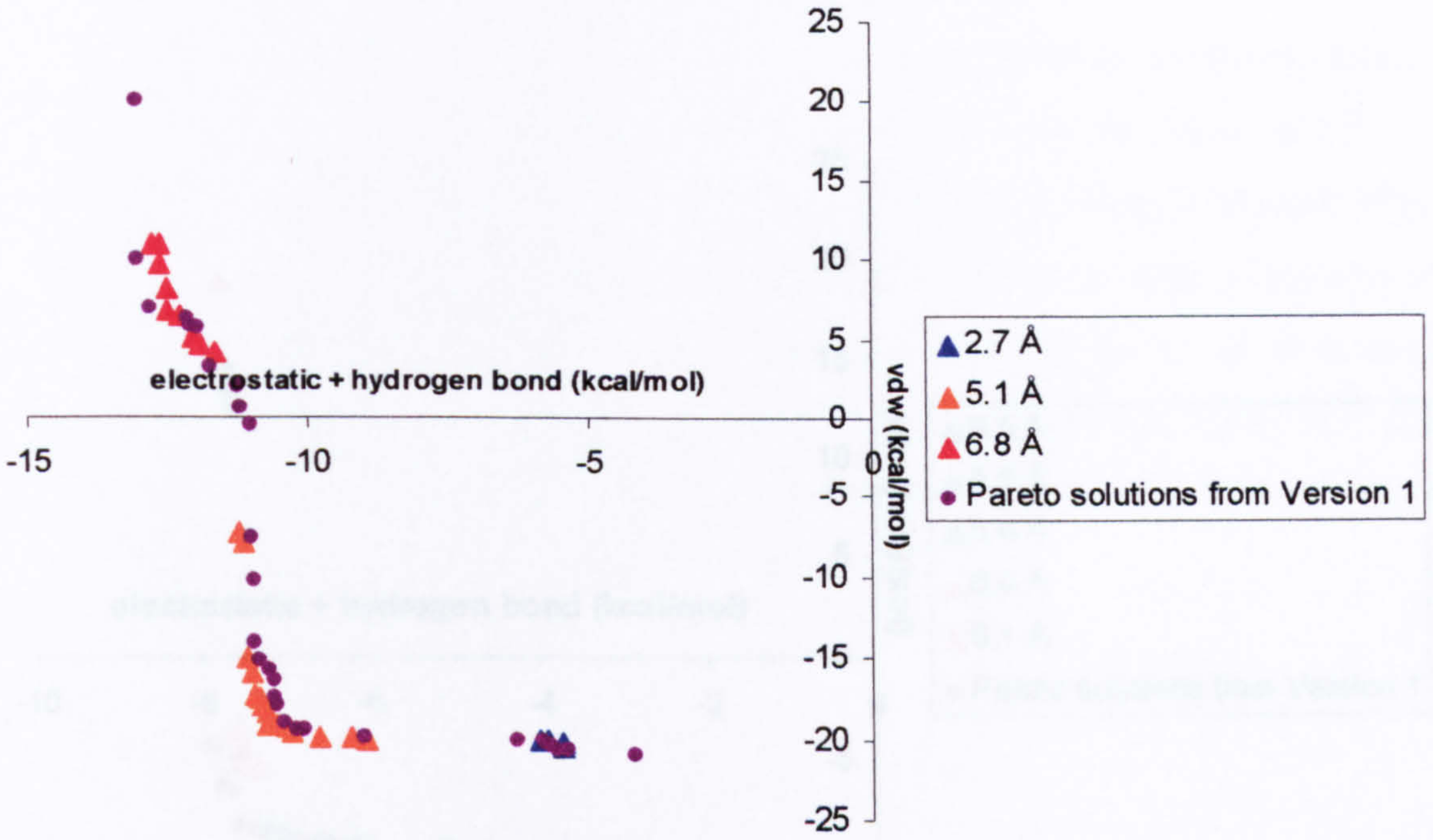


Figure 7-27 Pareto solutions generated by version 2 of NSGA-II for 4fab. Pareto solutions from Version 1 are also shown.

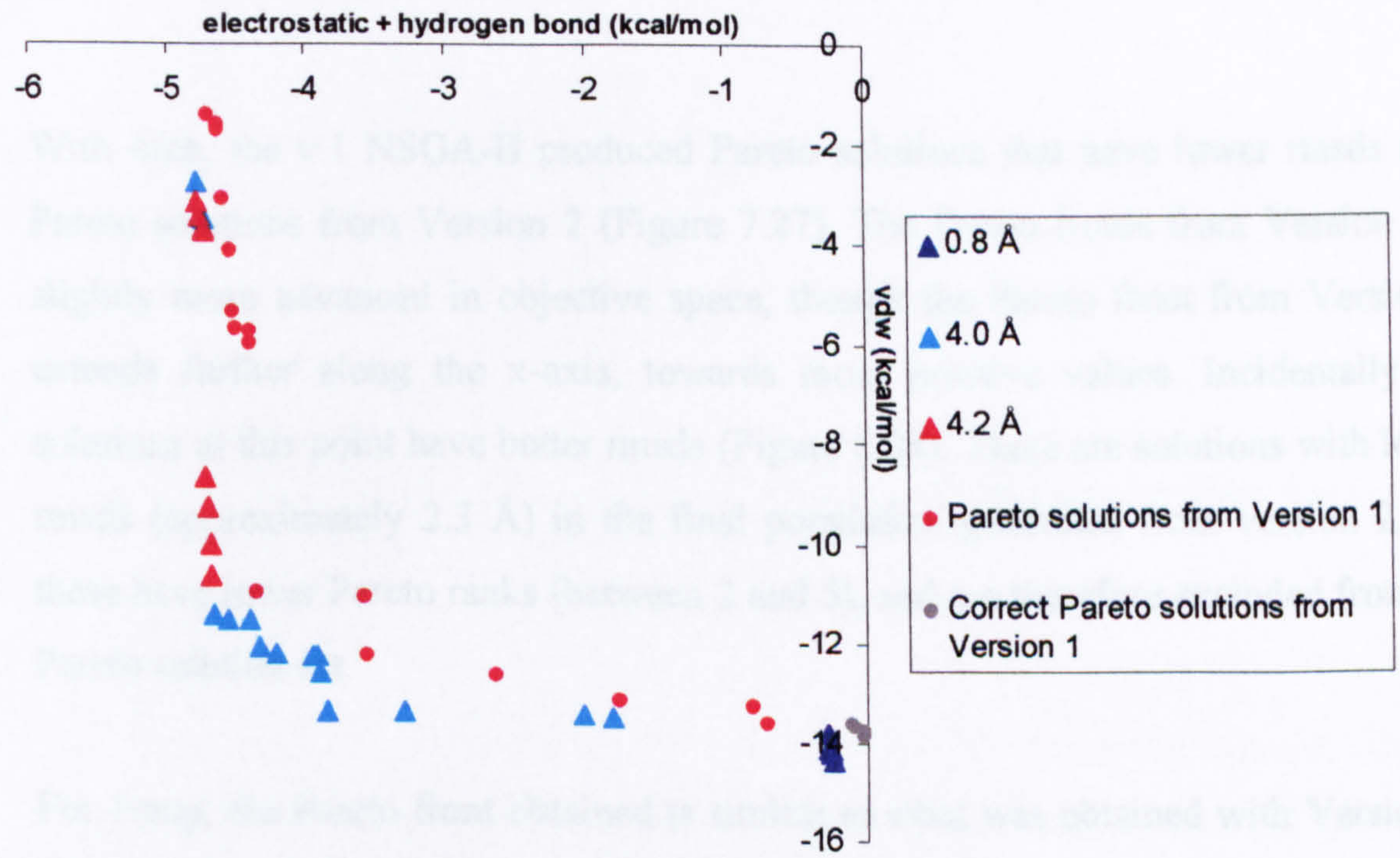


Figure 7-28 Pareto solutions generated by version 2 of NSGA-II for 1mup. Pareto solutions from Version 1 are also shown.

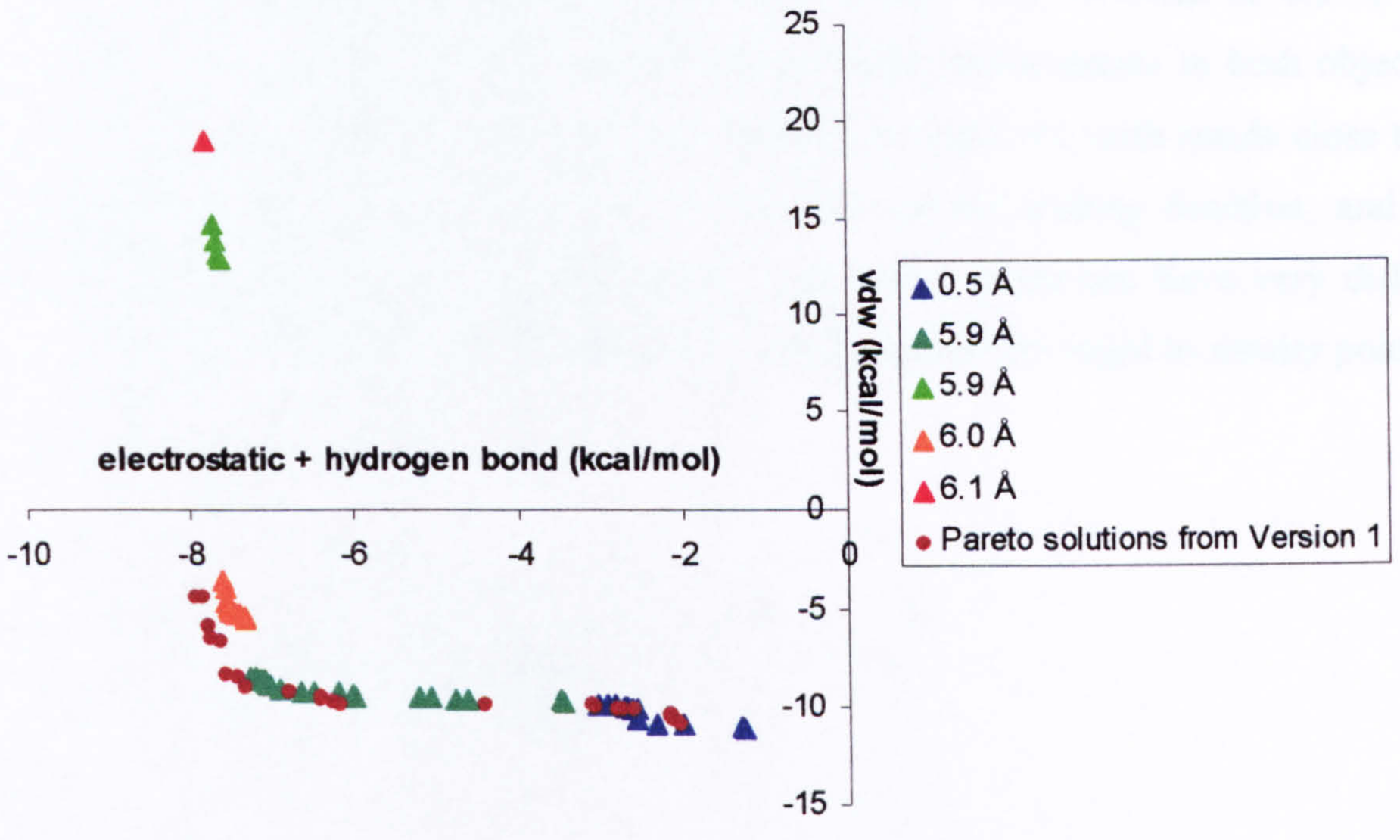


Figure 7-29 Pareto solutions generated by version 2 of NSGA-II for 2mth. Pareto solutions from Version 1 are also shown.

With 4fab, the v.1 NSGA-II produced Pareto solutions that have lower rmsds than Pareto solutions from Version 2 (Figure 7.27). The Pareto fronts from Version 1 is slightly more advanced in objective space, though the Pareto front from Version 1 extends further along the x-axis, towards more positive values. Incidentally the solutions at this point have better rmsds (Figure 6.28). There are solutions with lower rmsds (approximately 2.3 Å) in the final population generated from version 2, but these have lower Pareto ranks (between 2 and 5), and are therefore excluded from the Pareto solution set.

For 1mup, the Pareto front obtained is similar to what was obtained with Version 1, and it is slightly more advanced (Figure 7.28). Both versions produced one cluster with a good rmsd (0.4 Å with Version 1 and 0.8 Å with Version 2), and which were in similar positions of the Pareto space.

2mth was not docked by Version 1 of the algorithm (Figure 7.29). The lowest rmsd obtained by a cluster was 4.3 Å. Version 2 of the algorithm, however, was able to dock the ligand successfully and obtained a cluster with an rmsd of 0.5 Å. This cluster only has slightly better energies. This small improvement in both objectives has resulted in a change of orientation- resulting in solutions with rmsds close to the crystal structure. This illustrates the sensitivity of the scoring function, and how solutions which are close to each other in objective space can have very different orientations. The Pareto fronts from both versions have converged to similar points.

7.5.2.3 6rsa

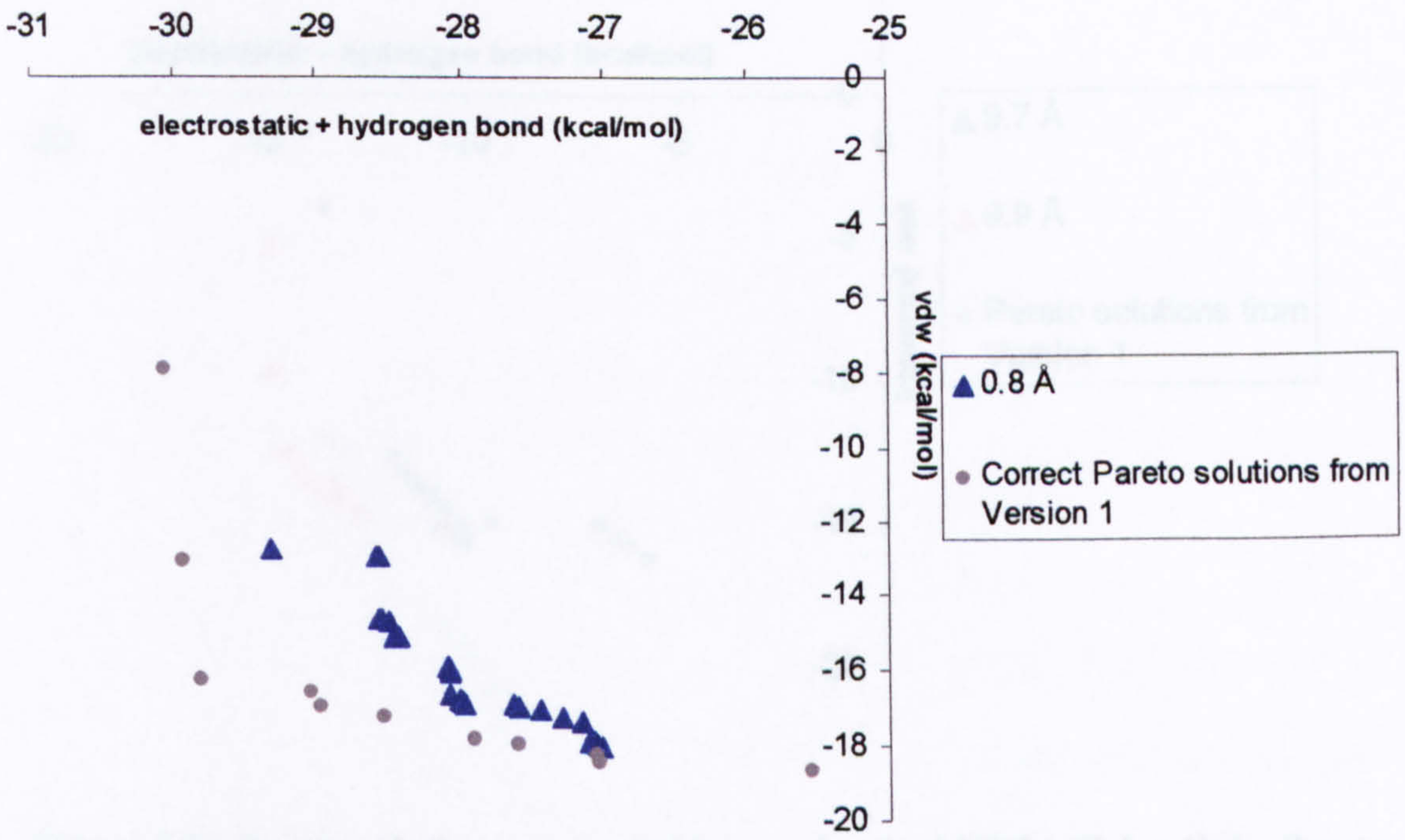


Figure 7-30 Pareto solutions generated by version 2 of NSGA-II for 6rsa. Pareto solutions from Version 1 are also shown.

Both versions of the algorithm produced Pareto sets containing correct solutions when tested with 6rsa (Figure 7.30). Both Pareto fronts obtained have varied distributions in objective space. The Pareto solutions from Version 1 (all of which have rmsds of less than 2.0 Å), are spread over a wide range of vdw energies (~-4 to -18 kcal/mol) and the electrostatics and hydrogen bond energies cover a smaller range (~-25 to -30 kcal/mol). The Version 2 Pareto solutions have vdw energies which are

overall as advanced as the Version 1 Pareto solutions, but cover a smaller range- (~ -13 to ~-18 kcal/mol). The electrostatic and hydrogen bond energies of the Pareto solutions from version 2 also cover a smaller range. This objective is slightly more advanced with the Version 1 Pareto solutions. The correct clusters from both versions have similar rmsds.

7.5.2.4 1hdc

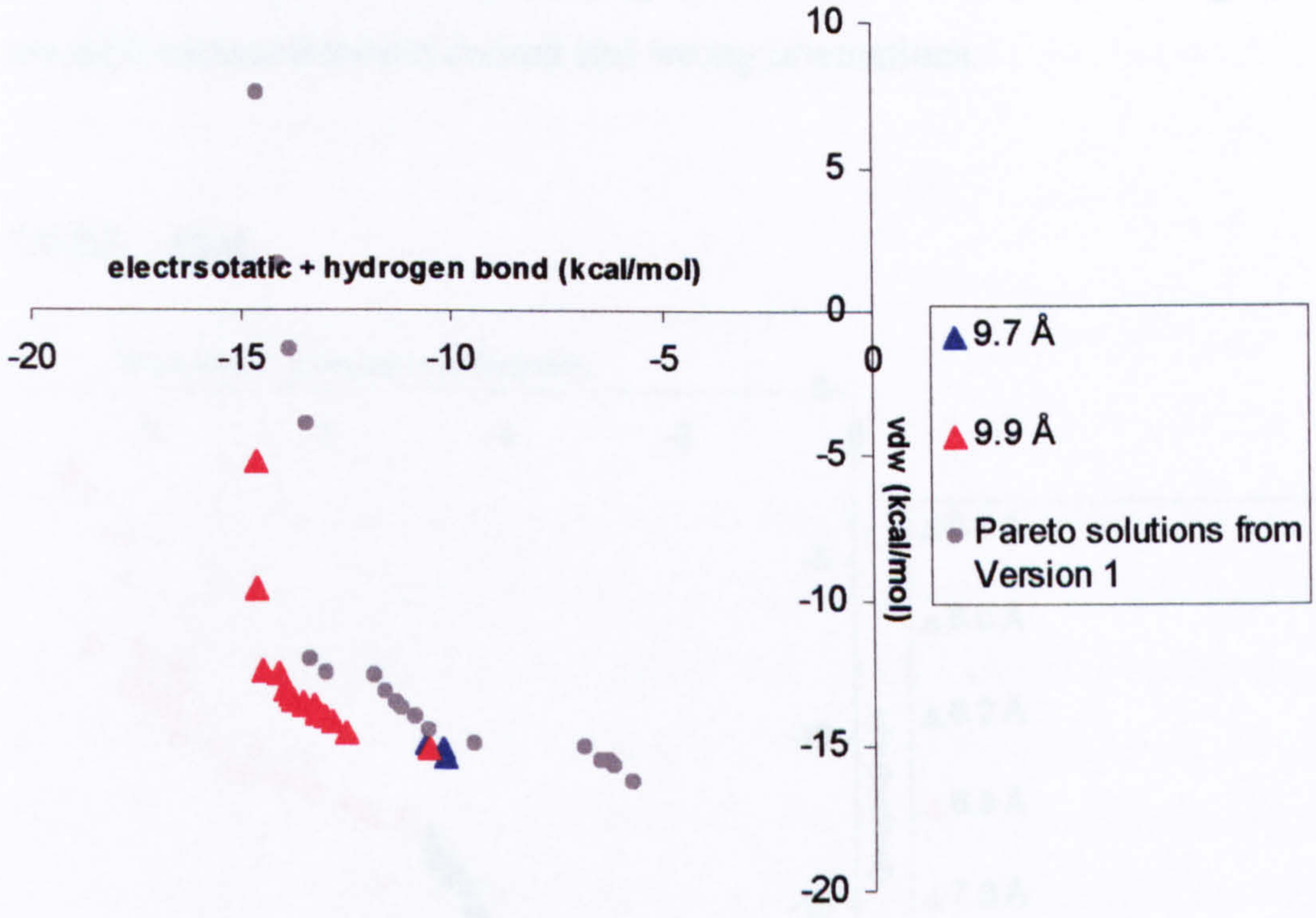


Figure 7-31 Pareto solutions generated by version 2 of NSGA-II for 1hdc. Pareto solutions from Version 1 are also shown.

1hdc was a problematic complex, which version 1 of the algorithm did not dock successfully. The v.2 NSGA-II was also unsuccessful at docking this complex- the lowest rmsd obtained for a cluster was 9.7 Å (Figure 7.31). The Pareto fronts from the two versions have converged to similar points in objective space. The Pareto front from version 1 extends into positive vdw energy space- the highest vdw energy value reached by the version 2 Pareto front is approximately -5 kcal/mol. Also the Version 1 Pareto front solutions cover a wider range of electrostatics and hydrogen bond energies (~-6 to ~-14 kcal/mol). The failure of this algorithm to dock this molecule

using either of the versions- and which Q-fit docked only at a lower rank of 18- indicates that this is a difficult complex to dock correctly. The solution with the lowest rmsd from the Q-fit output (with an rmsd of 1.5 Å from the crystal structure) has a total energy of -26.05 kcal/mol, which is higher than solutions with worse rmsds- hence its low rank. The magnitude of the scoring function therefore does not agree with the quality of the molecule's orientation- an orientation with a high rmsd has a more favourable energy than an orientation with a good rmsd. This makes 1hdc a complex that is difficult for the algorithm to dock- since the scoring function does not differentiate between correct and wrong orientations.

7.5.2.5 1baf

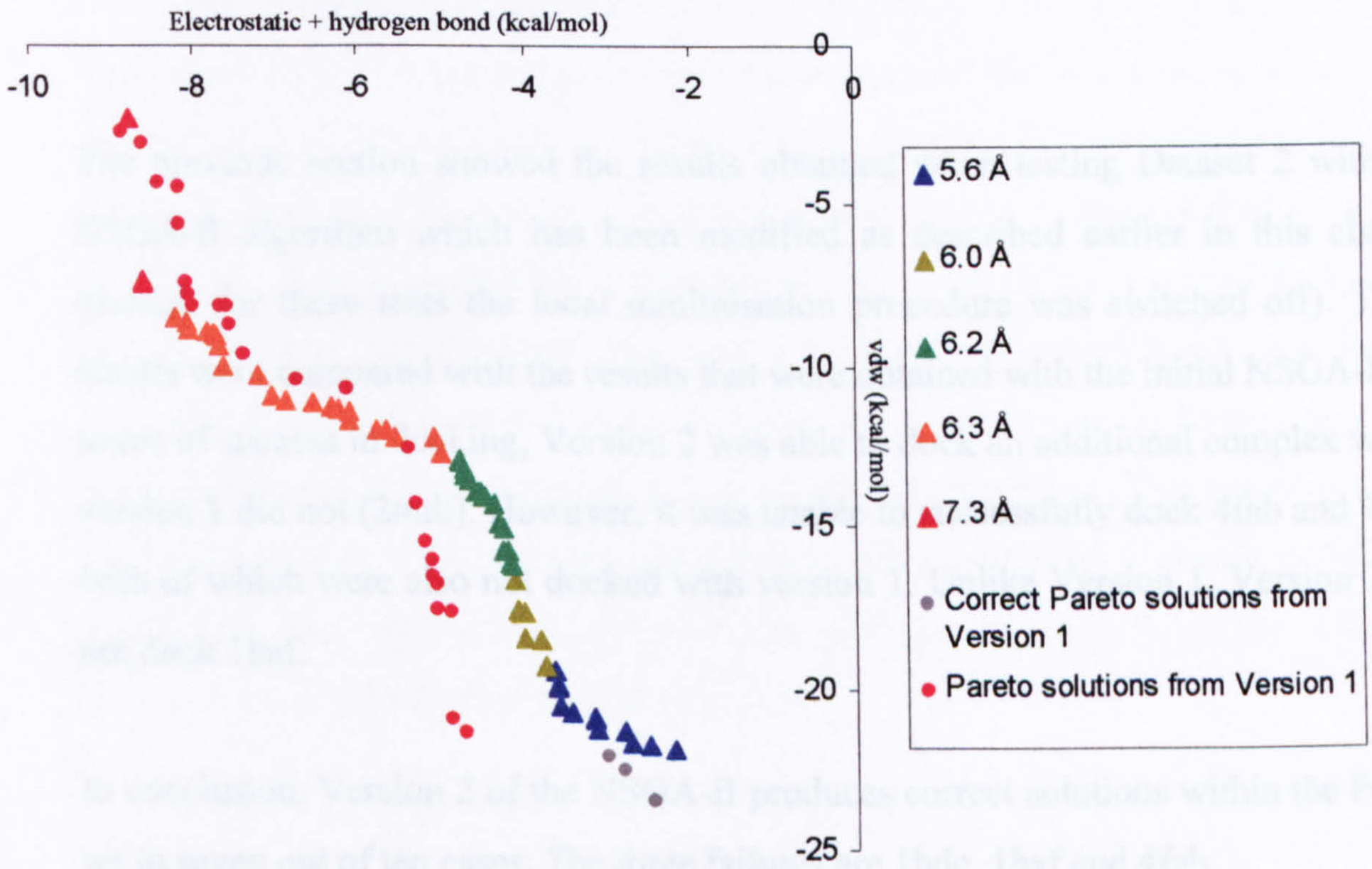


Figure 7-32 Pareto solutions generated by version 2 of NSGA-II for 1baf. Pareto solutions from Version 1 are also shown.

A test case which version 1 docked and version 2 did not is 1baf. The Pareto fronts obtained from the two versions when docking 1baf are in roughly similar positions of objective space, though a section of the Version 1 Pareto front is more advanced (Figure 7.32). The correct cluster from version 1 (0.6 Å), on the edge of the Pareto

front, has stronger vdw than electrostatics and hydrogen bond energies. The cluster labelled 5.6 Å from the version 2 Pareto front has very similar balances of objectives, yet its orientation is very different. This suggests the difficulty of this case: two energy minima, one local and one global, have very similar energies. Presumably it is more difficult to reach the global minimum, which may be down a narrow well. Version 2 of the algorithm seems to get stuck at the local minimum, whereas version 1 is able to reach the slightly more energetically favourable global minimum. The difficulty of this case suggests the possibility that version 1 may have "found" the correct solution by chance, and that different starting populations would not always produce good solutions.

7.5.2.6 Summary of results from Dataset 2

The previous section showed the results obtained when testing Dataset 2 with the NSGA-II algorithm which has been modified as described earlier in this chapter (though for these tests the local minimisation procedure was switched off). These results were compared with the results that were obtained with the initial NSGA-II. In terms of success in docking, Version 2 was able to dock an additional complex which version 1 did not (2mth). However, it was unable to successfully dock 4fab and 1hdc, both of which were also not docked with version 1. Unlike Version 1, Version 2 did not dock 1baf.

In conclusion, Version 2 of the NSGA-II produces correct solutions within the Pareto set in seven out of ten cases. The three failures are 1hdc, 1baf and 4fab.

7.6 Conclusions

The aim of this chapter was improve the performance of the algorithm on datasets 1 and 2, and particularly to dock 4dfr, the "model" test case. The three modifications implemented- controlled elitism, reducing E_{max} and local minimisation with Lamarckian element, succeeded in docking all ten cases from Dataset 1, and seven of

the ten cases from Dataset 2- giving a success rate of 85%. This is an improvement from the results obtained with version 1, where a success rate of 70% was achieved (taking into account that docking 1baf was not successful). Along with the fact that these enhancements have improved the results obtained from datasets, they have also provided the opportunity to explore the capabilities of these enhancements in the algorithm. The implementation of the local minimisation/Lamarckianism is a novel element which, as far as we are aware, has not been previously used in multiobjective optimisation. This has provided some benefit to the performance of the algorithm- as the Dataset 1 results show. Because this feature was detrimental on results from Dataset 2, but was beneficial to Dataset 1 results, it should be experimented with before deciding on whether to implement it on a particular case/dataset.

All three modifications allow the algorithm to be parameterised and tweaked to obtain the best results. The controlled elitism feature, which did not achieve a profound improvement of results, nevertheless provides the option of controlling the level of elitism in the algorithm. The E_{max} reduction feature can be manipulated by switching it off or on, or by changing the generation number at which it is reverted back to the default value of 5.0 kcal/mol. The local minimisation/Lamarckianism feature can have the termination criterion, λ_1 and λ_2 manipulated, as well as the proportion of the population to which this is applied at the end of each generation. This provides a user with the opportunity to tailor the algorithm to suit a particular problem, in order to achieve the best results. These features also give the algorithm flexibility- different parameters could be applied to different cases.

8 Testing of the NSGA-II on different datasets

In this chapter, and as specified in the aims of the thesis in section 3.5, the effect of running the NSGA-II on different datasets is explored. These are the FlexX dataset (Kramer *et al.*, 1999), and a dataset consisting of glycogen synthase kinase-3 beta (GSK-3 beta) co-crystallised with different ligands. Enhancements in the algorithmic structure of the NSGA-II were explored in the previous chapter. This chapter attempts to understand the biological capabilities of the algorithm.

As the previous chapters have shown, the NSGA-II is capable of satisfactorily docking various ligands into their respective co-crystallised protein binding sites. More importantly by observing the position of the correct solutions in objective space we are able to learn which interactions, if any, are having the dominating effect in obtaining the correct solutions. Many of the results showed that either one of the two objectives, at any one time, could have a dominating effect. For example 1mup had a stronger vdw interaction influence, whereas with 3ptb/4dfr the electrostatics and hydrogen bond energies are more influential. These variations lead to the question of what could be observed if different ligands, which bind to the same protein, were docked using the NSGA-II. Would there be a trend in the balance of the objectives? Are there any trends in the spread of the Pareto front or in the positions of the correct solutions on the Pareto front? To attempt to answer these questions, the algorithm was tested on the GSk-3 beta dataset.

8.1 Glycogen synthase kinase-3 beta

GSK-3 beta is a serine/threonine protein kinase which phosphorylates glycogen synthase, as well as being involved in a broad range of other biological processes. GSK-3 beta has been implicated in a number of conditions, including Alzheimer's disease, type 2 diabetes, cancer and chronic inflammatory conditions (Garcea *et al.*, 2007; Takashima, 2006). GSK-3 beta is known to phosphorylate many different kinds of structural, signalling and metabolic proteins. It is linked with Alzheimer's disease

because of its known interactions with the plaque-producing amyloid system, in assisting in the formation of neurofibrillary tangles and for its interactions with other proteins associated with the condition (Mudher and Lovestone 2002). GSK-3 beta plays a part in the Wnt and Hedgehog pathways, both of which are implicated in several forms of cancer (Doble and Woodgett, 2003).

These factors make GSK-3 beta a very attractive therapeutic drug target and it has therefore been studied extensively in recent years. Many structure based design experiments that attempt to find effective inhibitors for its function have also been published (Polgar *et al*., 2005, Lescot *et al*, 2005, Naerum *et al,* 2002).

The ten complexes of GSK-3 beta that were used in this study involved molecules bound to the ATP binding site of the kinase. The parameters used to dock these molecules were the same as those described in the previous chapter. For comparison, the molecules were also docked using the docking tool Q-fit (Jackson, 2002).

PDB Code	Structure Type/Inhibitor
1Q4L	I5
1Q3W	Alsterpaullone
1PYX	AMP-PNP
1Q3D	Staurosporine
1Q41	Indirubin
1UV5	Bromo-indirubin
1GNG	apo phosphorylated structure
1J1B	AMP-PNP
1J1C	AMP-PNP
1O9U	9-methyl-9H-purin-6-amine

Table 8.1 PDB codes and ligands of GSK-3 beta dataset

8.1.1 Results

Figures 8.1 and 8.2 show the Pareto fronts obtained when these molecules were docked into their protein binding sites. The top ranked solutions obtained when docking the ligand into the binding site using the docking tool Q-fit are also shown. The rmsd of these solutions from the crystal structure are indicated in the legend. If the rmsd of the top-ranked Q-fit solution is greater than 2.0 Å, then the solution with the lowest rmsd from the entire Q-fit output list is selected and also plotted in objective space. This solution's rank (within Q-fit's list of solutions), is quoted in the legend.

The NSGA-II was able to dock six out of ten of the GSK-3-beta complexes with rmsds that are lower than 2.0 Å. The Pareto fronts obtained with some of the successful cases (e.g. 1q41) show that the correct solutions in the Pareto solution set are more influenced by vdw energies than the electrostatic and hydrogen bond energies. Correct solutions from other Pareto fronts are spread more uniformly across both objectives (eg 1q3w)- but the vdw range covered appears to be larger than the electrostatics and hydrogen bond energies.

The overall results from Q-fit agreed with the results obtained by the NSGA-II, i.e. complexes for which NSGA-II obtained correct solutions in the Pareto set also obtained correct solutions in the top ranks of Q-fit's list of output, and vice versa. This was with the exception of 1q41, 1q3d and 1pyx. The NSGA-II docked these correctly (solutions with approximate rmsds of 1.0 Å), but Q-fit did not find a solution below 2.0 Å. The results of Q-fit also correlated with the negative results obtained by the NSGA-II: the four cases for which the NSGA-II did not obtain correct solutions among its Pareto sets were also not docked correctly by Q-fit.

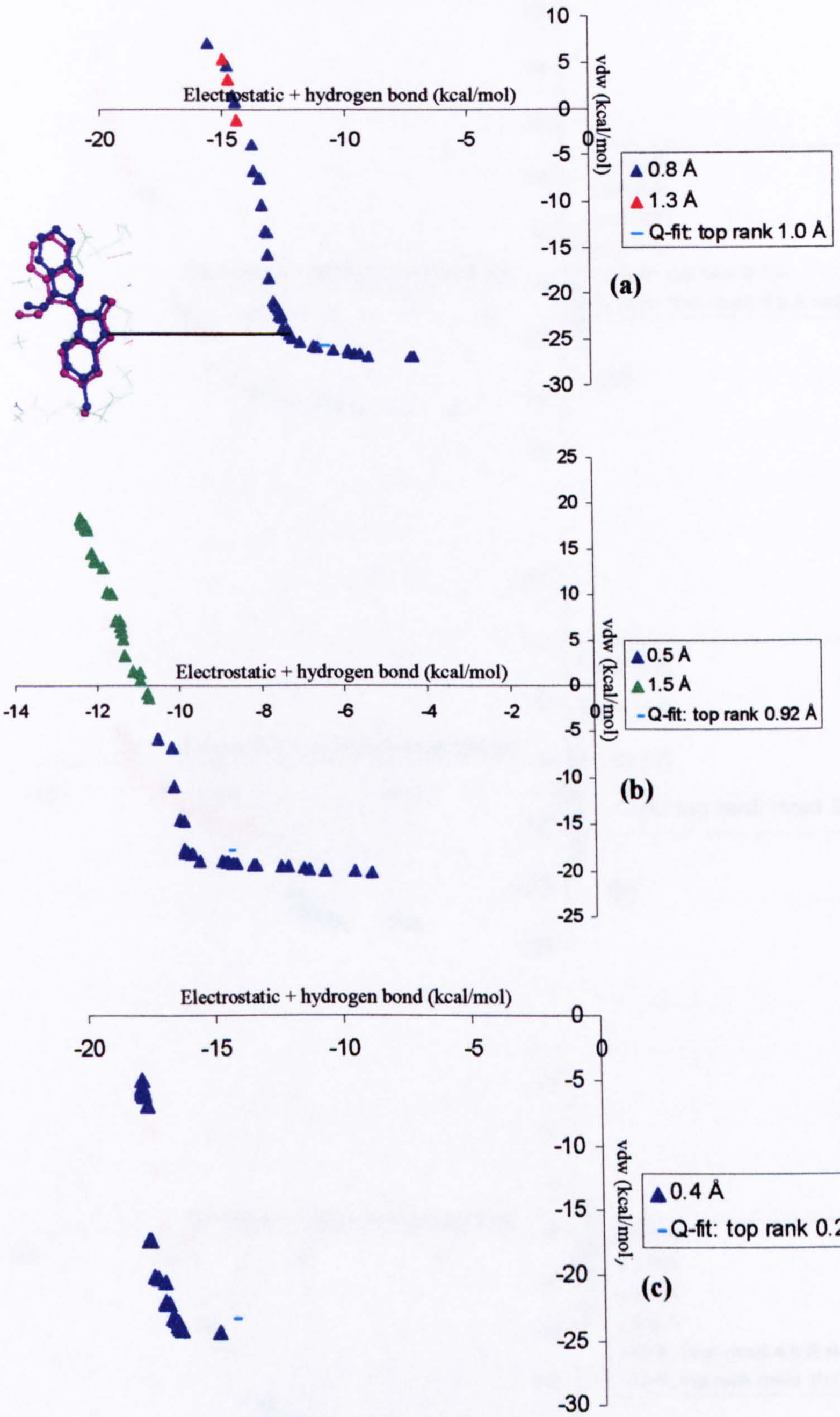
8.1.1.1 Successful cases: 1uv5, 1q3w, 1q41, 1q4L, 1q3d and 1pyx

The NSGA-II was able to successfully dock six of the ten GSK protein complexes, i.e. the Pareto solution sets generated all contained some solutions that had rmsds of

less than 2.0 Å from the crystal structure. With the exceptions of 1q3d, 1q41 and 1pyx, Q-fit obtained good rmsds for all its top-ranked solutions. All the Pareto solutions obtained with 1uv5 have good rmsds (less than 2.0 Å). The top-ranked Q-fit solution is also among the correct Pareto solutions; this implies that, presuming Q-fit has reached the global minimum, the NSGA-II has reached the true Pareto front. This scenario was also observed with 1q3w and 1q4L (Figure 8.1(b) (c)).

Q-fit did not dock the 1q41 complex successfully; the top-ranked solution has an rmsd of 6.2 Å, and the lowest rmsd for any solution was 3.2 Å (ranked at 35). Looking at the position of the top-ranked Q-fit solution relative to the Pareto front, it can be seen that this is in the same region as Pareto solutions with high rmsds (near Pareto solutions represented by green triangles) (Figure 8.1(d)). The correct Pareto solutions (blue triangles) are further down the Pareto front, towards the y-axis. This implies that, despite the Q-fit solution being amongst the Pareto solutions- it does not have a correct balance of the energy types- which results in a high rmsd from the crystal structure. Comparing the balance of the correct Pareto solutions and the Q-fit solution, it can be seen that the Q-fit solution has a slightly lower electrostatics and hydrogen bond energies. This case therefore highlights the importance of achieving a correct balance of energy types in order to find solutions with good rmsds.

The position of the top-ranked Q-fit solution for the 1q3d complex in objective space suggests that Q-fit did not reach the global minimum (Figure 8.1(e)). The rmsd of this solution is 2.6 Å from the crystal structure. The Pareto solutions (which also constitute a correct cluster) have more favourable electrostatics and hydrogen bond, and vdw energies and two of the clusters have rmsds that are below 2.0 Å (0.5 Å and 1.2 Å). This case indicates that the NSGA-II can be more effective at reaching the global minimum than Q-fit.


Electrostatic + hydrogen bond (kcal/mol)
vdw (kcal/mol)
0.8 Å
1.3 Å
Q-fit: top rank 1.0 Å
(a)
Electrostatic + hydrogen bond (kcal/mol)
vdw (kcal/mol)
0.5 Å
1.5 Å
Q-fit: top rank 0.92 Å
(b)
Electrostatic + hydrogen bond (kcal/mol)
vdw (kcal/mol)
0.4 Å
Q-fit: top rank 0.26 Å
(c)

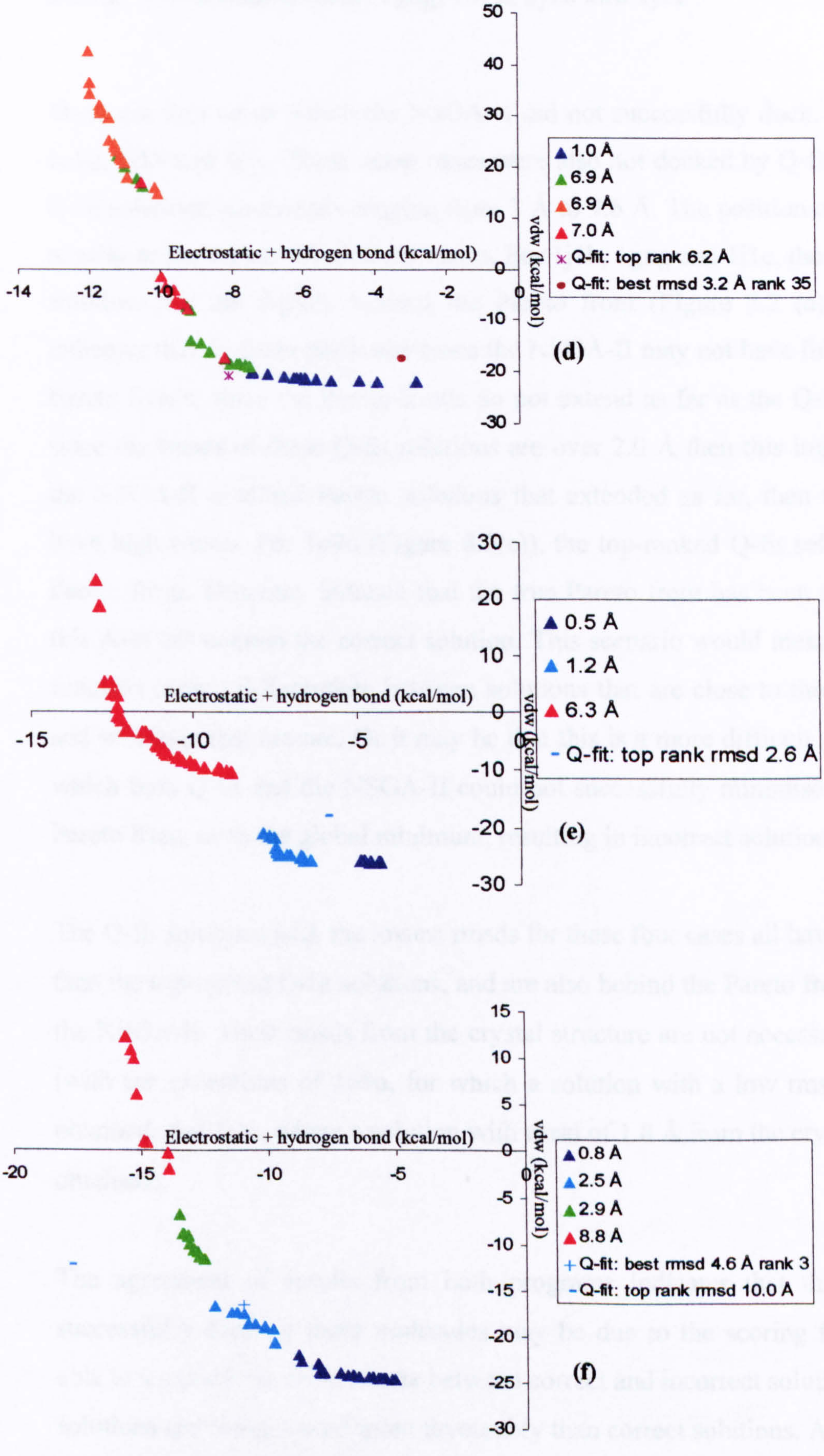


Figure 8-1 Correct Pareto solutions produced by NSGA-II when docking GSK-3 beta complexes. The plots also show Q-fit solutions. (a) 1uv5 (b) 1q3w (c) 1q4L (d) 1q41 (e) 1q3d (f) 1pyx

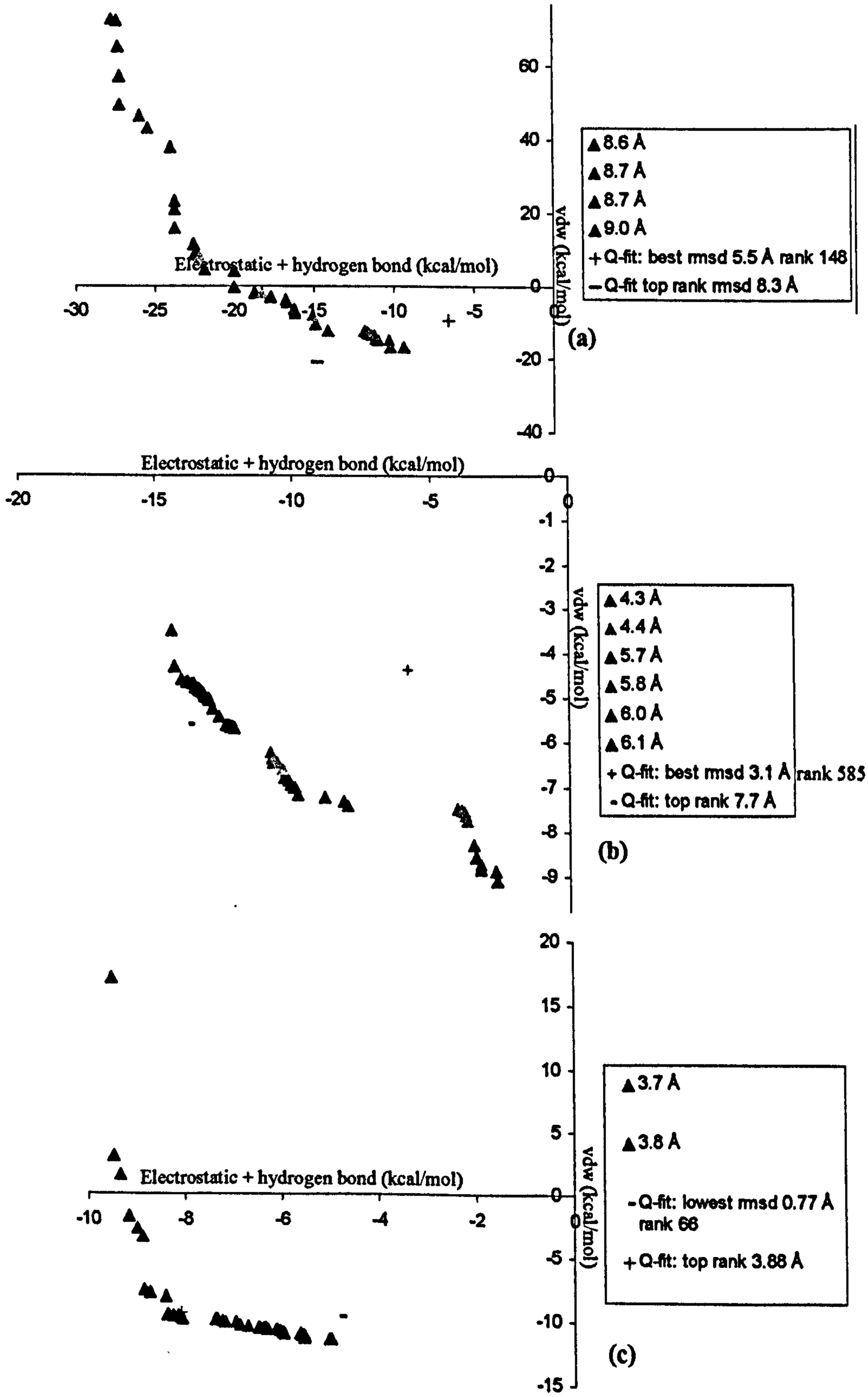
8.1.1.2 Unsuccessful cases: 1gng, 1o9u, 1j1b and 1j1c

There are four cases which the NSGA-II did not successfully dock. These are 1gng, 1o9u, 1j1b and 1j1c. These same cases were also not docked by Q-fit; the top-ranked Q-fit solutions have rmsds ranging from 3 Å to 9.6 Å. The position of these solutions relative to the Pareto solution set varies. For 1j1b, 1gng and 1j1c, the top-ranked Q-fit solutions are all slightly beyond the Pareto front (Figure 8.2 (a), (b), (d)). This indicates that, in these particular cases the NSGA-II may not have fully optimised the Pareto fronts, since the Pareto fronts do not extend as far as the Q-fit solutions. But since the rmsds of these Q-fit solutions are over 2.0 Å then this implies that even if the NSGA-II obtained Pareto solutions that extended as far, then these would still have high rmsds. For 1o9u (Figure 8.2(c)), the top-ranked Q-fit solution lies on the Pareto front. This may indicate that the true Pareto front has been reached, and that this does not contain the correct solution. This scenario would mean that the scoring function cannot differentiate between solutions that are close to the crystal structure and solutions that are not. Or it may be that this is a more difficult complex to dock, which both Q-fit and the NSGA-II could not successfully minimise towards the true Pareto front, or to the global minimum, resulting in incorrect solutions.

The Q-fit solutions with the lowest rmsds for these four cases all have higher energies than the top-ranked Q-fit solutions, and are also behind the Pareto fronts generated by the NSGA-II. Their rmsds from the crystal structure are not necessarily below 2.0 Å (with the exceptions of 1o9u, for which a solution with a low rmsd of 0.77 Å was obtained, and 1j1c, where a solution with rmsd of 1.8 Å from the crystal structure was obtained).

The agreement of results from both programs indicates that the reason for not successfully docking these molecules may be due to the scoring function: it is not able to successfully differentiate between correct and incorrect solutions and incorrect solutions are being scored more favourably than correct solutions. Another possibility is that the search procedure of both algorithms is not effectively sampling favourable areas of the search space so no correct orientations are ever produced during a search.

A reason for this may be that correct solutions for these cases lie down narrow energy wells, making it difficult for the algorithms to reach those points.



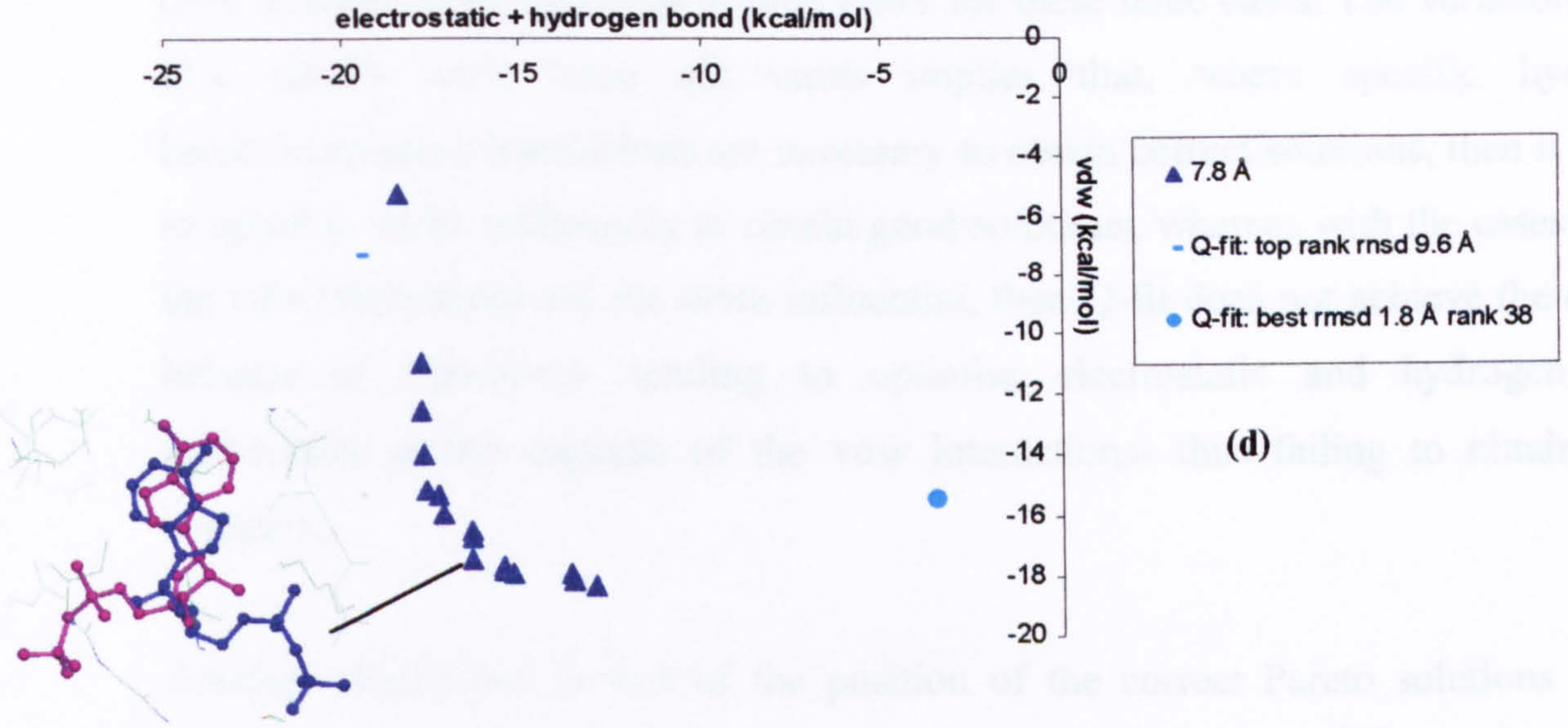


Figure 8-2 Pareto solutions produced by NSGA-II when docking GSK-3 beta complexes. The plots also show Q-fit solutions. (a) 1j1b (b) 1gng (c) 1o9u (d) 1j1c

8.1.2 Discussion of results obtained with GSK-3 beta dataset

The aim of this experiment was to observe Pareto fronts which have been generated by docking different molecules which have been co-crystallised with the same protein into their respective binding sites. Two trends were observed when looking at the distribution of the correct Pareto solutions in objective space. With 1pyx, 1qrd and 1q41, the vdw interactions are seen to be the dominating objective in obtaining the correct solutions- these interactions are more favourable relative to the electrostatics and hydrogen bond interactions. Q-fit failed to obtain good solutions at high ranks for these cases. With 1uv5, 1q3w and 1q4L, the correct Pareto solutions are distributed over a wider range of vdw interactions whereas the electrostatic and hydrogen bond energies of the same solutions have a narrower range, and in fact a few of the correct solutions appear to have very similar electrostatic and hydrogen bond energies

(demonstrated by the almost vertical portions of the Pareto fronts). This implies that making specific electrostatic and hydrogen bond interactions are important in obtaining correct solutions, but that the vdw interaction energies can be more varied. Q-fit obtained good solutions at high ranks for these three cases. The variation of Q-fit's results with these six cases implies that, where specific hydrogen bond/electrostatic interactions are necessary to obtain correct solutions, then it is able to optimise these sufficiently to obtain good solutions, whereas with the cases where the vdw interactions are the more influential, then Q-fit does not achieve the correct balance of objectives- tending to optimise electrostatic and hydrogen bond interactions at the expense of the vdw interactions- thus failing to obtain good solutions.

Another observation is that of the position of the correct Pareto solutions on the Pareto front. As the six correct Pareto fronts show, clusters containing the correct Pareto solutions are always on the right edge of the Pareto front. This may be a potentially useful application for prospective docking: if other molecules whose orientations with the protein are not known are docked using the NSGA-II into the GSK-3 beta binding site, it may be possible to infer that the Pareto solutions on the right edge of the Pareto front are more likely to be the correct orientations. This would require further verification by docking more ligands with GSK-3-beta and observing the Pareto fronts obtained, as well as performing more docking studies of proteins which have been co-crystallised with different ligands.

8.2 The Flexx Dataset

The results that have been discussed in this and previous chapters have shown that by comparing Pareto fronts and the relative positions of the top-ranked Q-fit solutions in objective space, it is possible to understand why Q-fit may fail to dock certain complexes. As the results with the GSK dataset have shown, the predominant cause for these failures is the inability of Q-fit to obtain a correct balance of objectives for solutions that have a favourable energy, and that have high ranks. These cases have therefore highlighted why it is important, for certain cases, to obtain a correct balance

of energy types, and the advantage that multiobjective optimisation possesses over single objective optimisation in explaining the different balance of energy types. To be able to explore the capabilities of multiobjective optimisation relative to single objective optimisation further, both the NSGA-II and Q-fit were run on a more extensive dataset, the Flexx dataset, which consists of 200 protein-ligand complexes (Kramer, *et al.*, 1999). By looking at cases within the Flexx dataset where the NSGA-II was successful and Q-fit was not, it may be possible, by observing Pareto and Q-fit solutions in objective space, to understand causes for Q-fit's failure in obtaining solutions with good rmsds. In the following section ten cases from the Flexx dataset, where the NSGA-II was successful and Q-fit was not, are presented and discussed.

8.2.1 Comparison of the NSGA-II with Q-fit

The FlexX ligands were docked into their corresponding protein binding sites (i.e. the binding sites with which they are co-crystallised) using Version 2 of the NSGA-II (with the local minimisation parameter switched off), and Q-fit. Comparing the results from both programs, there are 17 cases out of 200 where the NSGA-II was successful and Q-fit was not. Overall, Q-fit docks 104 of the 200 complexes successfully and the NSGA-II finds correct Pareto solutions for 84 of the 200 cases. The Pareto fronts obtained with the NSGA-II and Q-fit solutions with the best rmsd and that have the highest rank are shown in Figures 8.3 to 8.15. Examining these plots show that the results broadly follow two trends which are:

- Cases where Q-fit was unsuccessful because the highest ranked solutions obtained for a given complex did not achieve a correct balance of energies.
- Cases where Q-fit was unsuccessful because it did not obtain solutions whose energies are as minimised as the Pareto solutions.

These two situations are differentiated by the positions of the top-ranked Q-fit solution relative to the Pareto front. The first situation is observed if the top-ranked Q-fit solution is on the Pareto front, but is at a point that is away from correct Pareto clusters, i.e. its balance of energies differs from the correct Pareto solutions. The

second situation is observed if the top ranked Q-fit solution lies behind the Pareto front, therefore both its objectives (or energies) are not as optimised as the Pareto solutions. Because the Pareto solutions for these cases contain solutions which are below 2.0 Å, it may be assumed that these solutions have converged to (or very near to) the true Pareto front.

8.2.2 Q-fit solutions with incorrect balance of energies: 1xie, 2r07, 3hvt and 1igj

The Pareto plots in Figures 8.3 to 8.6 illustrate cases where the top ranked Q-fit solution did not obtain a good rmsd of less than 2.0 Å. With 1xie, the top ranked Q-fit solution, i.e. the solution with the lowest energy, has an rmsd of 4.3 Å from the crystal structure (Figure 8.3). Looking at the balance of energies for this solution, it can be seen that the electrostatics and hydrogen bond energies obtained are further optimised than those obtained by the NSGA-II clusters with good rmsds (labelled 1.2 Å and 2.0 Å). This implies that, to obtain a solution with a good rmsd, the electrostatics and hydrogen bond energies should not be as optimised as the top-ranked Q-fit solution, and that this solution's high rmsd is due to the over-optimised electrostatic and hydrogen bond interactions. The lowest rmsd obtained by any solution from within Q-fit's ranked list is 1.3 Å, and as Figure 8.3 shows, this solution does not have favourable energies that are minimised to the level of the top-ranked Q-fit solution and the Pareto solutions.

The top-ranked solution obtained by Q-fit for 2r07 has an rmsd of 9.8 Å, and the lowest rmsd obtained by any Q-fit solution is 6.3 Å. These results are in contrast to the rmsds of the clusters in the Pareto solution set where the lowest rmsds reached are 1.0 Å and 1.5 Å. Figure 8.4 shows that both the top ranked and the best rmsd solutions from Q-fit do not have optimised vdw interactions that are at the level of the 1.0 Å cluster. The lowest vdw energy value obtained by a solution in this cluster is -27.8 kcal/mol, while the vdw value of the top-ranked Q-fit solution is -18.0 kcal/mol.

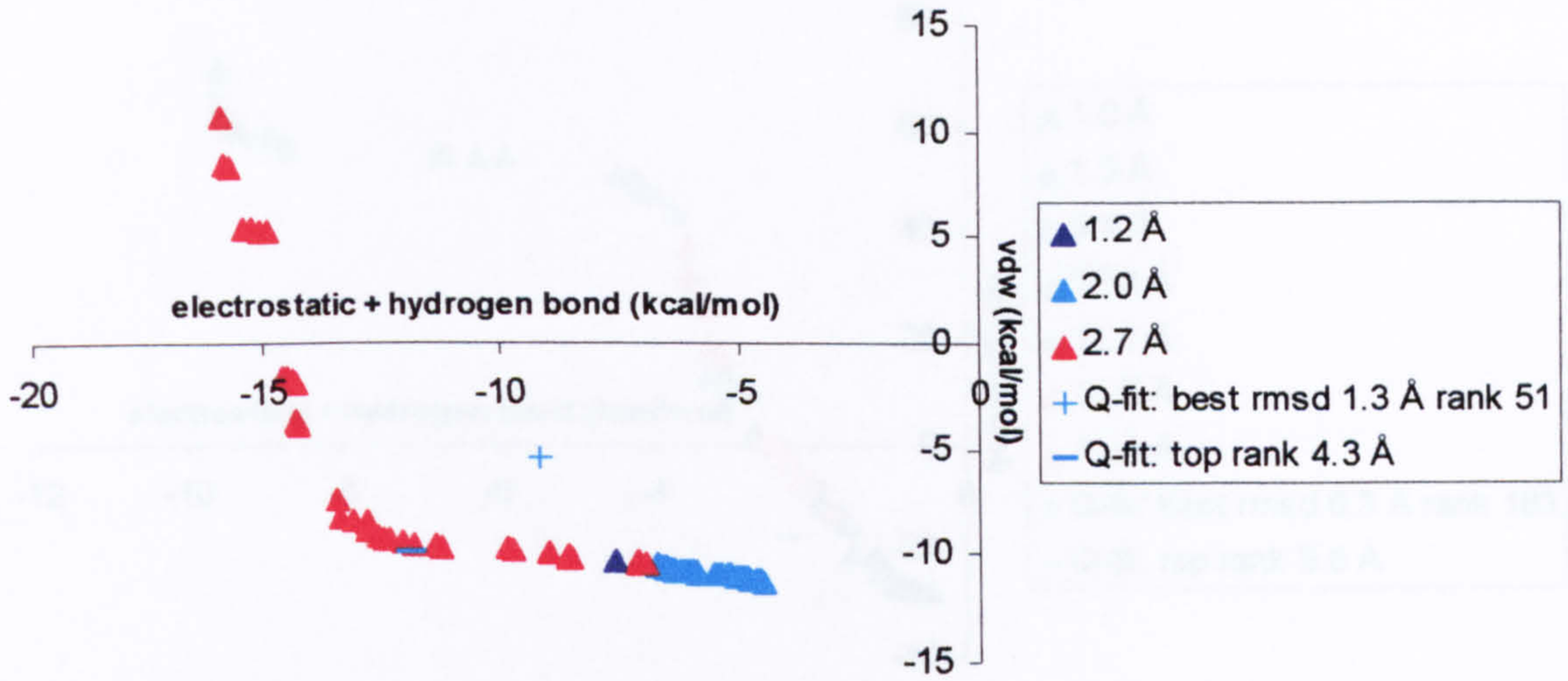


Figure 8-3 Pareto solutions produced by NSGA-II for 1xie in objective space. The positions of the top-ranked Q-fit solution and the solution with the lowest rmsd obtained by Q-fit are also shown.

The electrostatic and hydrogen bond energies on the other hand are further minimised by the top-ranked Q-fit solution, to -2.31 kcal/mol, while solutions in the 1.0 Å Pareto cluster have electrostatics and hydrogen bond energies in the range of -0.4 to -1.6 kcal/mol. By looking at the rmsd values of these solutions from the crystal structure, it is evident that a correct balance of the objectives needs to be achieved to obtain good solutions, and that the reason for Q-fit not obtaining a good rmsd solution in its top ranks is because it has optimised the electrostatic and hydrogen bond interactions at the expense of the vdw interactions. The correct NSGA-II solutions show that, in this case the vdw interaction energies are the predominant objective, and must be optimised to obtain correct solutions. The position of the top-ranked Q-fit solution shows, relative to the Pareto front, that this solution is an outlier. It therefore has a balance of energies (electrostatic and hydrogen bond energies of -2.31 kcal/mol and vdw energies of -17.98 kcal/mol) which none of the Pareto solutions are close to. This implies that at that particular point the Pareto front has not advanced to an optimum level.

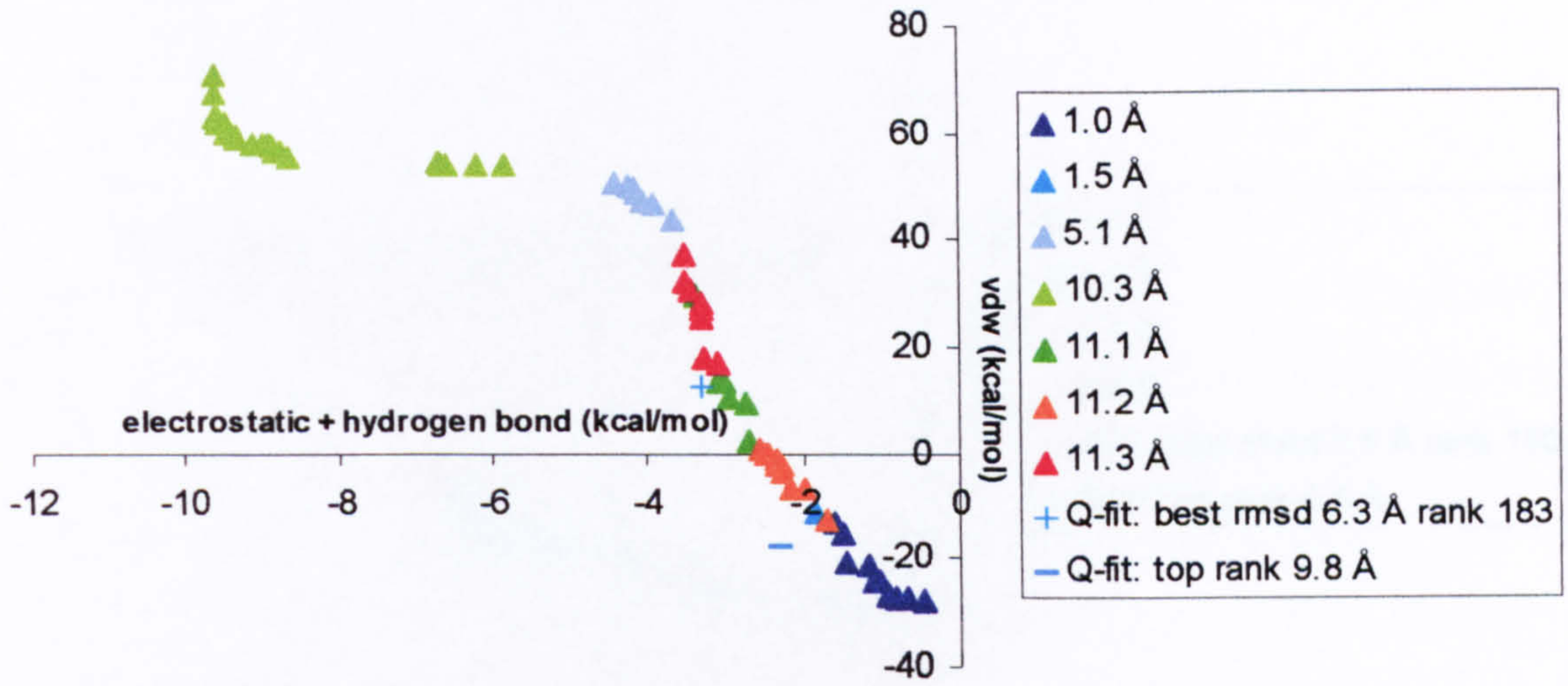


Figure 8-4 Pareto solutions produced by NSGA-II for 2r07 in objective space. The positions of the top-ranked Q-fit solution and the solution with the lowest rmsd obtained by Q-fit are also shown.

With 3hvt Q-fit did not produce any solutions with good rmsds from within its ranked list. As Figure 8.5 shows, the top-ranked Q-fit solution, with an rmsd of 4.5 Å has slightly lower electrostatic and hydrogen bond energies than the best Pareto solutions (with rmsds of 0.4 Å from the crystal), and higher vdw energies. The same scenario is observed with 1igj: the electrostatic and hydrogen bond energies of the top-ranked Q-fit solution are more negative than the Pareto solutions with the lowest rmsds (0.8 Å), whereas its vdw interactions are higher than the 0.8 Å Pareto cluster. The position of the Q-fit solution with the lowest rmsd is very close, in objective space, to the good Pareto solutions (Figure 8.6). This is reflected in its reasonably low rmsd value of 2.1 Å. It does, however, have a fairly low rank of 33.

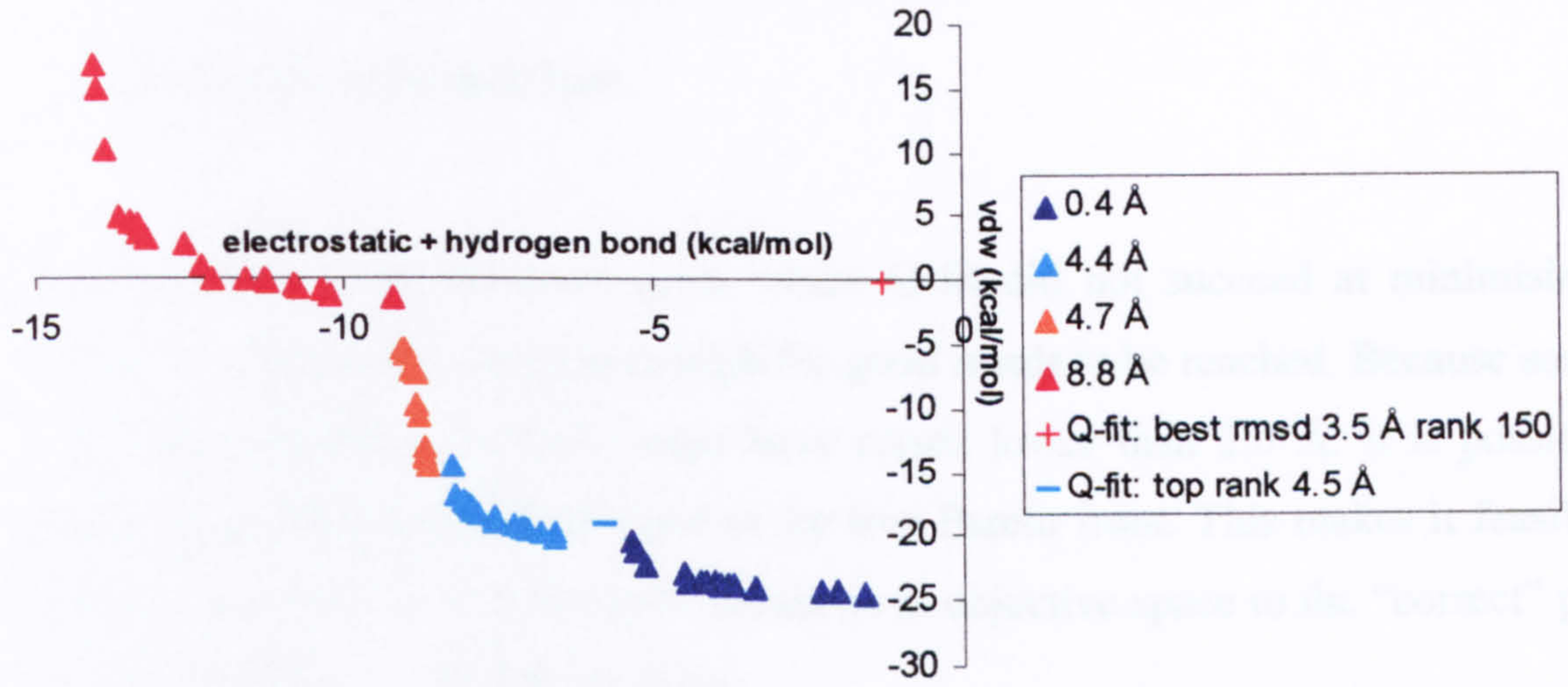


Figure 8-5 Pareto solutions produced by NSGA-II for 3hvt in objective space. The positions of the top-ranked Q-fit solution and the solution with the lowest rmsd obtained by Q-fit are also shown.

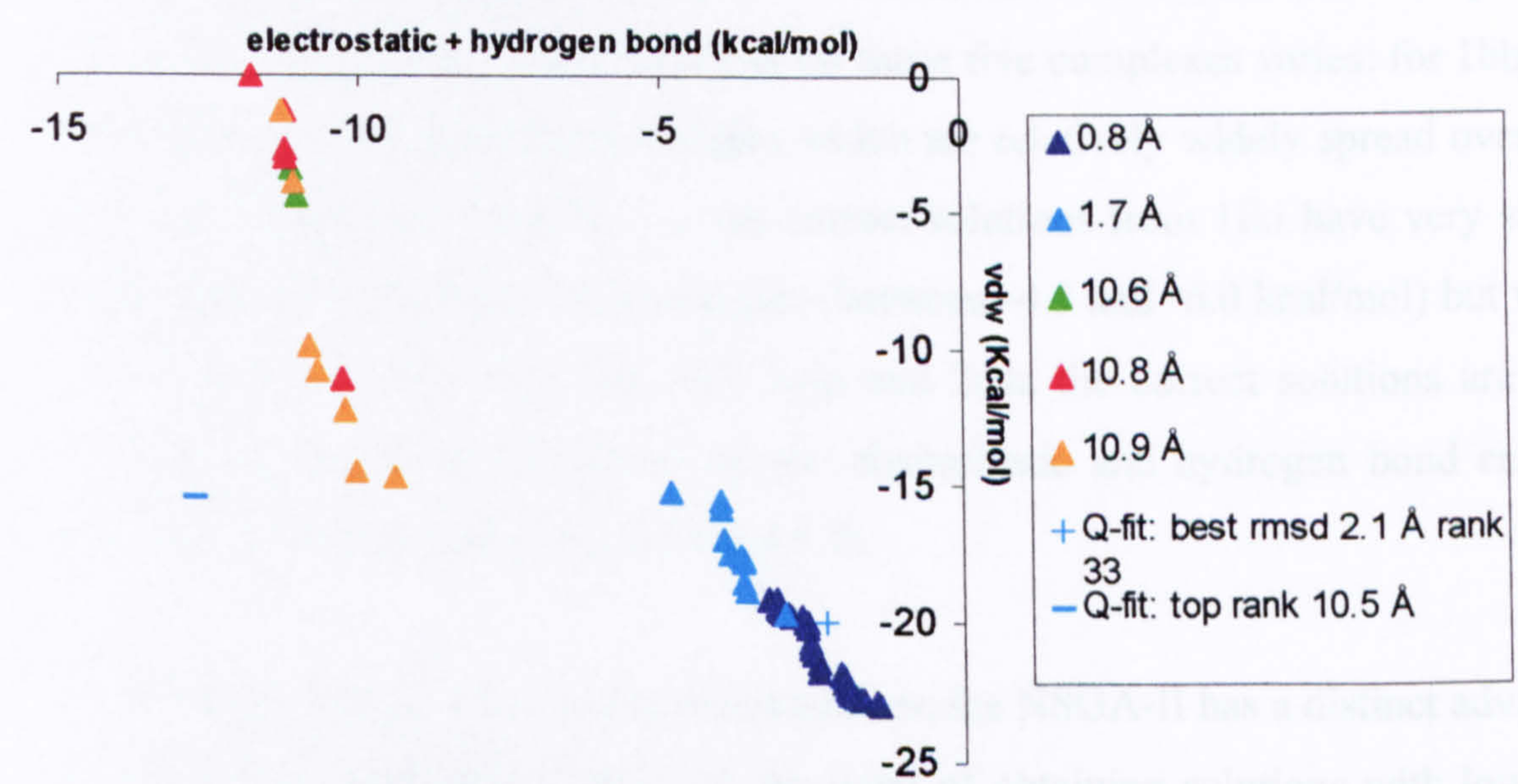


Figure 8-6 Pareto solutions produced by NSGA-II for 1igj in objective space. The positions of the top-ranked Q-fit solution and the solution with the lowest rmsd obtained by Q-fit are also shown.

8.2.3 Q-fit solutions with objectives not fully minimised: 1bbp, 1glp, 1fki, 2ada, 1rne and 1snc

The next few plots illustrate cases where Q-fit did not succeed at minimising its solutions' interaction energies enough for good rmsds to be reached. Because some of the Pareto solutions in these cases have rmsds lower than 2.0 Å, it is possible to presume that they have converged to the true Pareto front. This makes it feasible to compare the positions of the Q-fit solutions in objective space to the "correct" points of convergence, i.e. the Pareto front.

For the following five cases, 1bbp, 1glp, 1fki, 2ada, 1rne and 1snc Q-fit did not find solutions which have rmsds less than 2.0 Å. The positions in objective space of the solutions output from NSGA-II and Q-fit (Figures 8.7 to 8.12) show that both the top-ranked and the lowest rmsd Q-fit solutions have not advanced as far as the Pareto front. The high energies of the Q-fit solutions are observed in both objectives. With 2ada the two Q-fit solutions have positive values for both objectives. The spread of the clusters containing correct solutions for these five complexes varies: for 1bbp and 1rne the correct solutions have energies which are relatively widely spread over both objectives (Figures 8.7 and 8.11) , the correct solutions from 1fki have very similar electrostatic and hydrogen bond energies (between -4.5 and -6.0 kcal/mol) but varied vdw energies (Figure 8.8); and with 1glp and 2ada the correct solutions are close together and are more influenced by the electrostatic and hydrogen bond energies than vdw interactions (Figures 8.10 and 8.9).

These results indicate that, for these complexes, the NSGA-II has a distinct advantage over Q-fit in minimising energies to the point of obtaining solutions with low rmsd values from the crystal structure.

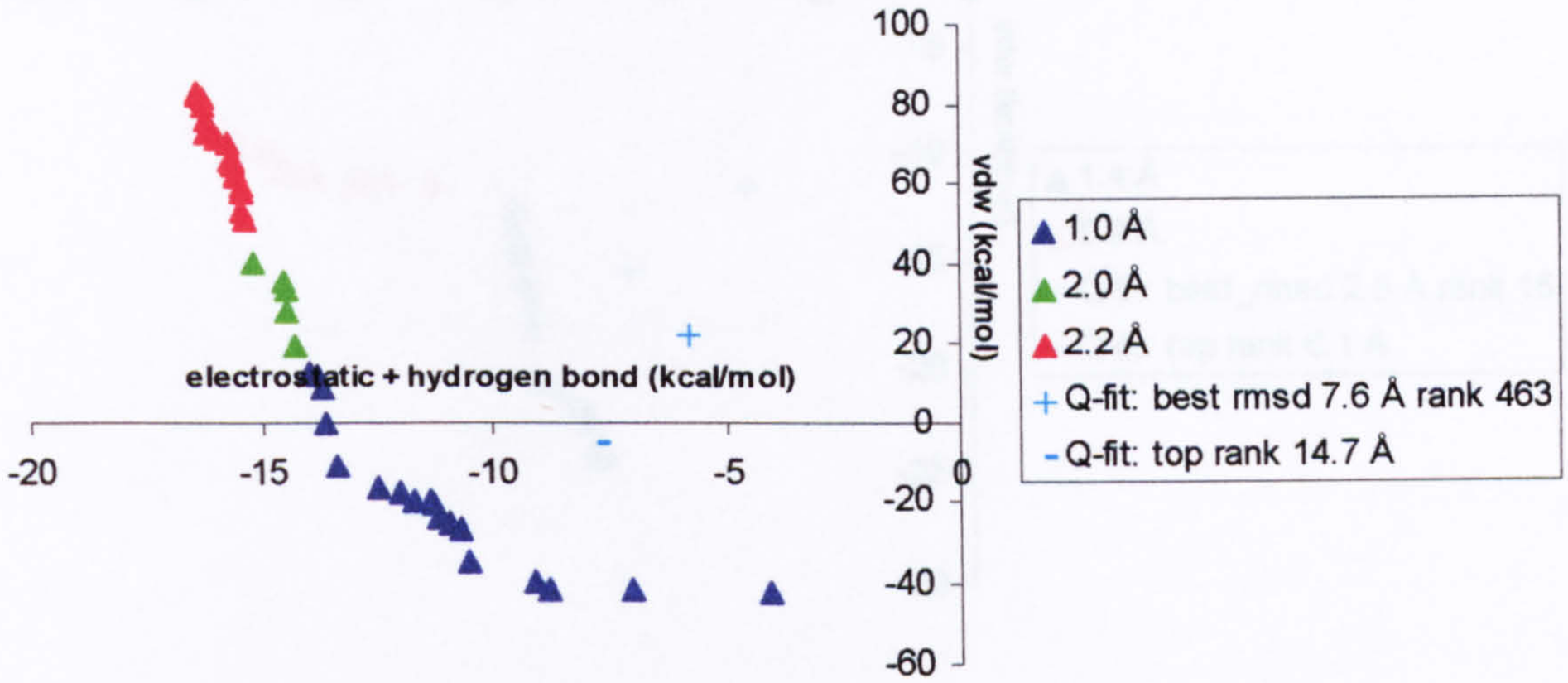


Figure 8-7 Pareto solutions produced by NSGA-II for 1bbp in objective space. The positions of the top-ranked Q-fit solution and the solution with the lowest rmsd obtained by Q-fit are also shown.

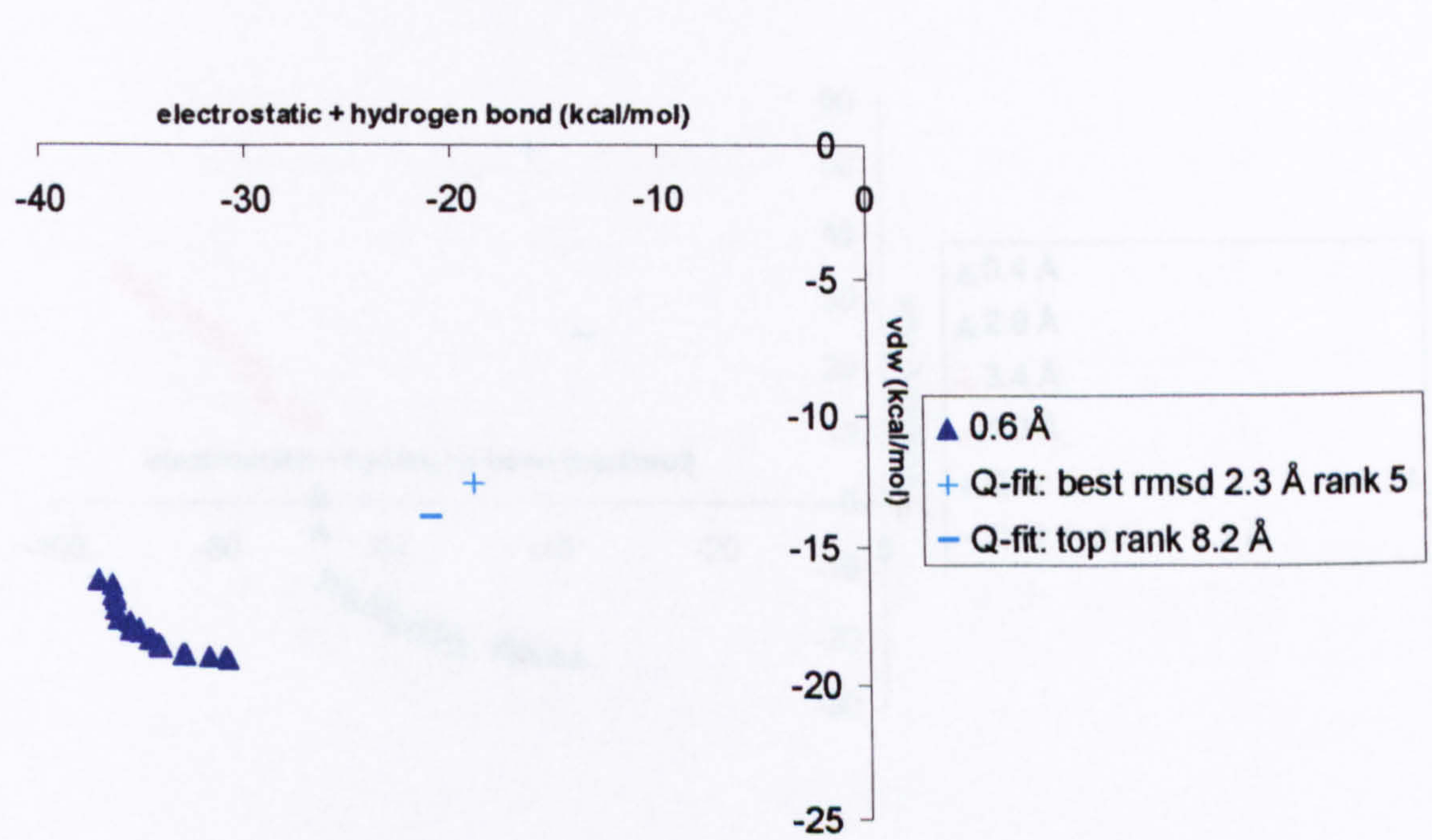


Figure 8-8 Pareto solutions produced by NSGA-II for 1glp in objective space. The positions of the top-ranked Q-fit solution and the solution with the lowest rmsd obtained by Q-fit are also shown

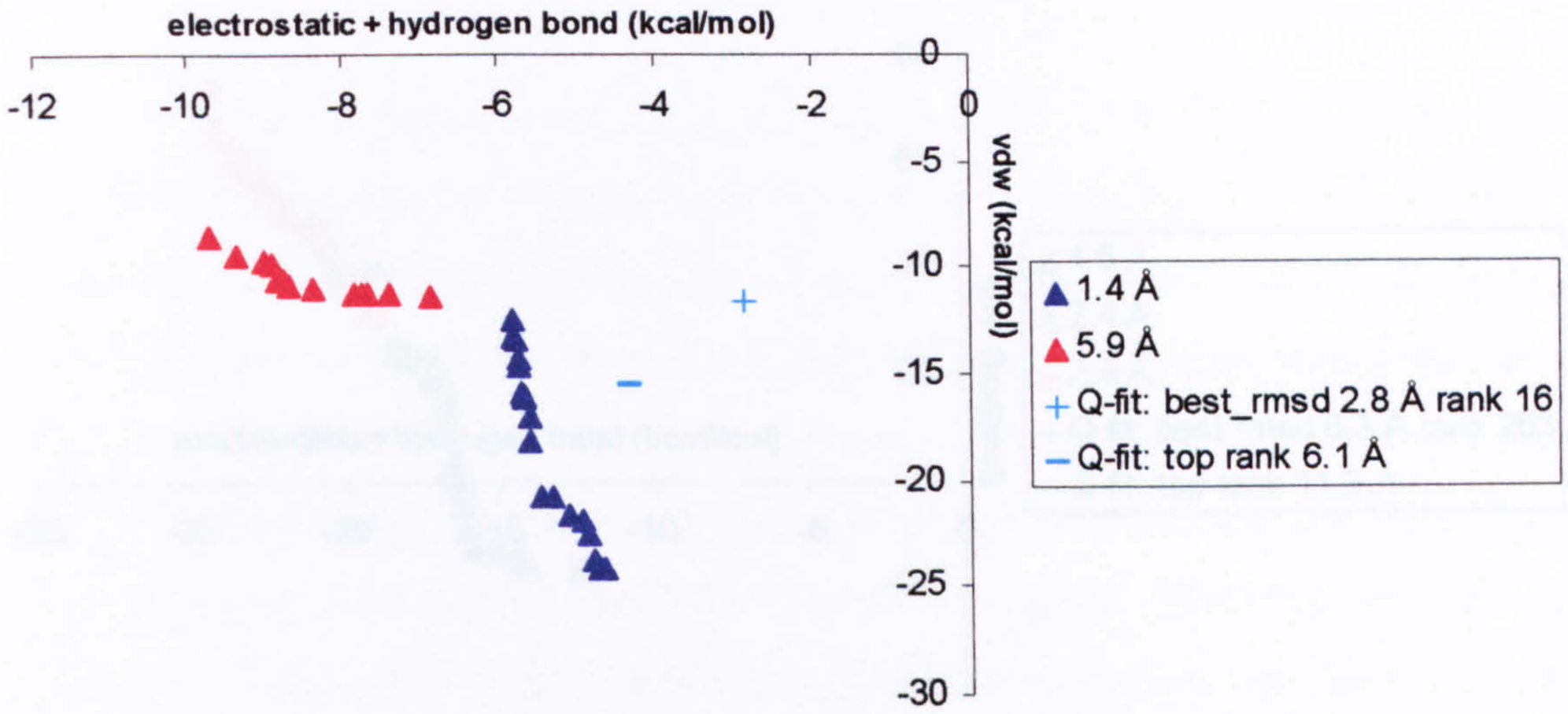


Figure 8-9 Pareto solutions produced by NSGA-II for 1fki in objective space. The positions of the top-ranked Q-fit solution and the solution with the lowest rmsd obtained by Q-fit are also shown.

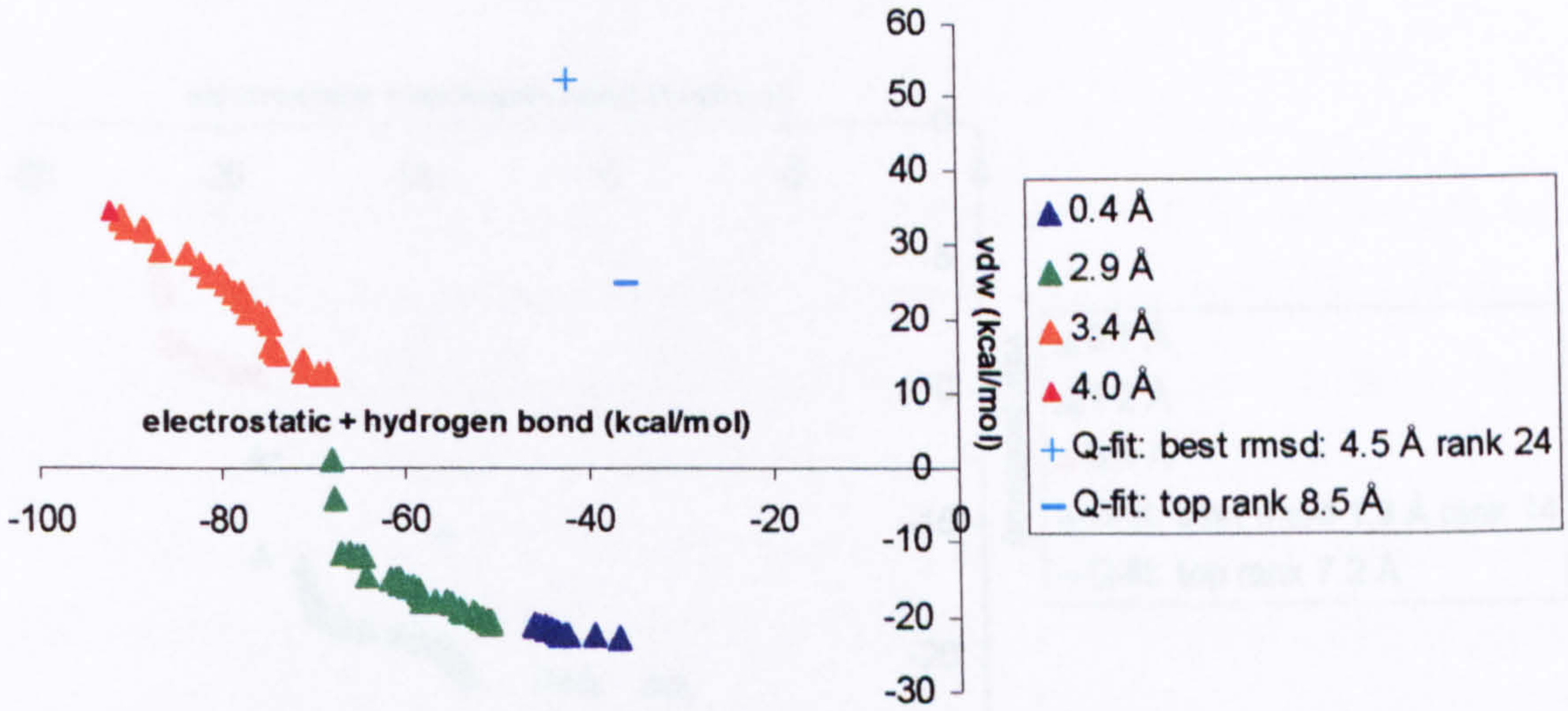


Figure 8-10 Pareto solutions produced by NSGA-II for 2ada in objective space. The positions of the top-ranked Q-fit solution and the solution with the lowest rmsd obtained by Q-fit are also shown.

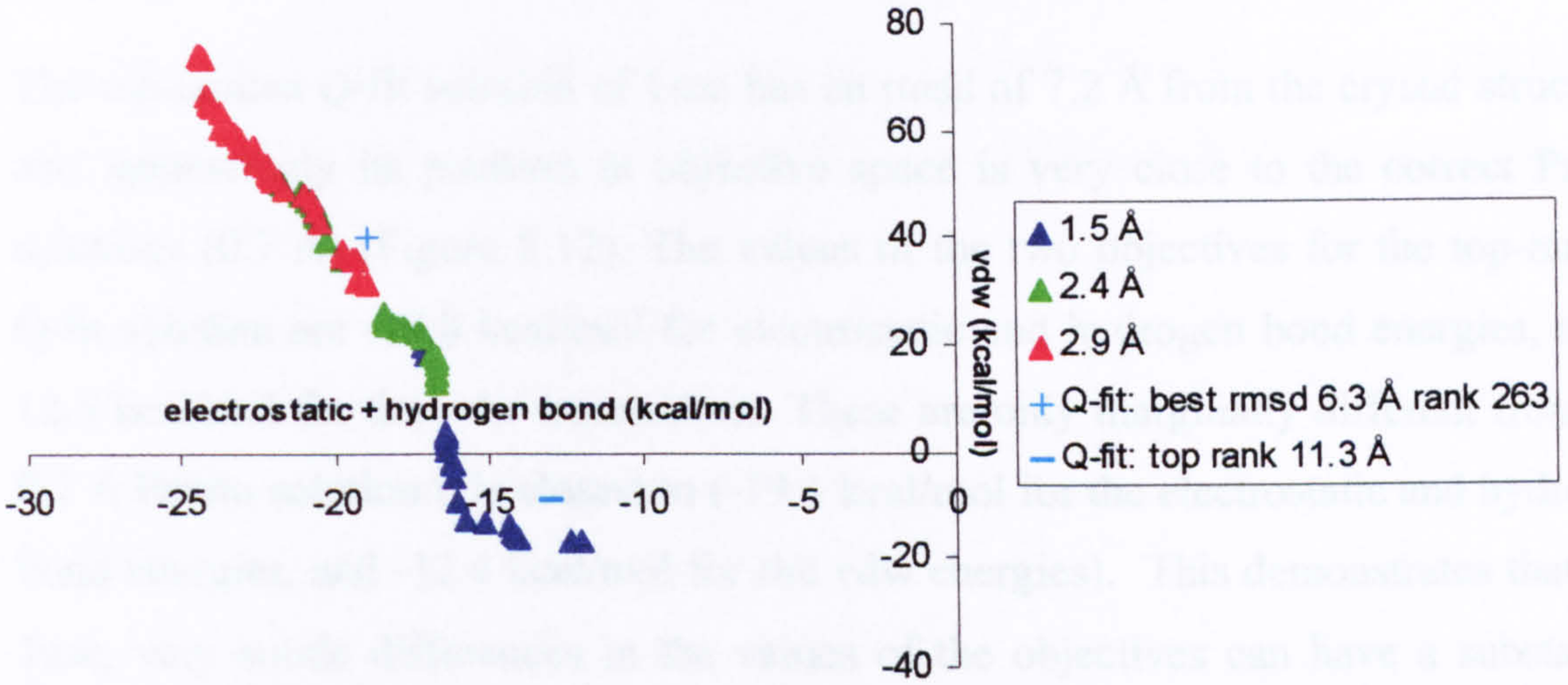


Figure 8-11 Pareto solutions produced by NSGA-II for 1rne in objective space. The positions of the top-ranked Q-fit solution and the solution with the lowest rmsd obtained by Q-fit are also shown.

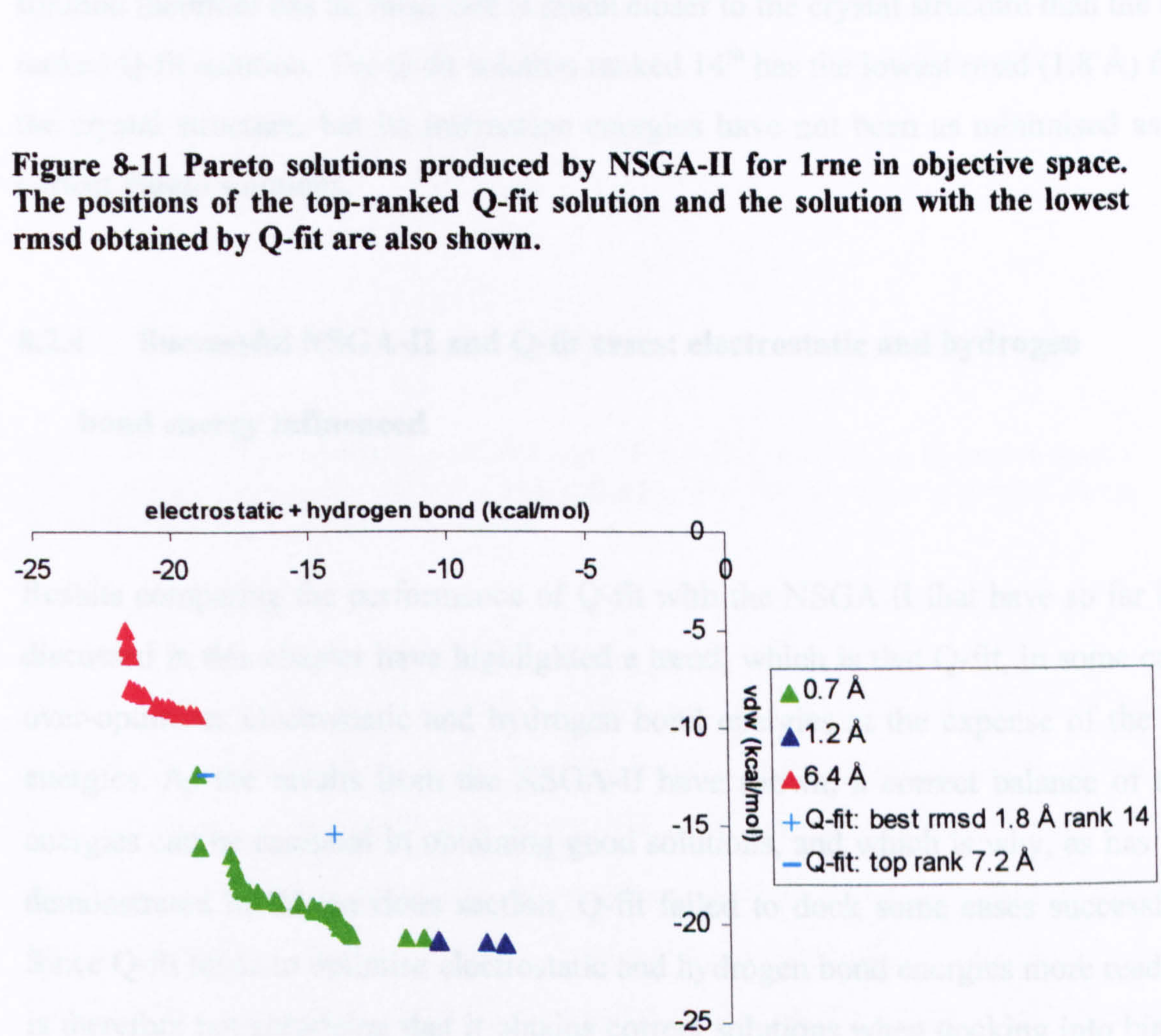


Figure 8-12 Pareto solutions produced by NSGA-II for 1snc in objective space. The positions of the top-ranked Q-fit solution and the solution with the lowest rmsd obtained by Q-fit are also shown.

The top-ranked Q-fit solution of 1snc has an rmsd of 7.2 Å from the crystal structure, and interestingly its position in objective space is very close to the correct Pareto solutions (0.7 Å) (Figure 8.12). The values of the two objectives for the top-ranked Q-fit solution are -18.8 kcal/mol for electrostatic and hydrogen bond energies, and -12.5 kcal/mol for the vdw interactions. These are only marginally different from the 0.7 Å Pareto solution it is closest to (-19.1 kcal/mol for the electrostatic and hydrogen bond energies, and -12.4 kcal/mol for the vdw energies). This demonstrates that, for 1snc, very subtle differences in the values of the objectives can have a substantial effect in the orientations of the solutions. The slightly more minimised Pareto solution therefore has an rmsd that is much closer to the crystal structure than the top-ranked Q-fit solution. The Q-fit solution ranked 14th has the lowest rmsd (1.8 Å) from the crystal structure, but its interaction energies have not been as minimised as the correct Pareto solutions.

8.2.4 Successful NSGA-II and Q-fit cases: electrostatic and hydrogen bond energy influenced

Results comparing the performance of Q-fit with the NSGA-II that have so far been discussed in this chapter have highlighted a trend, which is that Q-fit, in some cases, over-optimises electrostatic and hydrogen bond energies at the expense of the vdw energies. As the results from the NSGA-II have shown, a correct balance of these energies can be essential in obtaining good solutions, and which is why, as has been demonstrated in the previous section, Q-fit failed to dock some cases successfully. Since Q-fit tends to optimise electrostatic and hydrogen bond energies more readily it is therefore not surprising that it obtains correct solutions when docking into binding sites that are more influenced by these interactions.

The following plots illustrate cases where the electrostatic and hydrogen bond energies are the more influential objective, and where both Q-fit and the NSGA-II are successful in obtaining solutions with good rmsds. Figure 8.13 shows the Pareto solutions obtained when docking 1mld, and as their positions in objective space

indicate, the electrostatic and hydrogen bond energies have a stronger influence in obtaining these Pareto solutions than the vdw energies. Also the Pareto solutions are spread over a wider range with the electrostatic and hydrogen bond energies than with the vdw energies.

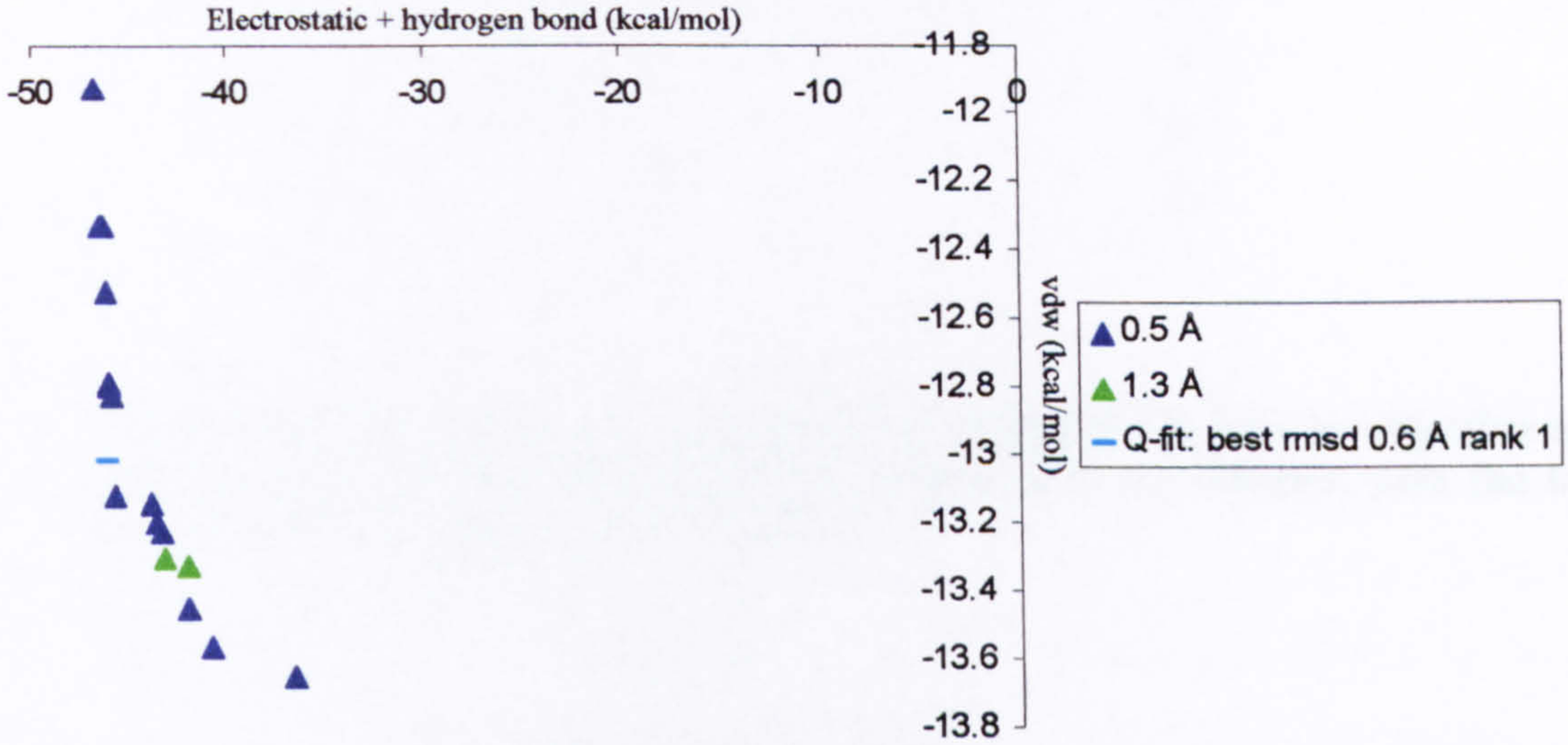


Figure 8-13 Pareto solutions produced by NSGA-II for 1mld in objective space. The positions of the top-ranked Q-fit solution and the solution with the lowest rmsd obtained by Q-fit are also shown.

With 1nis (Figure 8.14), one correct cluster was produced by the NSGA-II and this cluster is also more influenced by the electrostatic and hydrogen bond energies. With 5cts two clusters were produced by the NSGA-II with approximate rmsds of 1.9 Å and 3.0 Å (Figure 8.15). Both of these clusters are closer to each other in objective space, and they are also more influenced by electrostatic and hydrogen bond energies than by vdw interactions. It is interesting to note that the top-ranked Q-fit solution has lower electrostatic and hydrogen bond energies than the Pareto solutions, and that it also has a lower rmsd (0.5 Å). This implies that the Pareto front from 5cts may not have converged to the true Pareto front.

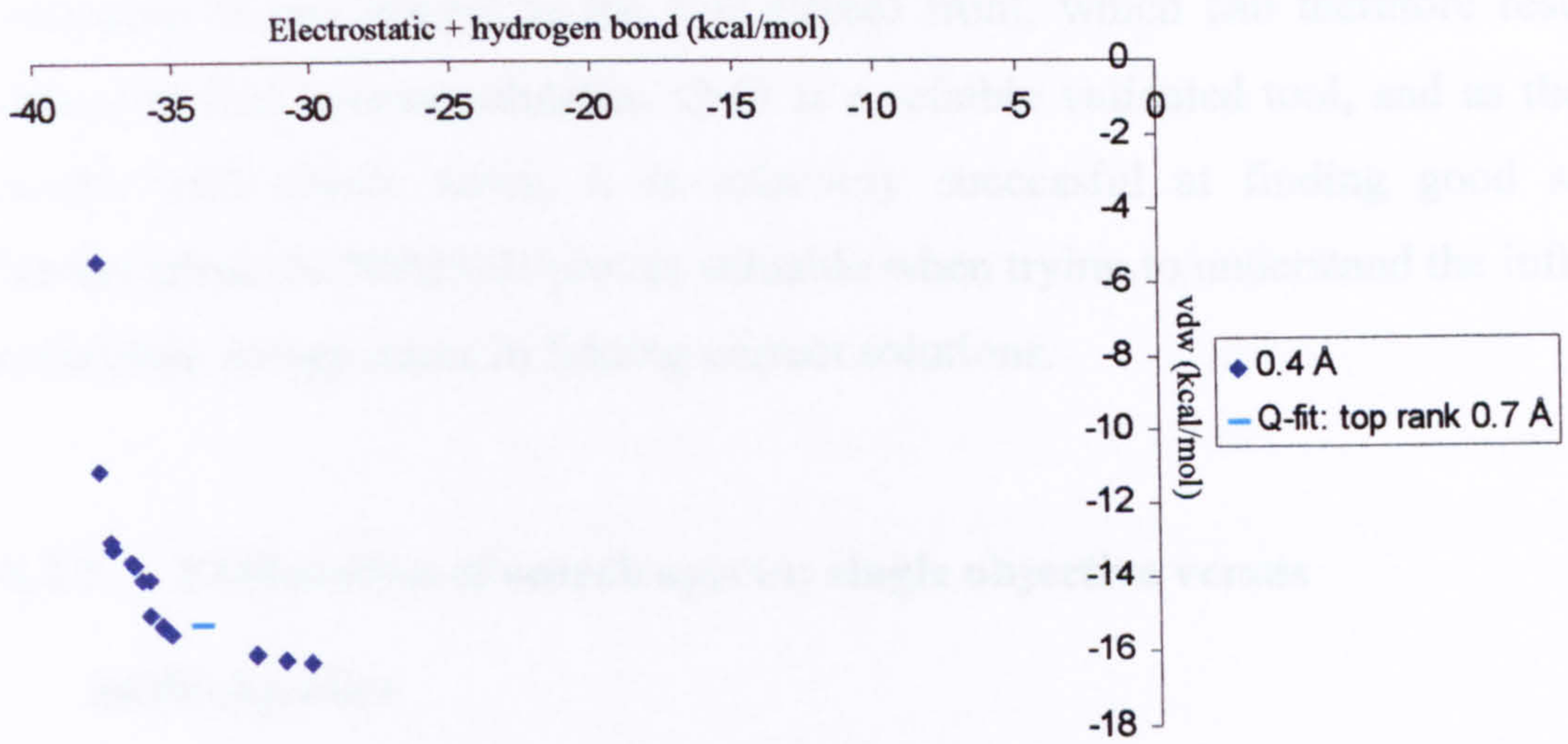


Figure 8-14 Pareto solutions produced by NSGA-II for 1nis in objective space. The positions of the top-ranked Q-fit solution and the solution with the lowest rmsd obtained by Q-fit are also shown.

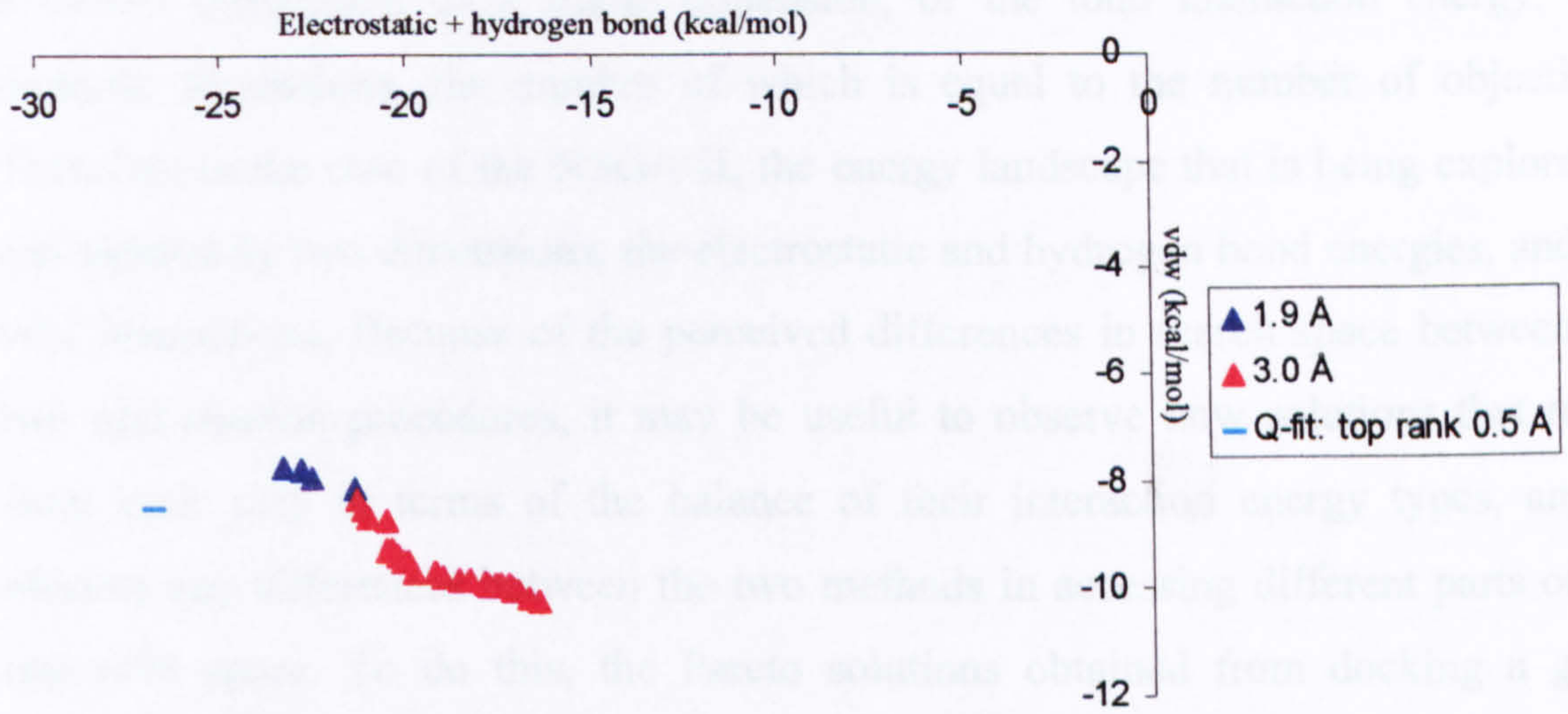


Figure 8-15 Pareto solutions produced by NSGA-II for 5cts in objective space. The positions of the top-ranked Q-fit solution and the solution with the lowest rmsd obtained by Q-fit are also shown.

There are several cases from within the FlexX dataset which Q-fit docks and the NSGA-II does not, and cases for which both algorithms fail to achieve good solutions. There are several reasons why the NSGA-II may have failed and Q-fit succeeded, the most likely reason being that the NSGA-II did not converge to the true Pareto front. As some previous test cases have shown, the NSGA-II is at times

incapable of converging to the true Pareto front, which can therefore result in its failure to find correct solutions. Q-fit is a reliable validated tool, and as the overall results with FlexX show, it is relatively successful at finding good solutions. Nevertheless the NSGA-II proves valuable when trying to understand the influence of individual energy terms in finding correct solutions.

8.2.5 Exploration of search spaces: single objective versus multiobjective

A given search space can be regarded differently depending on whether single or multiobjective optimisation is being employed. In single objective optimisation, changes in conformation explore an energy landscape in one dimension, which represents the total interaction energy. Multiobjective optimisation can be regarded as a further breakdown of a single dimension, or the total interaction energy, into multiple dimensions, the number of which is equal to the number of objectives. Therefore in the case of the NSGA-II, the energy landscape that is being explored is represented by two dimensions, the electrostatic and hydrogen bond energies, and the vdw interactions. Because of the perceived differences in search space between the two optimisation procedures, it may be useful to observe how solutions that result from each vary in terms of the balance of their interaction energy types, and to observe any differences between the two methods in accessing different parts of the objective space. To do this, the Pareto solutions obtained from docking a given complex are plotted in objective space along with an equal number of top-ranked Q-fit solutions obtained for the same complex. Complexes were selected which represent cases where:

a- both Q-fit and the NSGA-II were successful in finding solutions with good rmsds.

b- the NSGA-II found good solutions but Q-fit did not because it did not minimise solutions to the level of the Pareto front.

c- the NSGA-II succeeded in obtaining good solutions but Q-fit did not because the balance of energies of its top-ranked solutions is not correct.

Complexes with pdb codes 1fen, 1epb and 1xie were selected to represent cases (a), (b) and (c) respectively.

Figure 8.16 shows the Pareto solutions obtained when docking 1fen, as well as the top-ranked solutions obtained by Q-fit. The Q-fit solutions are evenly spread in objective space, and the top three Q-fit solutions, all of which have good rmsds, are at or very close to the Pareto front. The remaining Q-fit solutions' rmsds are above 2.0 Å.

1epb is a case where the top-ranked Q-fit solution was not minimised to the level of the Pareto front, and hence its top-ranked solution does not have a good rmsd (Figure 8.17). The correct Pareto clusters have lower vdw interactions than all of the Q-fit solutions, none of which have good rmsds. The spread of the Q-fit solutions indicates that Q-fit is attempting to optimise electrostatic and hydrogen bond interaction energies and not the vdw interactions, but, as the position of the correct Pareto cluster indicates, for this complex the vdw interactions need to be further minimised in order to obtain good solutions.

Figure 8.18 shows Pareto solutions and an equivalent number of Q-fit solutions obtained when docking 1xie. 1xie, as was described earlier, is a complex which NSGA-II docks but Q-fit does not. As was also discussed earlier, from the distributions of the solutions it has been inferred that Q-fit does not succeed at docking this molecule because it does not obtain a correct balance of energies for any of its solutions. The largest correct Pareto solution cluster has an rmsd of 2.0 Å (turquoise blue triangles in Figure 8.18- there is also a smaller correct Pareto cluster with an rmsd of 1.2 Å (dark blue triangle in Figure 8.18). None of the Q-fit solutions are close to the 1.2 Å cluster. There are however, two Q-fit solutions that are near the 2.0 Å cluster (orange circles). The rmsds of these solutions are not, as might be expected, close to 2.0 Å- they are 3.1 Å and 4.3 Å (ranked 26 and 25 respectively). The figure shows that the good Pareto solutions have slightly lower vdw energies than the two Q-fit solutions, and examining the energies of the Q-fit solutions more closely shows that these have vdw interaction energies that are very slightly higher in magnitude (by approximately 0.1 kcal/mol) than the Pareto solutions. Evidently the slight difference in energies between the two sets of solutions affects their

orientations, so that the Pareto solutions have rmsds that are closer to the crystal than the Q-fit solutions. The Q-fit solutions are spread out evenly in objective space, and, as the figure shows, the top-ranked Q-fit solution does not have the same balance of energies as the correct Pareto clusters, thus resulting in an rmsd that is high relative to the ligand crystal structure.

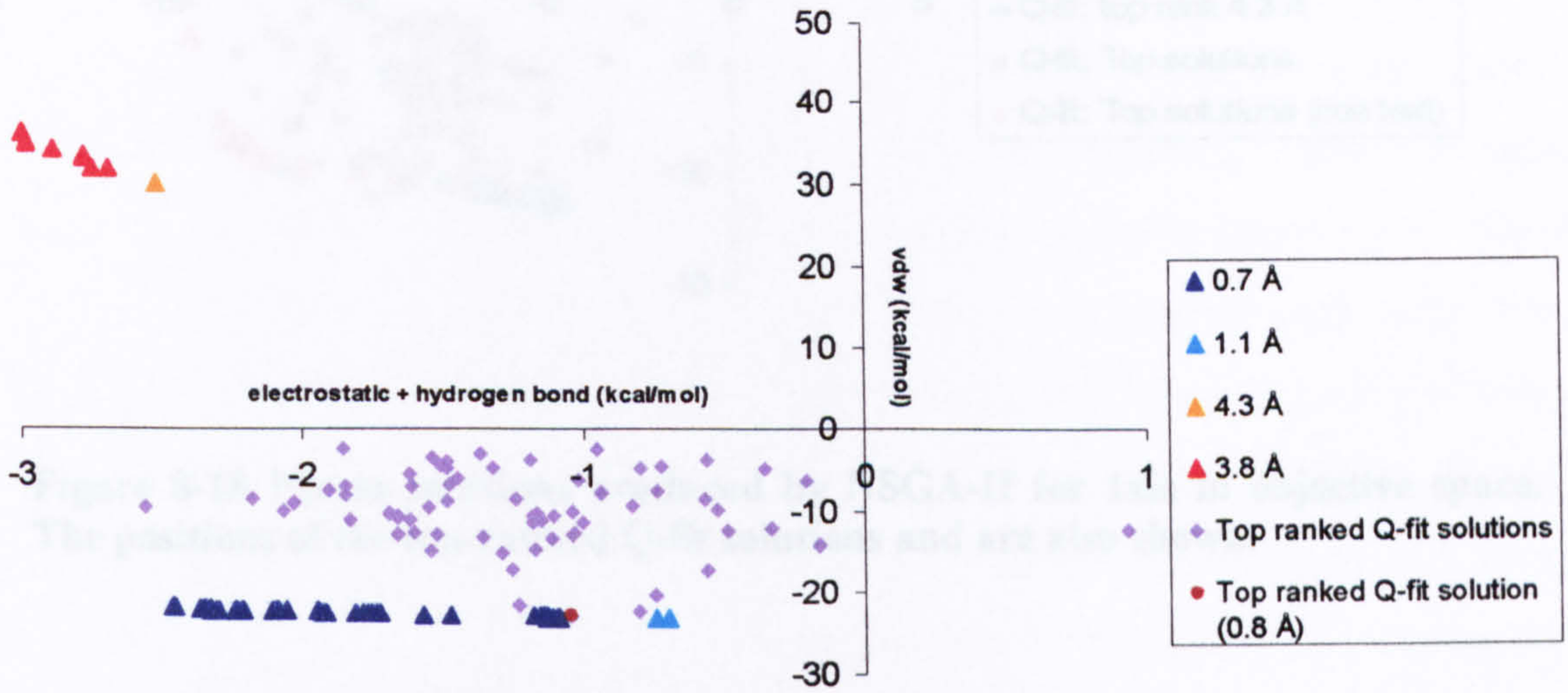


Figure 8-16 Pareto solutions produced by NSGA-II for 1fen in objective space. The positions of the top-ranked Q-fit solutions are also shown.

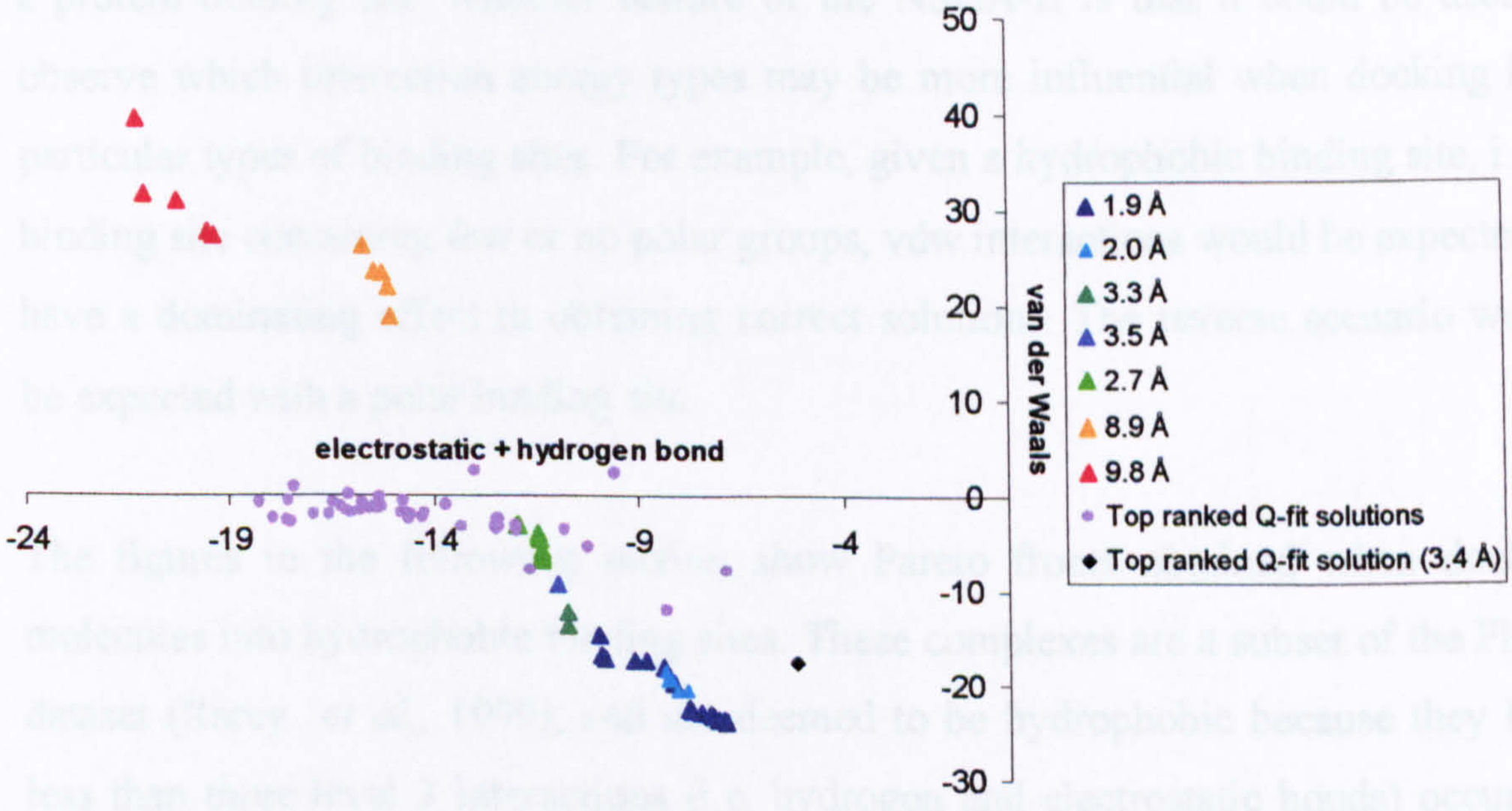


Figure 8-17 Pareto solutions produced by NSGA-II for 1epb in objective space. The positions of the top-ranked Q-fit solutions are also shown

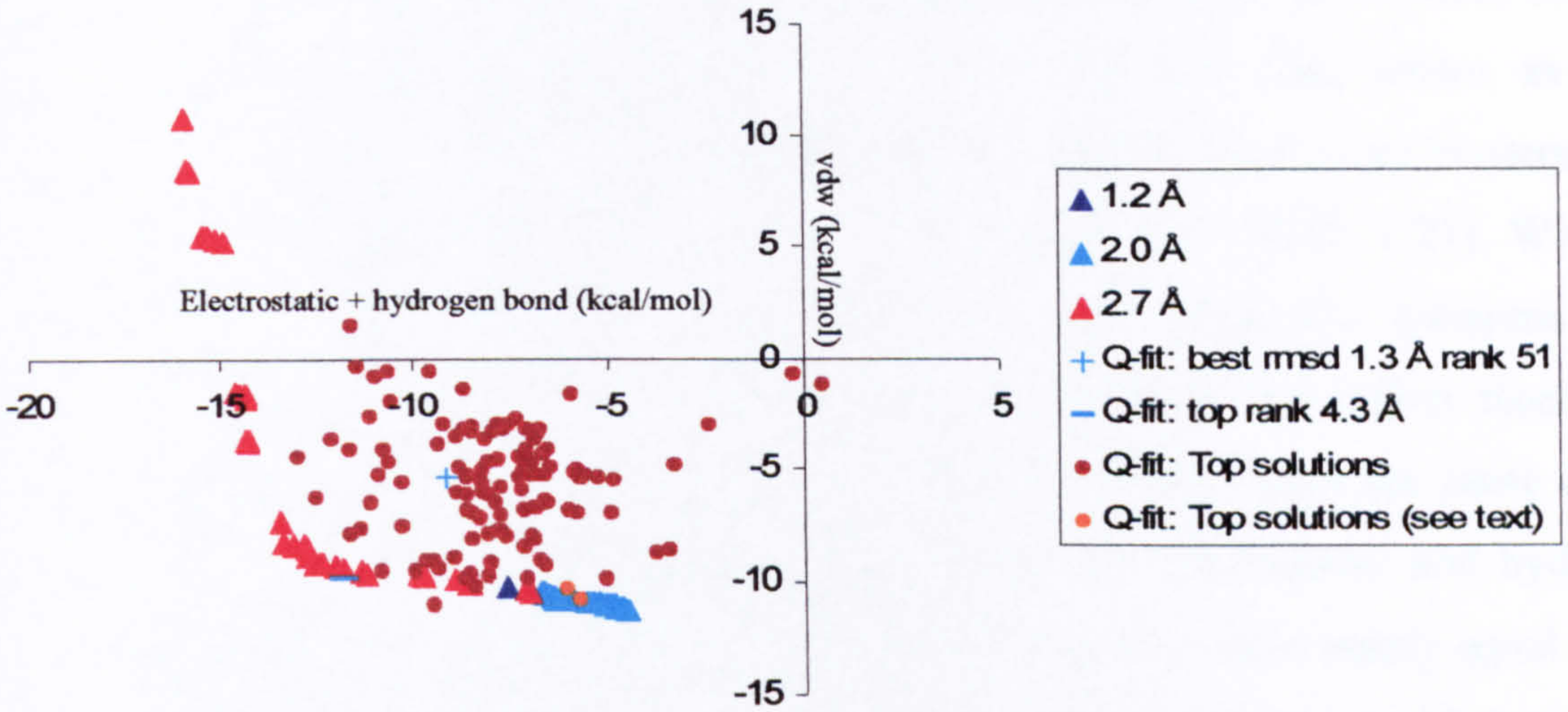


Figure 8-18 Pareto solutions produced by NSGA-II for 1xie in objective space. The positions of the top-ranked Q-fit solutions and are also shown.

8.2.6 Hydrophobic binding sites

The results discussed so far have shown that the NSGA-II is a useful tool for observing the influence specific interactions have in driving the binding of a ligand to a protein binding site. Another feature of the NSGA-II is that it could be used to observe which interaction energy types may be more influential when docking into particular types of binding sites. For example, given a hydrophobic binding site, i.e. a binding site containing few or no polar groups, vdw interactions would be expected to have a dominating effect in obtaining correct solutions. The reverse scenario would be expected with a polar binding site.

The figures in the following section show Pareto fronts obtained when docking molecules into hydrophobic binding sites. These complexes are a subset of the FlexX dataset (Rarey *et al.*, 1999), and are deemed to be hydrophobic because they have less than three level 3 interactions (i.e. hydrogen and electrostatic bonds) occurring between the ligand and protein.

Looking at the Pareto fronts obtained by docking these complexes, it can be seen that the vdw interactions do have a predominating effect on the correct clusters. This scenario is particularly noticeable with 1rbp 1dbb and 1fen, where, as the figures show, the electrostatic and hydrogen bond energies of the correct clusters are several magnitudes higher than vdw interactions (Figures 8.19, 8.20, 8.21). With 1epb the correct cluster (1.9 Å) is spread over a wider range- the solutions which are dominated by vdw interactions have low vdw energies and higher electrostatic and hydrogen bond energies, and the rest of the solutions within the same cluster have gradually increasing vdw energies and decreasing electrostatic and hydrogen bond energies, reaching a point where some of the solutions have nearly equal magnitudes in both objectives (Figure 8.22). The same situation is observed with 1mbi, where the correct clusters (1.1 Å and 1.7 Å) contain solutions with predominating vdw energies (at the right edge of the Pareto front), and also contain solutions which are spread so that, overall, the solutions have gradually increasing vdw energies, and decreasing electrostatic and hydrogen bond energies (Figure 8.23). With 1ack, unlike the other complexes within this subset, all of the Pareto solutions have good rmsds, and these are also spread in objective space: those on the right edge of the Pareto front are influenced by vdw energies, and other solutions are spread over a wider range in objective space (Figure 8.25).

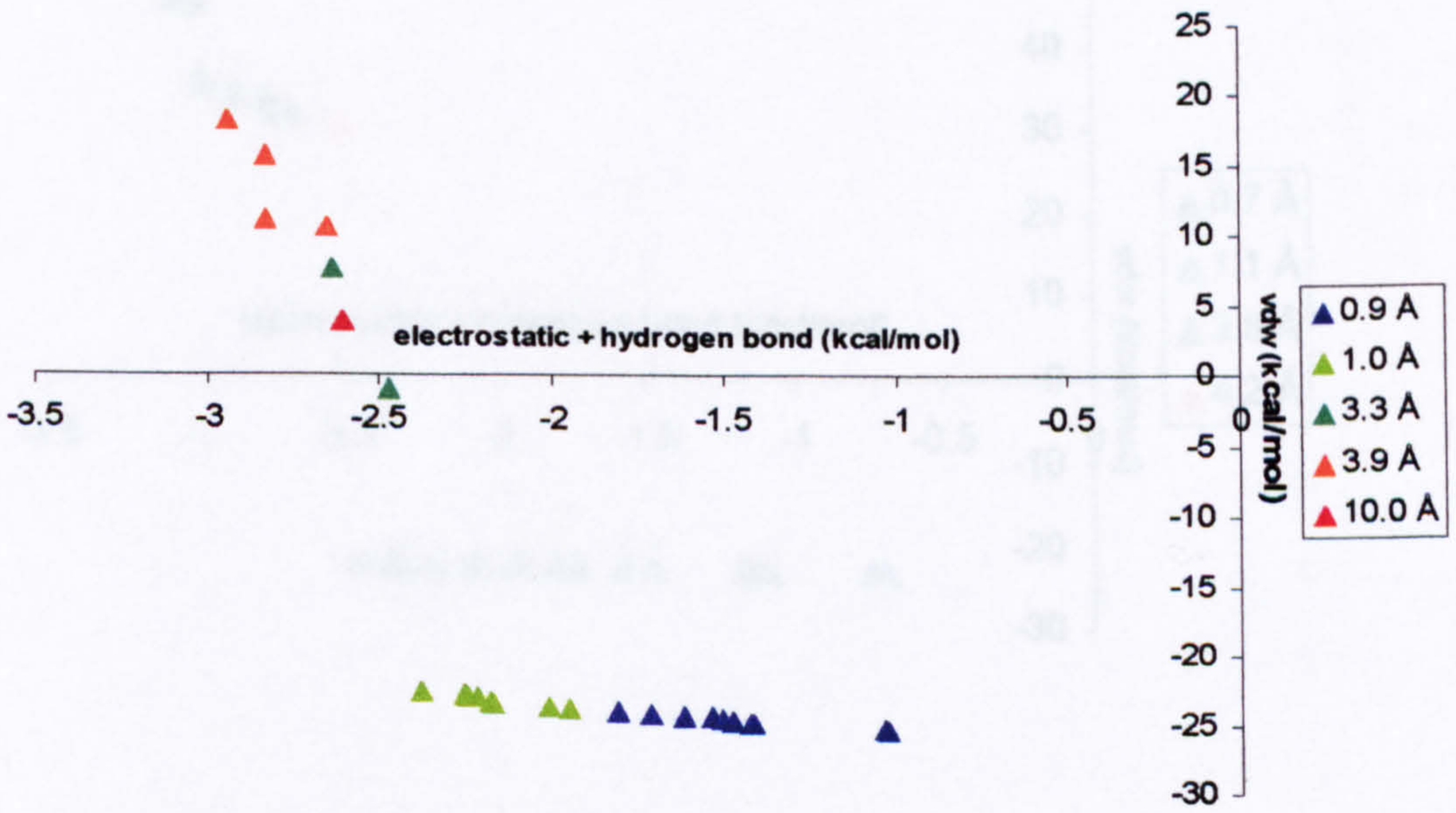


Figure 8-19 Pareto solutions produced by NSGA-II for 1rbp, a protein with a hydrophobic binding site, in objective space.

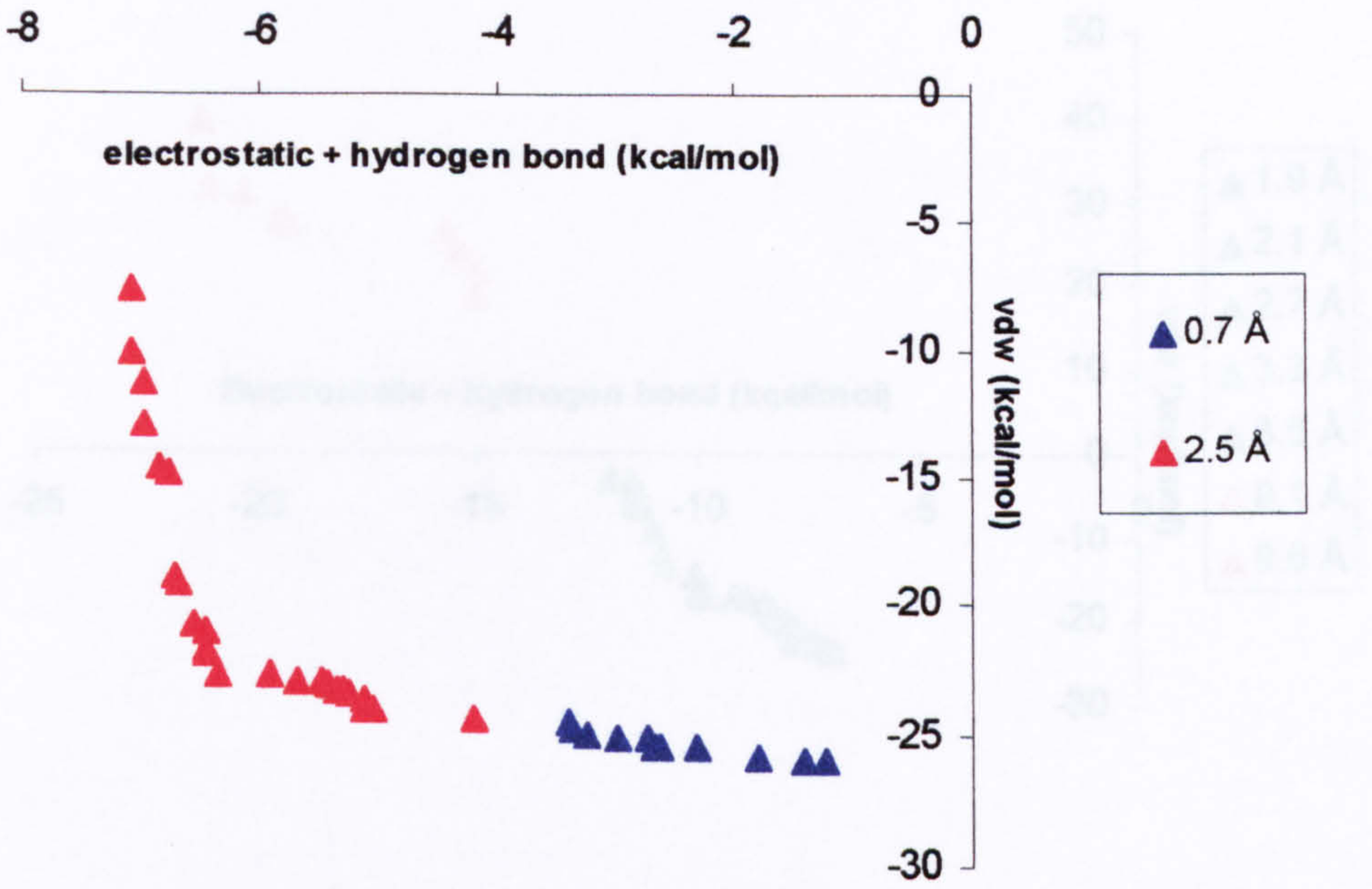


Figure 8-20 Pareto solutions produced by NSGA-II for 1dbb, a protein with a hydrophobic binding site, in objective space.

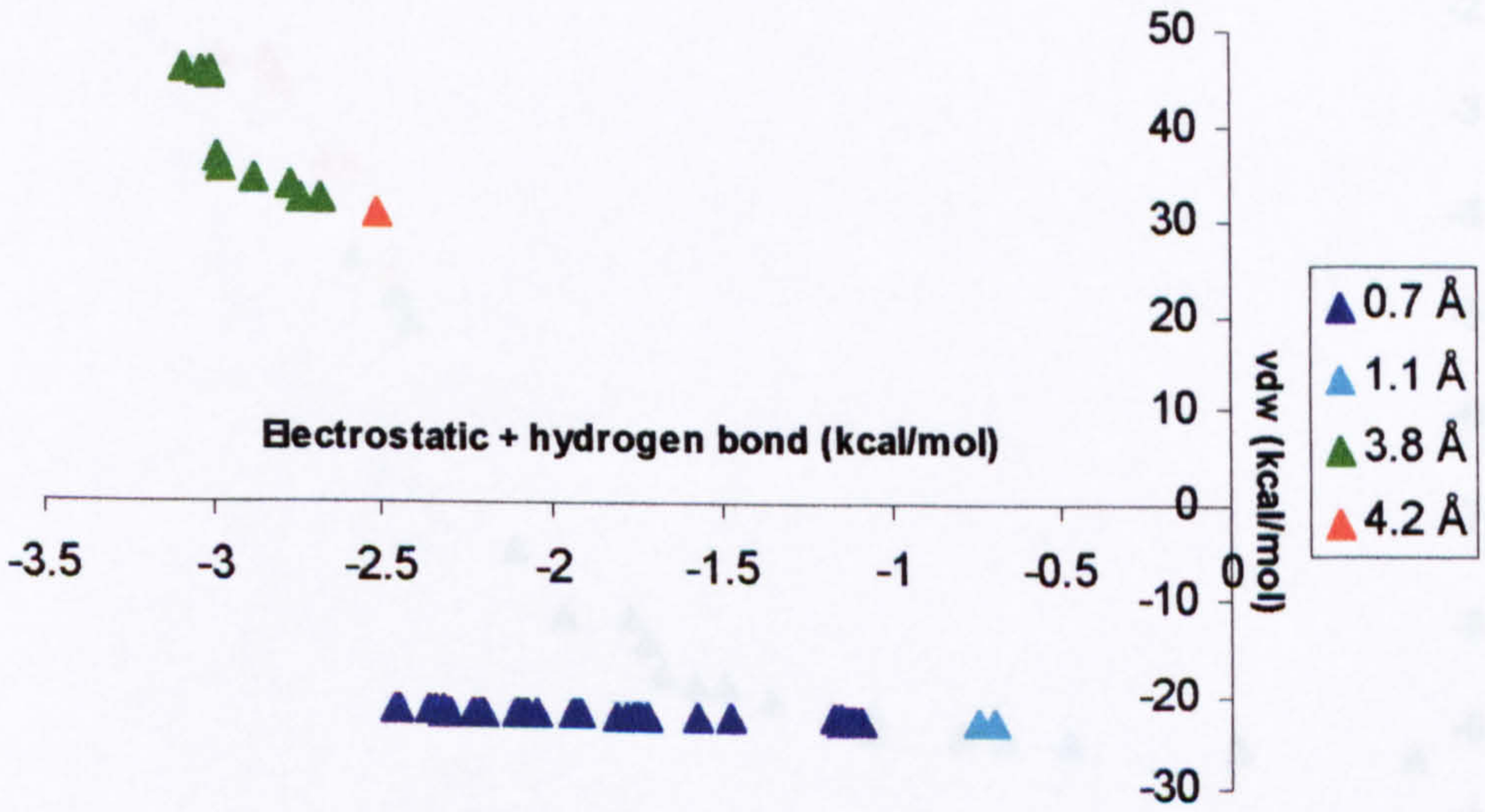


Figure 8-21 Pareto solutions produced by NSGA-II for 1fen, a protein with a hydrophobic binding site, in objective space.

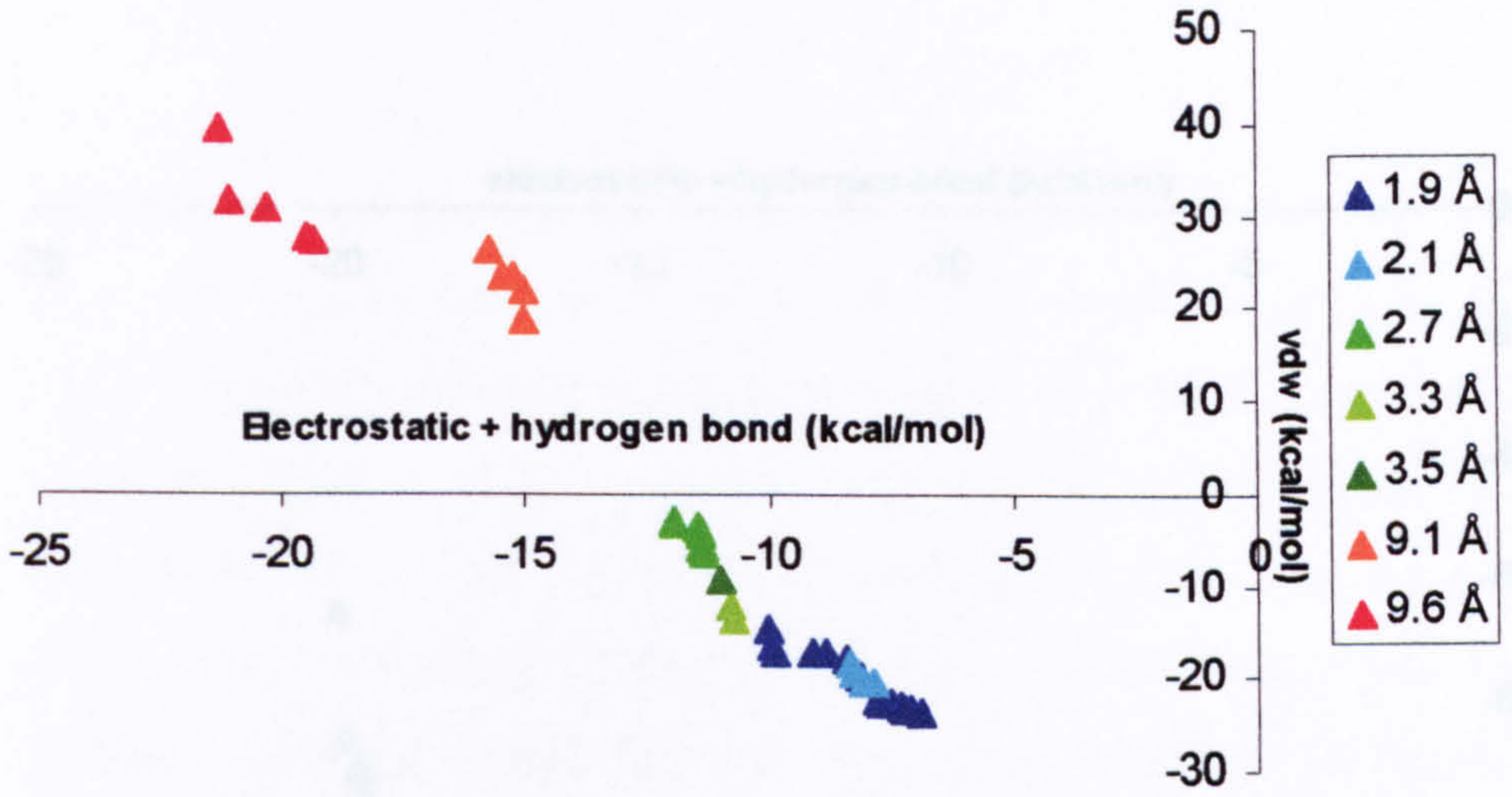


Figure 8-22 Pareto solutions produced by NSGA-II for 1epb, a protein with a hydrophobic binding site, in objective space.

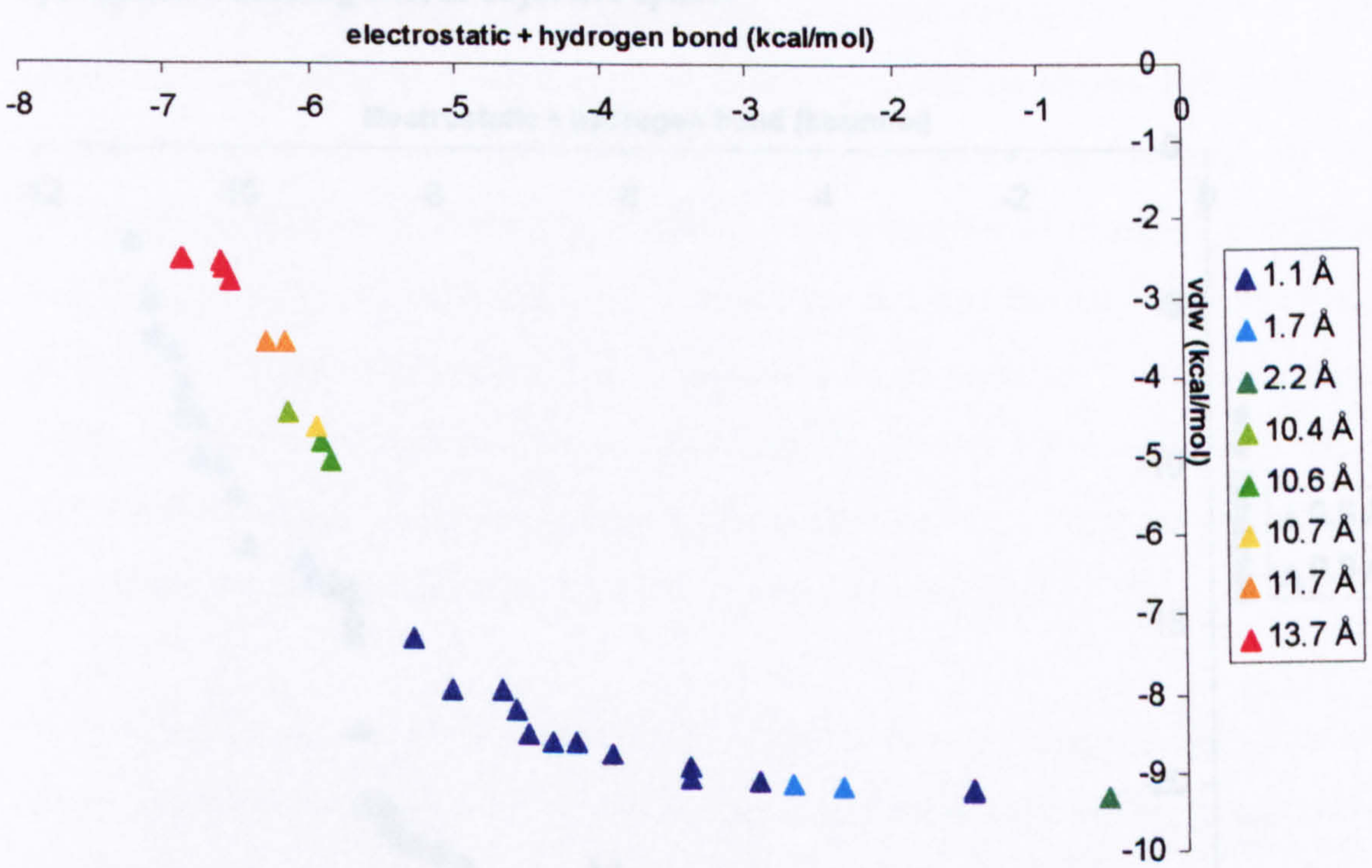


Figure 8-23 Pareto solutions produced by NSGA-II for 1mbi, a protein with a hydrophobic binding site, in objective space.

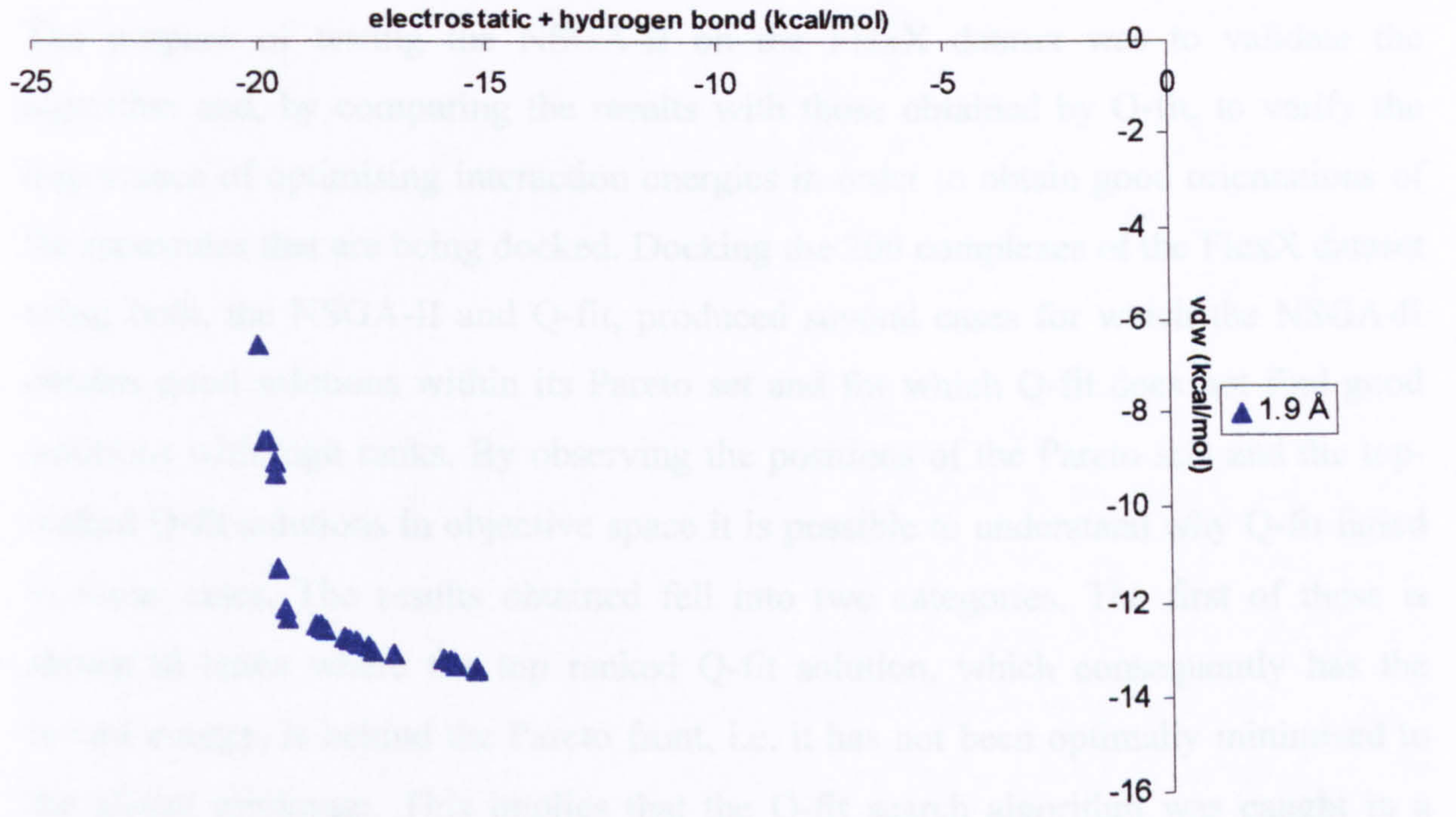


Figure 8-24 Pareto solutions produced by NSGA-II for 1pbd, a protein with a hydrophobic binding site, in objective space.

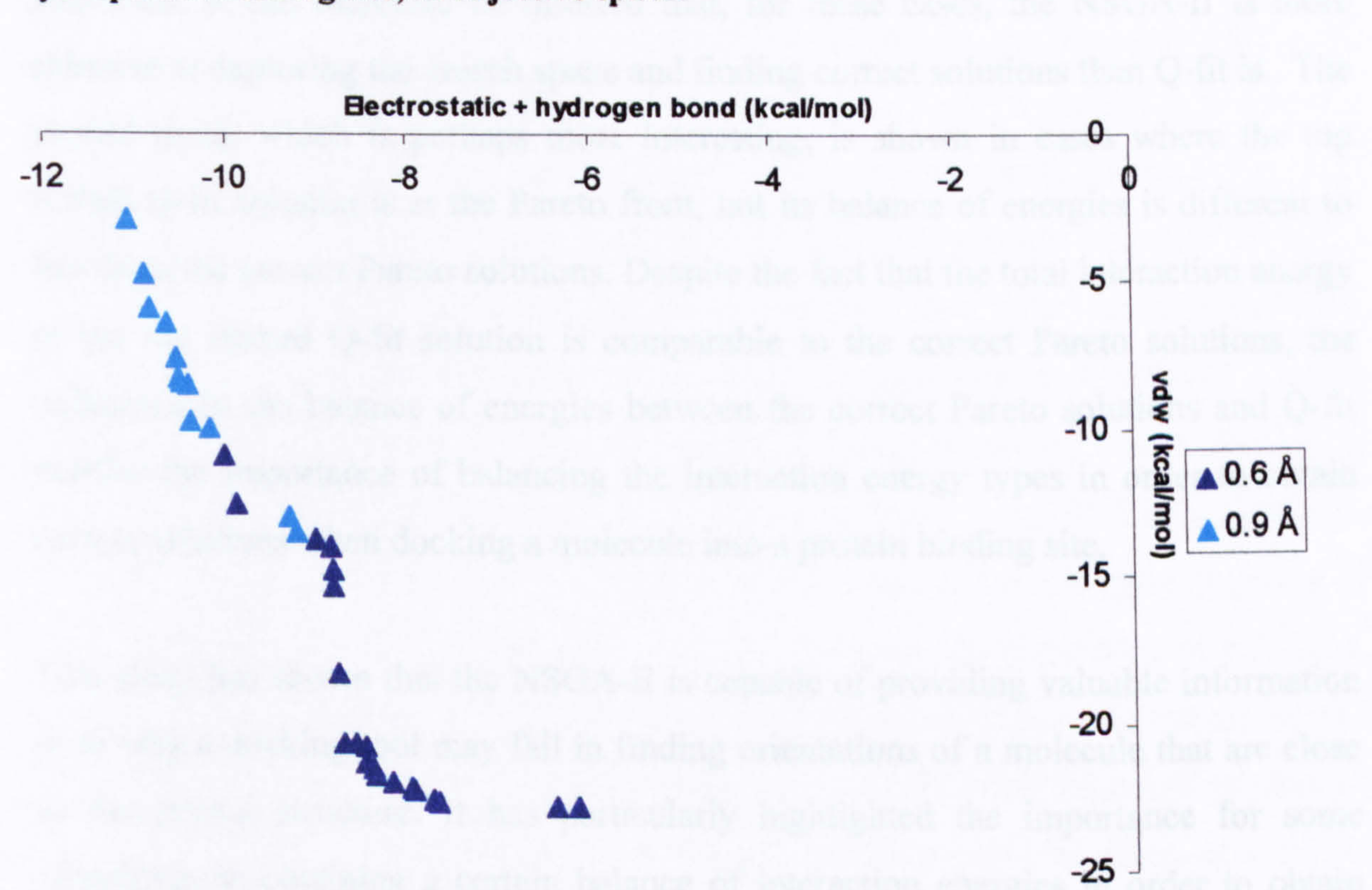


Figure 8-25 Pareto solutions produced by NSGA-II for 1ack, a protein with a hydrophobic binding site, in objective space.

8.2.7 Discussion of results obtained with FlexX Dataset

The purpose of testing the NSGA-II on the FlexX dataset was to validate the algorithm and, by comparing the results with those obtained by Q-fit, to verify the importance of optimising interaction energies in order to obtain good orientations of the molecules that are being docked. Docking the 200 complexes of the FlexX dataset using both, the NSGA-II and Q-fit, produced several cases for which the NSGA-II obtains good solutions within its Pareto set and for which Q-fit does not find good solutions with high ranks. By observing the positions of the Pareto sets and the top-ranked Q-fit solutions in objective space it is possible to understand why Q-fit failed in those cases. The results obtained fell into two categories. The first of these is shown in cases where the top ranked Q-fit solution, which consequently has the lowest energy, is behind the Pareto front, i.e. it has not been optimally minimised to the global minimum. This implies that the Q-fit search algorithm was caught in a local minimum. Because the Pareto sets in these cases contain solutions with good rmsds it is possible to assume that these solutions are very close to, or at the global minimum. It can therefore be inferred that, for these cases, the NSGA-II is more effective at exploring the search space and finding correct solutions than Q-fit is. The second trend, which is perhaps more interesting, is shown in cases where the top ranked Q-fit solution is at the Pareto front, but its balance of energies is different to that from the correct Pareto solutions. Despite the fact that the total interaction energy of the top ranked Q-fit solution is comparable to the correct Pareto solutions, the difference in the balance of energies between the correct Pareto solutions and Q-fit verifies the importance of balancing the interaction energy types in order to obtain correct solutions when docking a molecule into a protein binding site.

This study has shown that the NSGA-II is capable of providing valuable information as to why a docking tool may fail in finding orientations of a molecule that are close to the crystal structure. It has particularly highlighted the importance for some complexes in obtaining a certain balance of interaction energies in order to obtain correct solutions. The results from Q-fit have shown that the algorithm tends to optimise electrostatic and hydrogen bond energies more so than vdw interactions, and therefore Q-fit fails to obtain correct solutions in cases where the correct orientation

of a molecule is more influenced by vdw interactions. Section 8.2.4 describes cases where Q-fit and the NSGA-II are both successful at obtaining correct solutions, and where the influential objective is the electrostatic and hydrogen bond energies. These cases draw attention to the fact that the NSGA-II is not limited to successfully docking vdw interactions-influenced complexes, and that its versatility includes docking complexes that are influenced by either of the objectives, and complexes which are not particularly influenced by any one objective, such as 1rne and 1bbp.

Comparing the abilities of the two programs in exploring the search space (from section 8.2.5), demonstrated some differences in accessing the search space. The Pareto solutions from the NSGA-II are, of course, not dominating each other, which explains why they are spread linearly in objective space, and why each solution has minimised objectives, albeit with different balances. Q-fit, because it uses a single objective to drive its search, does not produce the same spread of solutions as the NSGA-II. Its top ranked solutions are instead distributed evenly in objective space and, as demonstrated with 1fen, the group of solutions may also contain solutions with low rmsds that have good ranks. 1xie is also a good case showing the even spread of the Q-fit solutions, though the solution with the lowest rmsd has a low rank of 51, and as its position shows, it does not have low energies relative to the Pareto solutions. The top ranked Q-fit solutions for 1epb are not as evenly spread as the Q-fit solutions for 1fen and 1xie and, as Figure 8.17 shows, are more concentrated on optimised electrostatic and hydrogen bond energies than vdw interactions. This results in the algorithm missing the area in objective space containing correct solutions, indicated by the position of the correct Pareto solutions, which have lower vdw energies and higher electrostatic and hydrogen bond energies. These plots also show that generally both algorithms cover similar areas of search space (keeping in mind that during earlier generations of an NSGA-II run, the population of solutions, at some point, most likely covered the same areas of objective space as the top ranked Q-fit solutions). 1epb is an exception- most of its Q-fit solutions cover an area of higher vdw energies and low electrostatic and hydrogen bond energies, and do not extend to the area around the correct Pareto solutions. Similarly the left edge of the Pareto front obtained by 1epb has not advanced as far as the Q-fit solutions. The Q-fit solutions dominate the Pareto clusters represented by the orange and red triangles- it can therefore be expected that the Pareto front should stretch as far as these solutions.

This implies that, for 1epb, the left edge of the Pareto front has not completely converged to the true Pareto front. Nevertheless further along the Pareto front correct solutions exist, which is sufficient for providing information on the correct balance of energies.

Finally the NSGA-II was tested on a dataset consisting of proteins with hydrophobic binding sites. As was described in section 8.2.6, and as was expected, many of the correct Pareto solutions were more influenced by vdw, rather than electrostatic and hydrogen bond interactions. This study reflects the ability of the NSGA-II to provide useful information on the nature of a particular binding site, and on the influence of individual interaction energies in obtaining solutions with good rmsds.

8.3 Chapter Summary

This chapter has provided an in-depth study of the capabilities of the NSGA-II. Testing the algorithm on a large dataset has re-iterated the variation observed in the balance of energies between different complexes, and the importance of obtaining a correct balance of energies to find good solutions. This is especially apparent when the results are compared with Q-fit. In situations where Q-fit did not obtain solutions at high ranks with good rmsds, a comparison with the positions of the correct Pareto solutions in objective space revealed that this may be because Q-fit solutions do not have the correct balance of energies. This information is extremely useful in trying to understand the reasons for docking failures, and for highlighting the weakness of a scoring function. The study with the GSK-3 beta dataset has shown that the NSGA-II has the potential to be developed as a prospective docking tool.

As has been shown throughout the previous chapters, the NSGA-II generates a set of "correct" Pareto solutions, which are clustered into groups of solutions, all of which are within 1.0 Å from each other. Unlike other docking algorithms, these solutions are not ranked and, algorithmically-speaking, are all considered equal. It is therefore not possible to choose the correct Pareto cluster without any further information (such as crystallographic data). However the results from the study on the GSK-3 beta dataset

have shown that the NSGA-II may potentially be used for prospective docking. All correct Pareto solutions obtained by docking molecules into GSK-3 beta binding sites were situated on the right edge of the Pareto front, which suggests that, if a known binder to GSK-3 beta (for which there is no crystallographic information) is docked into a GSK-3 beta binding site using the NSGA-II, it may be possible to infer that the correct orientation of the molecule is on the right edge of the Pareto front.

9 The incorporation of a third objective: desolvation energy

In this chapter, the NSGA-II is modified to incorporate a third objective, desolvation energy, alongside the vdw and combined electrostatic and hydrogen bond energies. The purpose of this chapter is to explore the capabilities of the algorithm in docking ligands into protein active sites when three objectives, rather than two, are utilised. By incorporating a third objective, which is also an energy contribution to the process of binding, it may be possible to gain insight into the importance of desolvation energy in obtaining correct docked solutions. Also by observing desolvation energy values relative to the vdw and electrostatic and hydrogen bond energies of solutions, it may be possible to observe relationships between the different energy types, and whether these have any effect in obtaining correct docked solutions.

As was discussed in section 3.2.3.1, desolvation energy plays an important role in protein-ligand binding, its physical model influences ranking in virtual screening (Huang and Caflisch, 2004) as well as docking geometry (Ferrara *et al.*, 2004). Incorporating solvation energies accurately in scoring functions continues to be problematic (Leach *et al.*, 2006) mainly because of the difficulty in modelling aqueous systems with many degrees of freedom without sacrificing computational time.

Several attempts at addressing desolvation effects in docking have been employed. Morreale *et al.*, (2007) recently developed an implicit solvent model for computing the electrostatics binding free energy in protein-ligand docking where the system is immersed in a continuum that permeates all space surrounding the molecules. Specific comparisons of different implicit solvation models have also been performed (Ferrara *et al.*, 2004). The placement of water molecules during the docking search process has also been implemented in some docking algorithms. As mentioned in section 3.4.3.5, a version of GOLD allows water molecules to switch on and off and to rotate around their three axes during the search procedure. This process accounts for the loss of rigid-body entropy when water molecules are switched on, thus rewarding water displacement (Verdonk *et al.*, 2005). Similarly the placement of

water molecules during the search process of the incremental construction method FlexX has also been explored (Rarey *et al.*, 1999). The aim of the latter method is to find water molecules at the protein-ligand interface that may assist in the process of accurate ligand placement.

The method for calculating desolvation implemented here is based on the solvent accessible surface area, or SASA. The desolvation energy is approximated as the change in SASA of each atom or atom type upon binding, using appropriate parameters, the atomic desolvation parameters (ADP), that reflect the given atom type's propensity for aqueous environments. The ADP values that have been used here were taken from Jain *et al.*, (2005), which were used to predict protein-ligand binding affinities. These were in turn derived from a combination of the Cornell *et al.*, force field (1995), for proteins and nucleic acids, and the GAFF force field for small molecules (2004) (see Appendix).

The method that has been implemented for calculating SASA of a given protein/ligand is that which was originally defined by Lee and Richards (1971), and which is described in the following section.

9.1 The atomic vdw surface

The protein and ligand surfaces are defined by the vdw radii of the atoms on the surface of the molecules. For the ligand, the entire surface of the molecule is assessed using the vdw radii of the constituent atoms. For the purposes of protein-ligand docking, only the surface of the atoms on the protein binding site needs to be defined, particularly as this is kept rigid throughout the run of the algorithm. These surfaces are estimated in the following way.

Spheres with radii that are equal to the appropriate vdw radius of an atom of a given chemical type are placed so that the centre of the sphere is at the x,y and z coordinates of the atom as determined in the pdb file. The surfaces of the spheres are defined by evenly distributed points such that a given density of points can be calculated per

square Angstrom for each sphere. The number of points on a given atofm's surface can therefore be calculated by multiplying the surface area of a sphere by the density of points; the number of points on the equators is the product of the equator's circumference and the square-root of the point density. Half of the equatorial points (those on the side of one sphere) are used to position rings of points around the sphere. The rings' elevations from the centre of the sphere, the numbers o points on the rings and the rings' radii can be calculated using basic trigonometric rules (Figure 9.1).

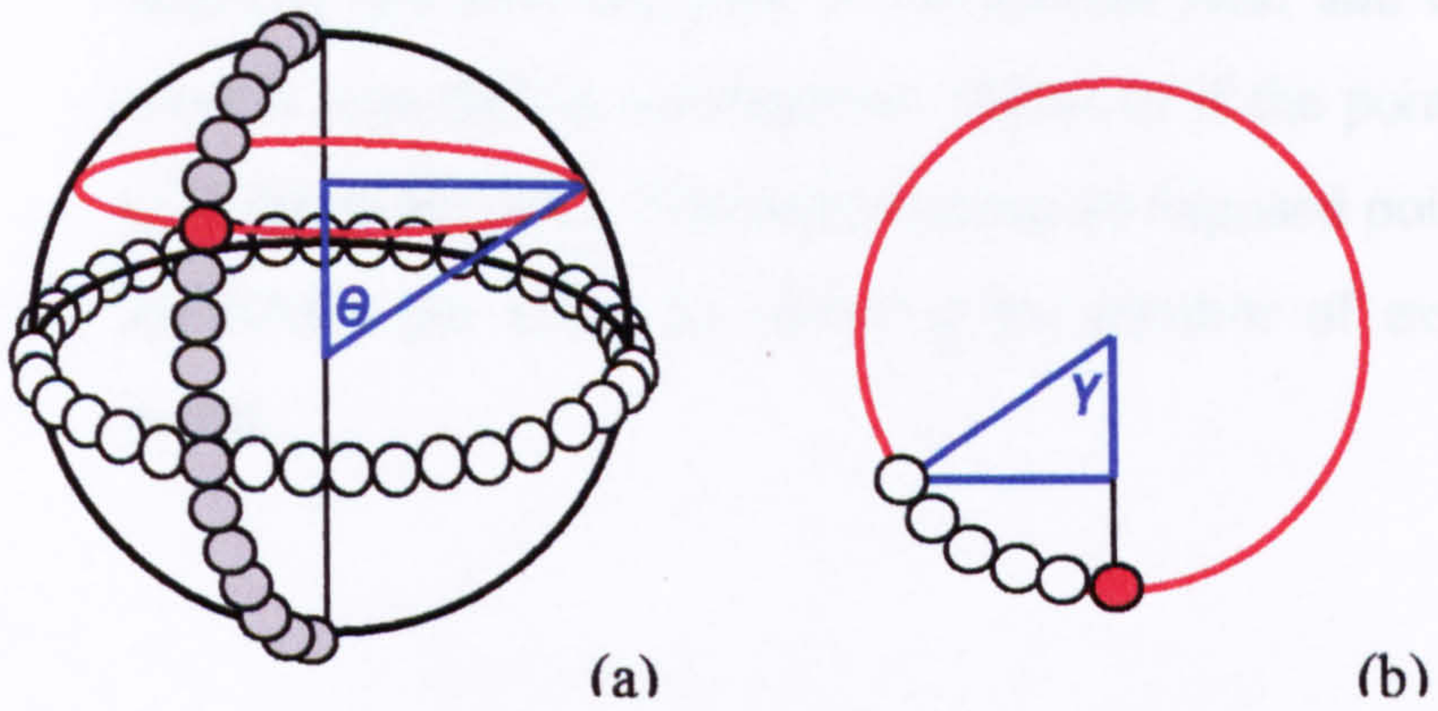


Figure 9-1 The method for positioning points on a sphere to determine an atom's vdw surface. (a) the number of points on the equator of the sphere (white circles), is calculated by multiplying the circumference of the equator and the square-root of the points density. Half of these points (grey circles), define rings around which the points are placed. The radius and elevation of the ring are the opposite and adjacent lengths of angle θ in the blue triangle (the sphere radius is the hypotenuse). The elevation is the z coordinates of points in the ring. For the n^{th} grey circle on the hemisphere, θ is the product of n and the fraction of 180° between adjacent grey circles on the hemisphere. (b) The number of points (white circles) on each ring (red circle, now viewed from above) is calculated as the product of the radius of the ring and square-root of the point density. The x and y coordinates of points on this circle are defined as the opposite and adjacent lengths the angle γ in the blue triangle (where the hypotenuse is the radius of the ring, as calculated previously).

In the manner described in Figure 9.1 above, an atom's vdw surface is determined. To determine which part of an atoms' vdw surface is exposed (in order to obtain the SAS), the molecule is first placed in a grid where the grid spacings represent the orthogonal limits of the SAS of the atom with the largest radii placed at the centre of the grid space. This allows for the rapid assessment of atoms that are likely to

influence the SAS of atoms within a given grid space. The atoms within the same or 26 surrounding grid spaces (6 orthogonally and 20 diagonally adjacent spaces) are the only atoms that can affect the SAS within the central grid space. This list can be reduced by selecting those atoms with overlapping SASs. The SAS of the protein binding site can be calculated by assessing the SAS of each atom in turn. The points on the SAS of the atom are defined using the points pre-calculated for a sphere, described in the previous section, of the same radii whose centre is superimposed on top of the atom's centre. Where the distance between point and neighbouring atom centre is less than the sum of the solvent radii and the vdw radius of the atom, the point is regarded as not exposed. However if the point is exposed, then it is regarded as lying on the SAS. The total number of exposed points per atom is used to calculate the SASA per atom, by dividing the number of exposed points by the density of points.

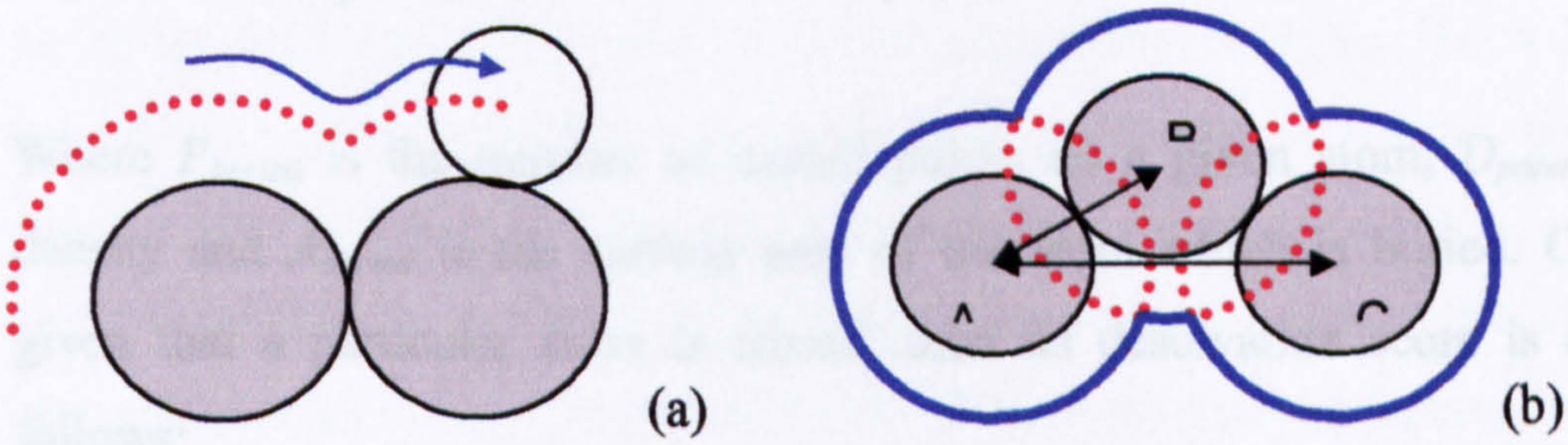


Figure 9-2 The Lee and Richards (1979) definition of SAS. (a) The SAS (solid line) around the protein/ligand (grey circles) is defined by rolling a sphere with the radius of water (1.4 Å- white circle) on the surface of the protein/ligand, and using the trace of the "path" of the centre of the sphere to represent the SAS. (b) The SAS (blue line) is defined by those points around atoms (grey spheres) which are not within the solvent radii of another atom (those which are, are shown in red). The neighbouring atom of A, atom B, is determined by calculating the distance between the centres of both atoms. If their centres are less than the combined vdw radii of the two atoms plus the diameter of the solvent probe, then the atoms are considered as neighbours and are used to calculate each other's SAS. Conversely the distance between the centres of A and C is over their combined radii and the probe diameter, and they are therefore not considered as neighbours.

9.2 Using SAS to calculate the buried surface area and the desovlation energy

To calculate the desolvation energy for a given ligand pose within the binding site, the buried surface area of both the protein and ligand, upon complex formation, must be established. This is calculated in the same way as determining the SAS described above, but rather than calculating distances between atom centres and the points on atoms of the same molecule, this is done between atoms on the different molecules, i.e. the protein and the ligand. A point on an atom is deemed buried if the distance between that point and a neighbouring atom centre from the other molecule is less than the sum of the solvent diameter and the neighbouring atom vdw radius. By counting the number of buried points on an atom, the buried surface area is calculated using the following equation:

$$A_{buried} = P_{buried} / D_{point} \qquad \text{Equation 9.1}$$

Where P_{buried} is the number of buried points on a given atom, D_{point} is the point density and A_{buried} is the surface area of the atom which is buried. Consequently, given that a particular atom is buried, then its desolvation score is calculated as follows:

$$E_{solv} = A_{buried} \times ADP_{atom} \qquad \text{Equation 9.2}$$

where ADP_{atom} is the atomic desolvation potential, or desolvation score, of the particular atom and E_{solv} is the given atom's contribution to desolvation energy.

Equations 9.1 and 9.2 are applied to all atoms of the ligand (since the position of the ligand is dynamic it is not possible to predetermine which of its atoms are in contact with the protein), and to the surface atoms of the protein binding site. The atoms on the surface of the binding site are determined at the start of a run, and these remain static throughout. The desolvation score of the ligand is calculated by first determining the distance between all points on its atoms to the protein surface atoms

to determine the molecule's buried surface area. This is followed by the application of equation 9.2 to calculate the desolvation score for each buried atom, and summing the desolvation scores for all buried atoms to give the desolvation energy of the ligand.

The process is reversed for the estimation of the protein's contribution to the desolvation energy. The distances between all the points on the protein surface atoms and the ligand atoms are calculated to find the binding site's buried surface area (using equation 10.1). Equation 10.2 is then used to calculate the desolvation score for each buried protein atom, all of which are summed to give the desolvation energy for the protein binding site. Finally the desolvation energies of both the protein and ligand are summed to give the desolvation energy of binding.

9.3 Incorporation of desolvation energy into NSGA-II

The NSGA-II was adapted to incorporate desolvation energy as the third objective, alongside vdw and electrostatic and hydrogen bond energies. Algorithmically the introduction of this objective requires the modification of the Parcto ranking function only; rather than using the two interaction energy objectives to determine the rank of every chromosome (as Figure 5.4 in Chapter 5 shows), the desolvation energy of each chromosome is also used, as the third objective, in the ranking process.

As was explained in Chapter 5 for the two-objective algorithm the combined parent and offspring population are assessed using all objectives, and this is followed by the Pareto ranking of the combined population. In the three-objective NSGA-II, the combined population's objectives are assessed by the two interaction energy objectives, as well as by the desolvation energy. Each chromosome's genes are therefore mapped into a ligand pose, and the desolvation energy between the protein and ligand pose is calculated. Specifically, and as described earlier, the buried surface area of the particular ligand pose is calculated relative to the atoms on the surface of the protein's binding site (Equation 9.1), followed by calculation of the desolvation energy using the constituent atoms' relevant desolvation parameters.

9.4 Preliminary Results

The three-objective NSGA-II was tested on Datasets 1 and 2, which were described in Chapter 6. Minimising the desolvation energy minimises the process of removing water from the interfaces of the molecules, thus contributing towards the binding of the molecules. The parameters used when testing Version 2 of the NSGA-II on Dataset 2 remained unchanged (Table 7.1). Out of the twenty complexes in Datasets 1 and 2, the three-objective NSGA-II obtained correct solutions for four complexes. These are from within Dataset 1, 1abe, 1ulb, 3ptb and 3tpi. Figures 9.3 to 9.6 are parallel coordinate plots, each plot showing the objective values of a chromosome in the final population. A single continuous line represents one chromosome, the objectives are plotted on the x-axis, and the values of the objectives for each chromosome on the y-axis, with the best value for each objective at zero on the y-axis. Non-dominated solutions are indicated by crossing lines which indicate a trade-off in the objectives. The solutions in the final population are clustered in terms of their rmsds. The raw values of all the objectives have been scaled, to allow for comparison, using the following equation:

$$scaledX_{obj} = (X_{obj}-X_{low})/(X_{high}-X_{low}) \qquad \text{Equation 9.3}$$

where X_{obj} is the raw objective value, X_{low} is the lowest value of that objective in the population, X_{high} is the highest value and $scaledX_{obj}$ is the scaled value of X_{obj}. This equation is applied to the objective values of all the chromosomes in the population.

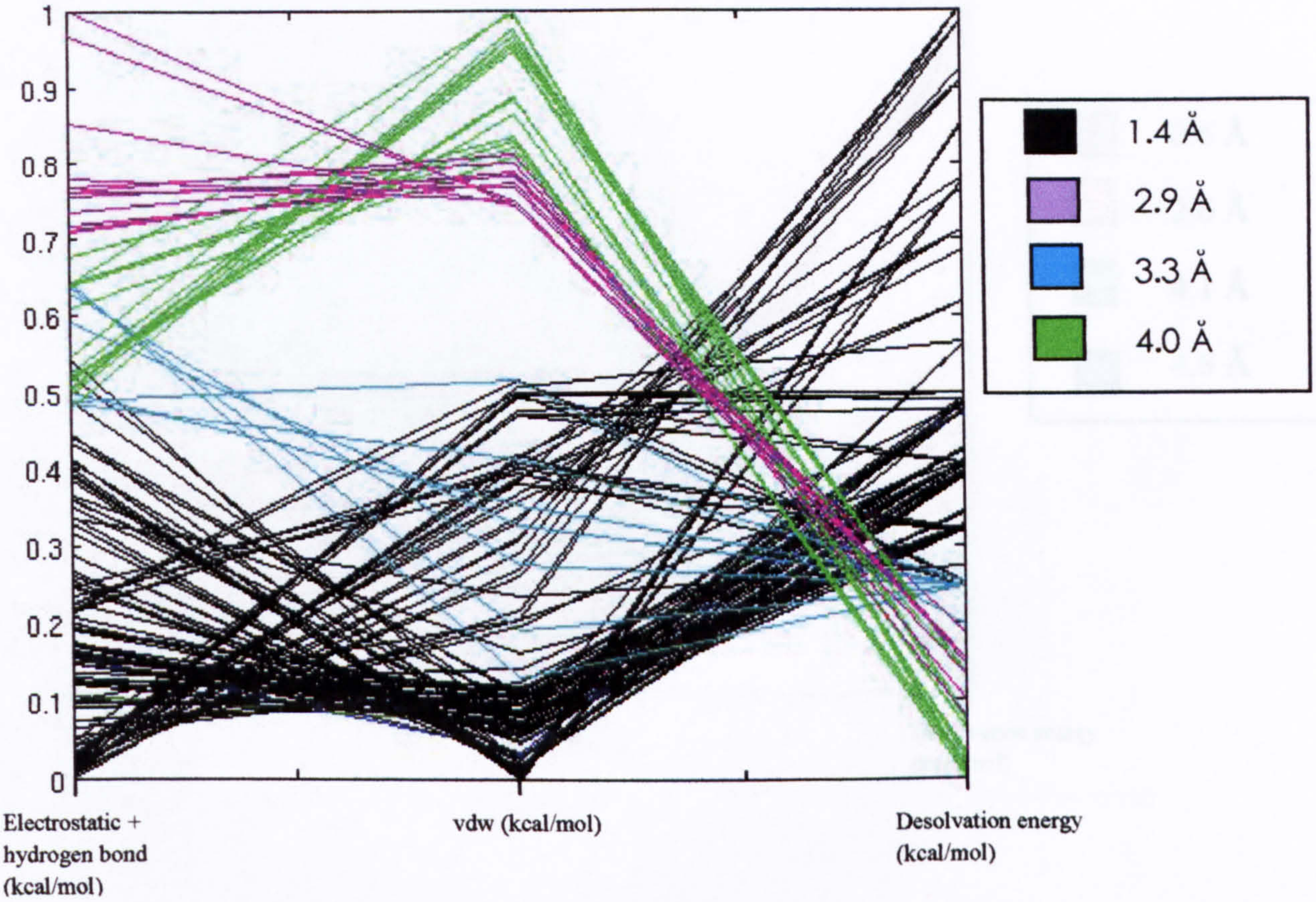


Figure 9-3 Parallel coordinate plots obtained when docking 1abe showing the objective values of the chromosomes in the final population.

Figure 9.3 shows that the three-objective NSGA-II succeeded in finding one cluster of correct solutions, with approximate rmsds of 1.4 Å, denoted by the black lines in the plot, for 1abe. The majority of the correct solutions, as the figure shows, have low vdw energies relative to the desolvation energies. Solutions with rmsds higher than 2.0 Å, especially the 2.9 Å and 4.0 Å clusters, show the inverse relationship between the two objectives; the vdw energies of these solutions are relatively higher than their desolvation energies. As the figure also shows, the electrostatic and hydrogen bond energies of the solutions with good rmsds are lower than solutions with worse rmsds. The vdw energies of the correct solutions are less spread out in objective space than the solution with higher rmsds.

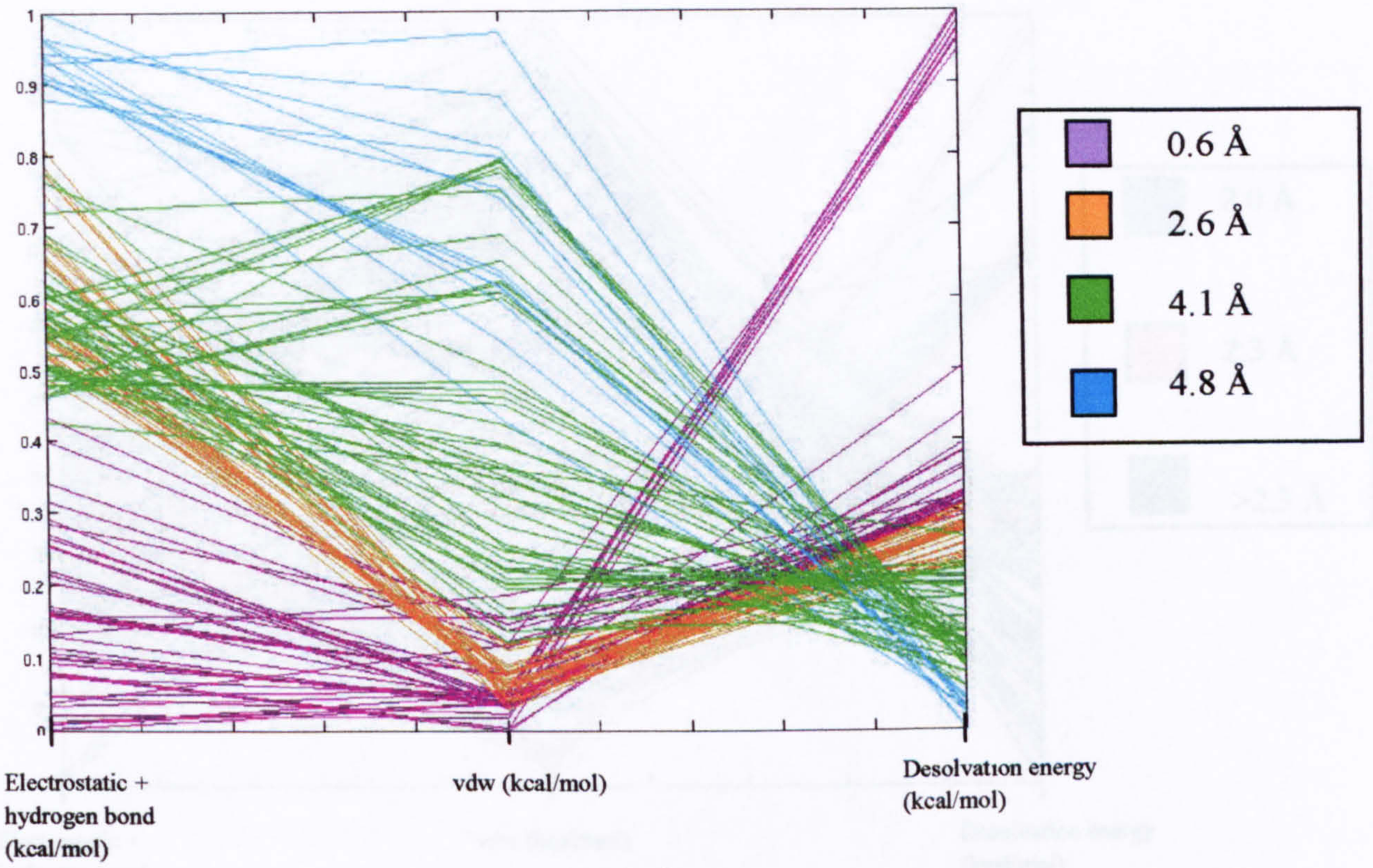


Figure 9-4 Parallel coordinate plots obtained when docking 1ulb showing the objective values of the chromosomes in the final population.

A similar scenario is observed with 1ulb. As figure 9.4 shows, solutions with low rmsds (in the 0.6 Å cluster shown in purple) have relatively high desolvation energies and low vdw energies; a similar situation is observed with the 2.6 Å cluster. The two clusters with the highest rmsds, 4.1 Å and 4.8 Å, have low desolvation energies and higher vdw energies. Unlike 1abe, where the desolvation energy of the solutions is largely uniformly distributed, the desolvation energies of 1ulb's solutions are spread over two discontinuous groups. The majority of solutions from the 0.6 Å clusters are clustered together on the higher end of the scale in terms of their desolvation energies, whereas the remaining solutions, covering a larger area on the scale, are discreetly grouped at the lower end of the scale. In fact, as the figure shows, several of these solutions appear to have converged to a singular desolvation energy value. As with 1abe, the electrostatic and hydrogen bond energies of the 0.6 Å solutions are, in general, lower than solutions in the clusters with higher rmsds. The vdw energies of solutions with high rmsds also cover a larger range than correct solutions.

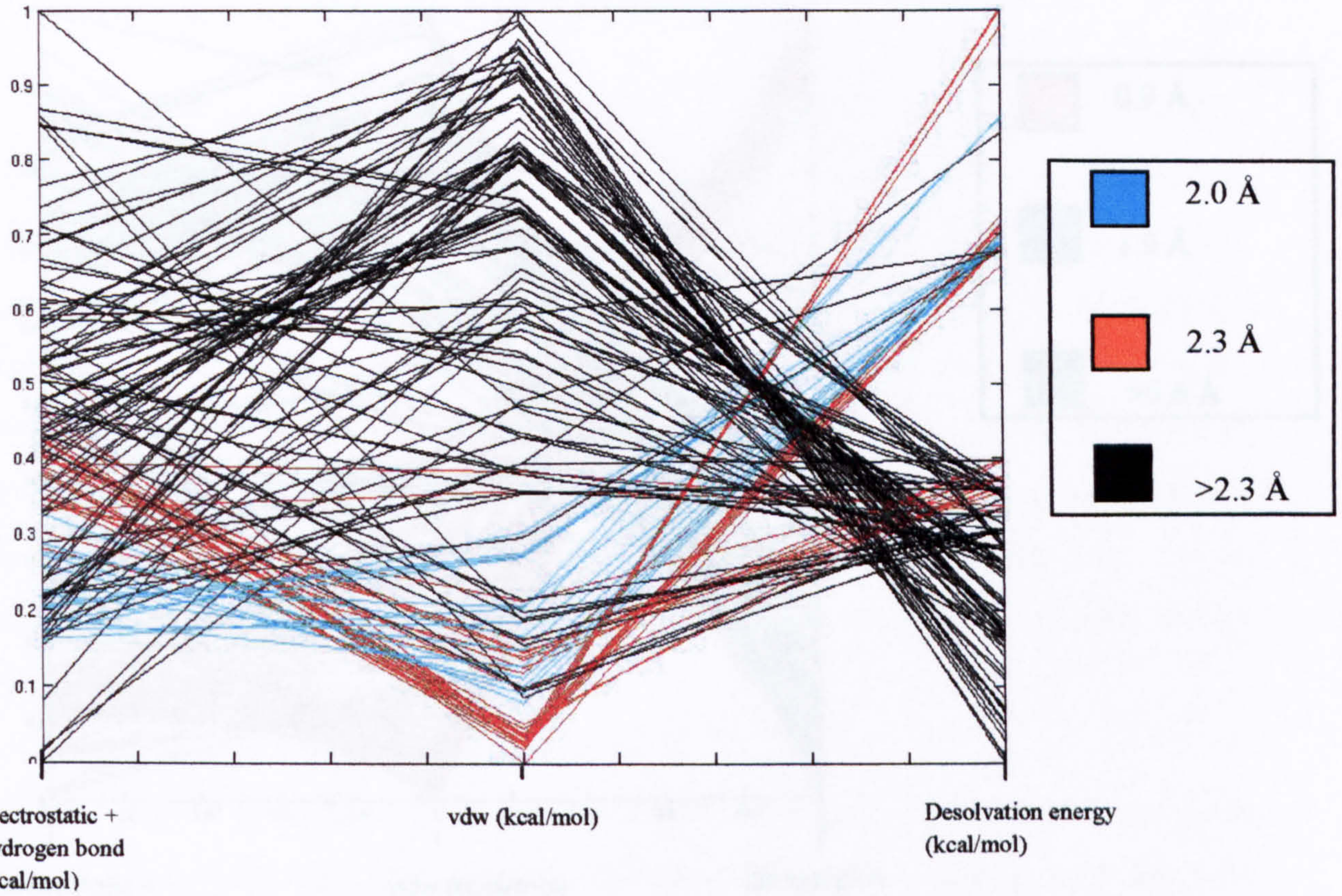


Figure 9-5 Parallel coordinate plots obtained when docking 3ptb showing the objective values of the chromosomes in the final population.

With 3ptb, a similar situation is observed; in general the solutions which have docked well have high relative desolvation energies and lower vdw energies, and vice versa for solutions with higher rmsds (Figure 9.5). In terms of electrostatic and hydrogen bond energies, solutions from the 2.0 Å cluster have lower values than solutions from the 2.3 Å cluster. Solutions with higher rmsds are distributed across the scale. As with the previous case, the values for the desolvation energies also fall into reasonably discreet groups and are not continuous, as shown with 1abe.

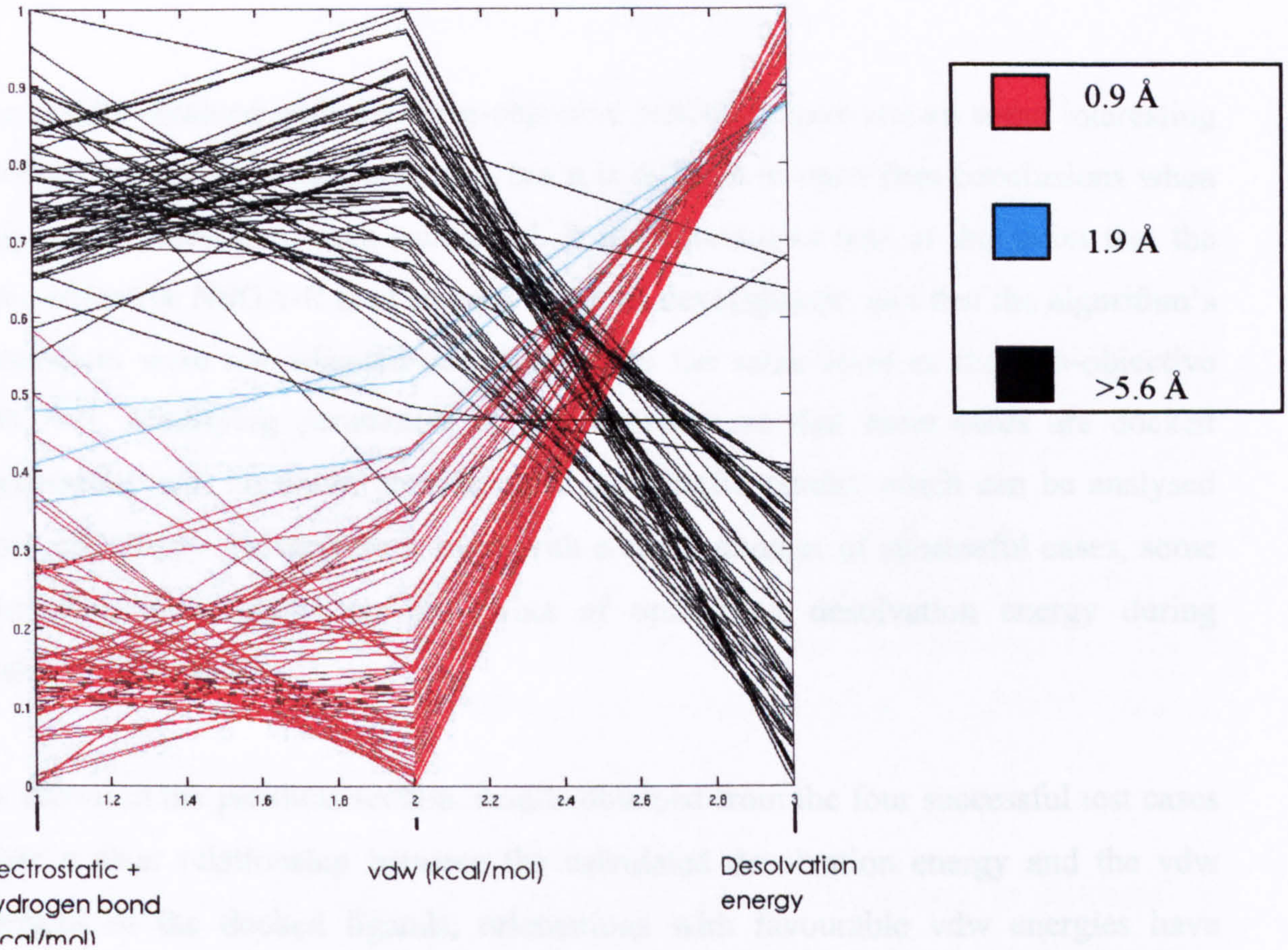


Figure 9-6 Parallel coordinate plots obtained when docking 3tpi showing the objective values of the chromosomes in the final population.

For 3tpi, the same observation can be made in terms of the relationship of the vdw to the desolvation energies. Solutions with good rmsds have high desolvation energies and low vdw energies, and vice versa for solutions with high rmsds. As Figure 9.6 shows, this effect is more apparent with this case than with the previous three cases. The desolvation energies of the solutions also fall into discreet groups, the correct solutions' desolvation energy values appear to have converged to a singular point, whereas the incorrect solutions' desolvation energy is more distributed across the scale. The electrostatic and hydrogen bond energies are, as shown with the other cases, lower for the correct solutions than for solutions with higher rmsds.

9.5 Discussion

The results obtained with the three-objective NSGA-II have shown some interesting relationships between the objectives, but it is difficult to draw firm conclusions when only a few cases have been successful. It is important to note at this point that the three-objective NSGA-II is at an early stage of development, and that the algorithm's parameters were not adjusted and tweaked to the same level as the two-objective NSGA-II. Modifying parameters of the algorithm so that more cases are docked successfully will no doubt provide more meaningful results which can be analysed more effectively. Nevertheless, even with a small number of successful cases, some interesting observations on the effect of optimising desolvation energy during docking can be made.

As shown in the previous section, results obtained from the four successful test cases show a clear relationship between the calculated desolvation energy and the vdw energies of the docked ligands; orientations with favourable vdw energies have unfavourable desolvation energies. The nature of the desolvation score that has been implemented here assigns more unfavourable energies to the loss of water from polar and charged groups than it does to the loss of water from aromatic and aliphatic surfaces. Orientations with favourable desolvation energies will be those that lose water from the surface of non-polar groups. These same orientations have unfavourable vdw interaction energies. This suggests that the desolvation potential optimises for orientations that bind to non-polar surface, and not for orientations that make polar-polar hydrogen and electrostatic interactions. The four successful cases have relatively polar interfaces, which would explain why all good orientations have high, or unfavourable, desolvation energies. The fact that these correct solutions have low vdw energies is possibly because correct orientations will inherently make optimal and favourable vdw interactions with the protein. The inverse of this argument could be applied to the incorrect solutions; the low desolvation scores of these orientations may be due to non-polar contacts between the protein and ligand; the high vdw energies of these solutions implies unfavourable vdw interactions, or steric clashes between the ligand and protein. These are speculative arguments and it is difficult to draw further conclusions between the relationships of the objectives

without further study of the systems and more successful cases, preferably cases consisting of non-polar protein binding sites.

Other issues to be aware of is that this method of calculating desolvation does not model a given system ideally, and makes certain assumptions to simplify the process of estimating the desolvation energies. One of these is that it treats the entire surfaces of molecules as lined with one layer of water. In practice this may not be the case. It has recently been shown by Barratt *et al.*, (2005), that in hydrophobic pockets the density of water is very much lower than expected, with water molecules only relatively static around polar groups. A second assumption made by the implemented model of desolvation is that water molecules form spanning networks of interactions on protein surfaces (Oleinikova *et al.*, 2005), implying that the desolvation energy is based on several layers of interacting water molecules rather than just the first to contact the protein surface.

The purpose of this study was to show that the NSGA-II can be extended to incorporate a third objective that is important to the process of docking, namely desolvation energy. The application of an optimised version of the algorithm to a larger dataset should provide information on whether including desolvation energy improves docking results. This should also enable the study of the effect of desolvation energy relative to electrostatic and hydrogen bond, and vdw energies in obtaining correct solutions. A comparison to Q-fit, as performed in previous chapters with the two-objective NSGA-II, may also provide an insight on whether the inclusion of the desolvation term has an effect on improving docking performance.

9.6 Chapter Summary

In this chapter the NSGA-II was extended to include a desolvation energy term as a third term, alongside electrostatic and hydrogen bond, and vdw interactions, thus fulfilling one of the aims outlined in section 3.5. Preliminary results obtained when the three-objective NSGA-II was tested on cases in Datasets 1 and 2 showed that the algorithm docked four cases successfully. These results showed that the desolvation

energies of the correct solutions were high, and appear to be inversely correlated with vdw energies. However it is recognised that these are preliminary results and further optimisation and testing of the algorithm is needed to determine the ultimate benefit of the three-objective NSGA-II.

10 Discussion and Conclusions

Limitations of scoring functions are repeatedly attributed to the hindrance of developing robust and reliable docking algorithms, which is why scoring function development remains an active area of research (Jain, 2006). The importance of this area motivated the aims of this work, which is the application of a multiobjective optimisation method to a particular scoring function. This work has focused on understanding the roles of individual energy contributions in finding good poses within a protein binding site, which can effectively be performed using multiobjective optimisation. To our knowledge, this type of study has not been performed previously, and has thus provided a valuable insight to our knowledge of scoring functions and their roles in docking algorithms.

10.1 Summary of Results and Discussion

The approach taken to develop a multiobjective optimisation algorithm for protein-ligand docking was to start by developing a single-objective, standard genetic algorithm (SGA), which uses a single objective, or fitness function, to score chromosomes, and to adapt that into a multiobjective genetic algorithm. This is done by introducing functions specific to multiobjective optimisation, including adapting the scoring procedure, so that chromosomes are scored by multiple, rather than a single objective. The production of a multiobjective algorithm for protein ligand docking fulfils one of the major aims of this work.

The experimental work begins in Chapter 4, within which the development of the SGA is described. This is a "classic" genetic algorithm, implementing the popular genetic operators, roulette wheel selection, single-point crossover and mutation, to perform rigid-body docking of ligands into protein binding sites. The algorithm used the GRID scoring function to assess the chromosomes that represent the different poses of the ligand. To test the algorithm's capabilities, the SGA was run on a dataset consisting of ten protein-ligand complexes, known as Dataset 1. Dataset 1 was also

used to run Q-fit, a docking tool that uses GRID as its scoring function, thus forming a good basis for comparison. The SGA docked four out of the ten complexes successfully, i.e. the top-ranked solutions of these four complexes had rmsds of less than 2.0 Å. Q-fit was, overall, more successful at obtaining solutions with good rmsds and with lower interaction energies; one reason for this may be that Q-fit performs a local minimisation of the solutions listed in its output of ranked solutions, and the SGA does not. Also the performance of the SGA would likely improve if the algorithm's parameters were further experimented with and tweaked. However, the current performance of the SGA was deemed satisfactory for its modification into a multiobjective genetic algorithm. The view behind this was that it is more reasonable to improve the algorithm's performance by modifying the multiobjective optimisation algorithm, whose development is, after all, the purpose of this work, rather than the SGA. This chapter is therefore followed by a methods chapter, describing the conversion of the algorithm from a single-objective into a multiobjective algorithm (Chapter 5).

From the many different modes of multiobjective optimisation algorithms (some of which are described in section 2.5), it was decided to implement a version of the highly elitist NSGA-II (Deb *et al.*, 2000) for performing protein-ligand docking. As mentioned earlier, the different energy components of the scoring function were selected for multiobjective optimisation. These are the vdw interactions and the combined electrostatic and hydrogen bond interactions of the GRID scoring function. Some of the major changes implemented to enable the algorithm to perform multiobjective optimisation (apart from scoring chromosomes with two objectives) include Pareto ranking and niching. The NSGA-II was tested on two datasets, the results of which are described in Chapter 6.

The results described in Chapter 6 were produced when the NSGA-II was tested on two datasets, termed Datasets 1 and 2. These datasets were selected primarily because they were used to validate Q-fit (Jackson, 2002), and the ligands forming the complexes are deemed suitable for rigid-body docking. Also, both datasets contain complexes of varying difficulties for docking algorithms. Dataset 1 contains relatively straightforward complexes for docking whereas Dataset 2 is considered a more problematic dataset, consisting of more challenging complexes.

Of the ten complexes in Dataset 1, the NSGA-II obtained correct solutions within the Pareto sets for eight out of the ten complexes. The most prominent observation made from this data is that the successful docking of different complexes is influenced by different objectives. This can be inferred by observing the position of the correct solutions in objective space and the contributions that the different objectives are making to the interaction energies of the correct solutions. Some of the complexes within Dataset 1 are influenced relatively equally by both objectives, and some are found to be more influenced by electrostatic and hydrogen bond energies. These results were compared to results obtained when Q-fit docks the same complexes. As was discussed in section 3.5, a multiobjective optimisation approach to docking can be used to understand why a single objective optimisation approach may fail at finding correct solutions. Given that a single objective optimisation method has failed because it did not optimise individual energy terms adequately, then a multiobjective approach will confirm this, and also reveal which of the energy terms need to be optimised to obtain correct solutions. The top-ranked Q-fit solutions were, for the majority of the complexes, among the correct Pareto solutions, indicating that both algorithms have found an optimal balance of the two objectives. The two complexes which the NSGA-II did not dock successfully, 2phh and 4dfr, Q-fit did succeed in docking, and by observing the position of these solutions in objective space, it is clear that Q-fit was successful at docking these because it was capable of minimising their interaction energies further than the NSGA-II. This indicates that the NSGA-II did not minimise these solutions effectively.

With Dataset 2, seven out of ten complexes contained correct solutions among the Pareto set. The influence of the objectives in this dataset varied. Two of the complexes were influenced equally by both objectives, one complex was influenced by electrostatic and hydrogen bond interactions and, unlike the Dataset 1 results, three complexes are also influenced by vdw interactions. An interesting observation with this dataset is that the shape of some of the Pareto fronts is not as "smooth" as those observed with Dataset 1, indicating a more "rugged" energy landscape, which in turn implies that it is easier for the algorithm's search component to get stuck at a local minimum.

Considering Q-fit's results, the top-ranked Q-fit solutions were among the correct Q-fit solutions for several of the complexes. An interesting case was that of 1baf, for which the NSGA-II obtained a good solution, but Q-fit did not. This was clearly a case of the NSGA-II obtaining a correct balance of energies for a subset of its Pareto solutions that resulted in solutions with good rmsds. It could therefore be argued that Q-fit was unable to find a low rmsd solution at a high rank because the correct balance of interaction energies was not achieved. This case demonstrates the potential usefulness of a multiobjective approach relative to a single objective one, and confirms the ability of the NSGA-II to understand *why* a single objective optimisation approach may have failed . The three complexes that the NSGA-II did not succeed in docking were also not successfully docked by Q-fit. Therefore it was inferred that these complexes are problematic for docking, with global minima that are difficult to access. The overall inference of the results from the two datasets is that to obtain correct poses, optimising objectives so that a correct balance is obtained is important for certain complexes. Secondly, these results show that that the different energy terms, or objectives, have varying influences between different complexes. It is not possible to determine, with single objective optimisation, which of the energy terms is having the strongest influence in optimising the search, whereas a multiobjective optimisation approach to retrospective docking can determine the influence of different objectives by observing the position of correct solutions in objective space.

In Chapter 7, the NSGA-II is modified from the version described in Chapter 5; the main reason for this was to determine the effect of certain algorithmic enhancements on the performance of the algorithm. The three major modifications made to the algorithm were; the application of controlled elitism; reduction of the E_{max} parameter to smooth the energy landscape and the implementation of a local search by using simplex minimisation with a Lamarckian element. Reducing the value of E_{max} softens the energy landscape, and allows for the existence of chromosomes that may otherwise be clashing with the protein, but that are nevertheless close to the crystal structure. Finally performing the simplex minimisation procedure allows chromosomes with orientations relatively close to the crystal structure and with low Pareto ranks to reach a local minimum, thus increasing their Pareto ranks and increasing their prospects of surviving in the population. The Lamarckian element to

this method ensures that the newly minimised chromosomes are passed on to the next generation of the population.

The new version of the algorithm was tested on Datasets 1 and 2, and the overall consensus was that the modified algorithm performs better than the original NSGA-II. Comparing the Pareto fronts from both versions of the algorithm shows that the modified version advances the Pareto front further than the original version. Also, the modified version successfully docked all ten complexes in Dataset 1, an improvement from Version 1 of the algorithm which only docked eight complexes. With Dataset 2, Version 2 succeeded in docking a complex which Version 1 could not, and also failed to a dock a different complex that was successfully docked by Version 1. Therefore the overall success rate for Dataset 2 from both versions of the algorithm remained unchanged though the Pareto fronts from several of the dataset's complexes had advanced further with Version 2.

The results obtained with Version 2 indicate that the modifications implemented have been beneficial to the performance of the algorithm. Also, they add a more exploratory note to the algorithm, giving the flexibility to modify the parameters controlling the modifications as deemed necessary to a particular test case. To more thoroughly understand the capabilities of Version 2, the algorithm was tested on several, more extensive datasets, the results of which are described in Chapter 8.

In Chapter 8 the NSGA-II (Version 2) was tested on two datasets. One of these consists of complexes of one protein (glycogen synthase kinase-3 beta, or GSK-3 beta) co-crystallised with several ligands, and the second is a large dataset (FlexX) that is routinely used in different docking experiments. Testing the algorithm on these different types of datasets fulfils two of the aims outlinsed in section 3.5. The purpose of the GSK-3 beta study was to see whether any trends could be observed when the NSGA-II is used to dock different ligands into the same protein. The Pareto fronts generated by the GSK-3 beta study revealed two trends in the interactions made between the protein and its ligands. The vdw interactions were seen as the dominating interactions for three of the complexes. For three other complexes the very narrow range of electrostatic and hydrogen bond energies of the correct solutions implied that these ligands need to make specific electrostatic and hydrogen bond interactions in

order to find correct solutions. Interestingly Q-fit was unable to successfully dock the vdw-influenced complexes, whilst succeeding in docking the electrostatic and hydrogen bond-specific ligands. This implies that Q-fit is capable, for this set of complexes, of optimising specific electrostatic/hydrogen bond interactions, which is beneficial if a correct pose is influenced by these interactions. However if vdw interactions are the more influential of the objectives, then Q-fit can fail in finding correct solutions. Again this demonstrates the ability of the NSGA-II in highlighting the importance of individual energy terms and in explaining why a single objective optimisation algorithm may have failed in obtaining correct solutions, thus fulfilling one of the aims introduced in section 3.5. Another interesting observation made through this study is that the complexes' correct clusters obtained by the NSGA-II are all on the right edge of the Pareto front. Though more validation studies are needed to confirm this, this information could be used to potentially perform prospective docking using the NSGA-II.

The purpose of testing the NSGA-II on a large dataset like FlexX was to compare the algorithm's performance to Q-fit at a larger scale, and to observe whether the multiobjective approach could provide extra information on Q-fit's capabilities. Though overall Q-fit docked more of the FlexX test cases successfully, there were 17 cases that the NSGA-II successfully docked which Q-fit could not. Examining these further revealed that Q-fit may have failed in these cases for one of two reasons, (1) it did not minimise the total interaction energy to the level of the NSGA-II, (2) it did not achieve a correct balance of energies to find correct solutions. From the first point, it can be inferred that, for these particular complexes, the NSGA-II was more successful at finding solutions at the global minimum than Q-fit, whose search may have stopped at a local minimum. The more interesting cases are those which failed because of the second point, once again highlighting the importance, for any algorithm, of finding the correct balance of energies in order to succeed at finding good solutions. Overall, the FlexX study has shown that multiobjective optimisation can provide insight into why a docking tool may fail in finding correct solutions for certain complexes, exemplifying possible scoring function weaknesses.

Finally Chapter 9 describes achievement of the final aim of this thesis (section 3.5), which is the implementation of a third objective, desolvation energy, within the

NSGA-II. Although the results produced are preliminary only, they have demonstrated the algorithm's capabilities to use three objectives, rather than two, in optimisation. The preliminary results revealed some interesting relationships between the objectives, though further analyses and improvements to the algorithm are necessary before any firm conclusions can be drawn.

10.2 Conclusions and Future Directions

Scoring functions continue to be a problematic aspect of protein-ligand docking. This work, applying a multiobjective approach to the scoring function of a docking algorithm, has highlighted the importance of correctly weighting the individual energy terms constituting a scoring function. Ideally scoring functions should have a level of transferability between complexes, and although this can be achieved for the most part, the results from this thesis have shown that certain failures of a docking algorithm could be attributed to the varied influences of different interaction energies among complexes. Multiobjective optimisation has provided a new "dimension" to understanding the optimisation of scoring functions during a search. As shown with the results from Q-fit, the NSGA-II can be used to understand a specific docking algorithm more thoroughly, and to be aware of any weaknesses. Although the algorithm cannot perform prospective docking, a potential avenue to explore, as shown with the GSK-3 beta data, is to learn where correct solutions lie on the Pareto fronts obtained from docking co-crystallised ligands into the same protein binding site, and to use that information to predict orientations of ligands for which no x-ray crystal information is available.

In its current form, the NSGA-II is a tool for gaining understanding of the relative importance of the different factors that contribute to a scoring function, and consequently a docking algorithm. As this discussion has shown, the NSGA-II has provided useful insight into the performance of Q-fit, its implementation of the GRID scoring function, and has shown why the algorithm fails to dock certain test cases successfully. This approach could potentially be extended to other docking programs; a multiobjective procedure may be used to understand the effect of the different

components of other scoring functions, assess their influence in finding correct solutions and may also reveal their ability to balance the different components. To do this, the NSGA-II must be adapted to incorporate the same scoring function, where the objectives are the relevant scoring function components. Results obtained from the algorithm can then be compared with results from the docking tool under study, in the same manner used here to assess Q-fit.

Though the current NSGA-II generally had a good docking success rate, there is the potential to increase its level of performance. The NSGA-II currently performs rigid-body docking, and though for the purposes of comparison of its performance with Q-fit this was adequate (Q-fit is also a rigid-body docking tool), introducing ligand flexibility may be desirable. This could potentially improve the algorithm's performance since more optimal interactions could be made if rotatable bonds have the freedom to move. Also this would allow the algorithm to be compared to the more popular docking tools, which all implement flexible ligand docking. A further improvement to the algorithm is to optimise the three-objective version of the algorithm so that the influence of desolvation energy in protein-ligand docking can be understood. Finally the algorithm could be used to help in the parameterisation, or tailoring of a scoring function, so that the weights for individual energy components can be more reliably assigned.

Appendix

121p	1eta	1nis	2cpp	8gch
1aaq	1etr	1nsc	2ctc	9hvp
1abe	1fen	1pbd	2dbl	
1abf	1fkg	1pha	2er6	
1acj	1fki	1phd	2gbp	
1ack	1frp	1phf	2lgs	
1acm	1ghb	1phg	2mcp	
1aec	1glp	1poc	2mth	
1aco	1glq	1ppc	2phh	
1aha	1hdc	1pph	2pk4	
1ake	1hef	1ppi	2r04	
1apt	1hfc	1ppk	2r07	
1ase	1hgg	1ppl	2sim	
1atl	1hgh	1ppm	2tmn	
1avd	1hgi	1pso	2yhx	
1azm	1hgj	1rbp	3aah	
1baf	1hri	1rne	3cla	
1bbp	1hsl	1rnt	3cpa	
1blh	1hti	1rob	3gch	
1bma	1hvr	1slt	3hvt	
1byb	1hyt	1snc	3ptb	
1cbs	1icn	1srj	3tpi	
1cbx	1ida	1stp	4cts	
1cde	1igj	1tdb	4est	
1cdg	1imb	1thy	4fbp	
1cil	1ivb	1tka	4dfr	
1com	1ivc	1tlp	4fxn	
1coy	1ivd	1tmn	4hmg	
1cps	1ive	1tng	4hvp	
1ctr	1ivf	1tnh	4phv	
1dbb	1ldm	1tni	4tim	
1dbj	1lah	1tnj	4tln	
1dbk	1lic	1tnk	4ts1	
1dbm	1lpm	1tnl	5abp	
1did	1lcp	1tph	5cpp	
1die	1lmo	1tpp	5cts	
1dr1	1lna	1trk	5p2p	
1dwb	1lst	1tyl	5tim	
1dwc	1mbi	1ukz	5tmn	
1dwd	1mcr	1ulb	6abp	
1eap	1mdr	1wap	6cpa	
1eed	1mld	1xid	6rnt	
1ela	1mmq	1xie	6rsa	
1elb	1mnc	2ak3	6tim	
1elc	1mrg	2ada	6tmn	
1eld	1mrk	2cgr	7cpa	
1ele	1mup	2cht	7tim	
1epb	1nco	2cmd	8atc	

List of PDB codes of FlexX dataset

Description	Parameters
sp^2 carbonyl	-0.1209
sp carbon	-0.2522
sp^2 carbon aliphatic	0.0283
sp^2 carbon aromatic	-0.0141
sp^3 carbon	-0.1276
Halogens (Fl, Cl, Br, I)	-0.0081
Hydrogen bonded to aliphatic carbon	-0.0005
Hydrogen bonded to aromatic carbon	0.0040
Hydrogen bonded to nitrogen	-0.0051
Hydroxyl group	-0.0013
Hydrogen bonded to sulfur	-0.0595
sp^2 nitrogen in amide groups	0.0232
sp^2 nitrogen in aliphatic systems	0.0311
sp^2 nitrogen in aromatic systems	0.0111
sp nitrogen	-0.0037
sp^3 nitrogen	0.0478
Amine nitrogen connected to one or more aromatic rings	-0.0077
Oxygen with one connected atom	0.0074
Oxygen in hydroxyl group	0.0094
Ether and ester oxygen	0.0147
Phosphate	-0.7097
Sulfur	-0.0109

Desolvation parameters used in desolvation energy calculations in Chapter 9 (Jain and Jayaram, 2005)

Bibliography

Abagyan, R. & Totrov, M. (2001). High-throughput docking for lead generation. *Curr Opin Chem Biol* **5**, 375-82.

Acharya, K. R. & Lloyd, M. D. (2005). The advantages and limitations of protein crystal structures. *Trends Pharmacol Sci* **26**, 10-4.

Ahlstrom, M. M., Ridderstrom, M., Luthman, K. & Zamora, I. (2005). Virtual screening and scaffold hopping based on GRID molecular interaction fields. *J Chem Inf Model* **45**, 1313-23.

Ajay & Murcko, M. A. (1995). Computational methods to predict binding free energy in ligand-receptor complexes. *J Med Chem* **38**, 4953-67.

Andersson, C. D., Thysell, E., Lindstrom, A., Bylesjo, M., Raubacher, F. & Linusson, A. (2007). A multivariate approach to investigate docking parameters' effects on docking performance. *J Chem Inf Model* **47**, 1673-87.

Barratt, E., Bingham, R. J., Warner, D. J., Laughton, C. A., Phillips, S. E. & Homans, S. W. (2005). Van der Waals interactions dominate ligand-protein association in a protein binding site occluded from solvent water. *J Am Chem Soc* **127**, 11827-34.

Bash, P. A., Singh, U. C., Brown, F. K., Langridge, R. & Kollman, P. A. (1987). Calculation of the relative change in binding free energy of a protein-inhibitor complex. *Science* **235**, 574-6.

Baxter, C. A., Murray, C. W., Clark, D. E., Westhead, D. R. & Eldridge, M. D. (1998). Flexible docking using Tabu search and an empirical estimate of binding affinity. *Proteins* **33**, 367-82.

Berman, H. M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T. N., Weissig, H., Shindyalov, I. N. & Bourne, P. E. (2000). The Protein Data Bank. *Nucleic Acids Res* **28**, 235-42.

Bohm, H. J. (1994). The development of a simple empirical scoring function to estimate the binding constant for a protein-ligand complex of known three-dimensional structure. *J Comput Aided Mol Des* **8**, 243-56.

Brenk, R., Vetter, S. W., Boyce, S. E., Goodin, D. B. & Shoichet, B. K. (2006). Probing molecular docking in a charged model binding site. *J Mol Biol* **357**, 1449-70.

Brooijmans, N. & Kuntz, I. D. (2003). Molecular recognition and docking algorithms. *Annu Rev Biophys Biomol Struct* **32**, 335-73.

Brooks, B. R., Bruccoleri, R.E., Olafson, B.D., States, D.J., Swaminathan, S., Karplus, M. (1983). CHARMM: A program for macromolecular energy, minmimization, and dynamics calculations. *J. Comp. Chem.* **4**, 187-217.

Brown, N., McKay, B., Gilardoni, F. & Gasteiger, J. (2004). A graph-based genetic algorithm and its application to the multiobjective evolution of median molecules. *J Chem Inf Comput Sci* **44**, 1079-87.

Campbell, S. J., Gold, N. D., Jackson, R. M. & Westhead, D. R. (2003). Ligand binding: functional site location, similarity and docking. *Curr Opin Struct Biol* **13**, 389-95.

Catana, C. & Stouten, P. F. (2007). Novel, customizable scoring functions, parameterized using N-PLS, for structure-based drug discovery. *J Chem Inf Model* **47**, 85-91.

Cecchini, M., Kolb, P., Majeux, N. & Caflisch, A. (2004). Automated docking of highly flexible ligands by genetic algorithms: a critical assessment. *J Comput Chem* **25**, 412-22.

Charifson, P. S., Corkery, J. J., Murcko, M. A. & Walters, W. P. (1999). Consensus scoring: A method for obtaining improved hit rates from docking databases of three-dimensional structures into proteins. *J Med Chem* **42**, 5100-9.

Clark, R. D., Strizhev, A., Leonard, J. M., Blake, J. F. & Matthew, J. B. (2002). Consensus scoring for ligand/protein interactions. *J Mol Graph Model* **20**, 281-95.

Claussen, H., Buning, C., Rarey, M. & Lengauer, T. (2001). FlexE: efficient molecular docking considering protein structure variations. *J Mol Biol* **308**, 377-95.

Cole, J. C., Murray, C. W., Nissink, J. W., Taylor, R. D. & Taylor, R. (2005). Comparing protein-ligand docking programs is difficult. *Proteins* **60**, 325-32.

Cornell, W. D., Cieplak, P., Bayly, C.I., Gould, I. R., Merz, Jr.,K. M., Ferguson, D.M., Spellmeyer, D.C.,Fox, T., Caldwell, J.W and Kollman, P. A. (1995). A Second Generation Force Field for the Simulation of Proteins, Nucleic Acids, and Organic Molecules. *J Am Chem Soc* **117**, 5179-5197.

Cottrell, S. J., Gillet, V. J. & Taylor, R. (2006). Incorporating partial matches within multi-objective pharmacophore identification. *J Comput Aided Mol Des* **20**, 735-49.

Cottrell, S. J., Gillet, V. J., Taylor, R. & Wilton, D. J. (2004). Generation of multiple pharmacophore hypotheses using multiobjective optimisation techniques. *J Comput Aided Mol Des* **18**, 665-82.

Cram, D. J. (1988). The design of molecular hosts, guests, and their complexes. *Science* **240**, 760-7.

Davis, A. M., Teague, S. J. & Kleywegt, G. J. (2003). Application and limitations of X-ray crystallographic data in structure-based ligand and drug design. *Angew Chem Int Ed Engl* **42**, 2718-36.

Deb, K., Agrawal, S., Pratap, A. & Meyarivan, T. (2000). A fast elitist non-dominated sorting genetic algorithm for multi-objective optimisation: NSGA-II. *Proceedings of the Parallel Problem Solving from Nature VI*, 849-858.

Deb, K. & Goel, T. (2001). Controlled Elitist Non-dominated Sorting Genetic Algorithms for Better Convergence. *EMO* **2001**, 385-399.

DesJarlais, R. L., Sheridan, R. P., Dixon, J. S., Kuntz, I. D. & Venkataraghavan, R. (1986). Docking flexible ligands to macromolecular receptors by molecular shape. *J Med Chem* **29**, 2149-53.

Dill, K. A., Truskett, T. M., Vlachy, V. & Hribar-Lee, B. (2005). Modeling water, the hydrophobic effect, and ion solvation. *Annu Rev Biophys Biomol Struct* **34**, 173-99.

Doble, B. W. & Woodgett, J. R. (2003). GSK-3: tricks of the trade for a multi-tasking kinase. *J Cell Sci* **116**, 1175-86.

Eldridge, M. D., Murray, C. W., Auton, T. R., Paolini, G. V. & Mee, R. P. (1997). Empirical scoring functions: I. The development of a fast empirical scoring function to estimate the binding affinity of ligands in receptor complexes. *J Comput Aided Mol Des* **11**, 425-45.

Ewing, T. J., Makino, S., Skillman, A. G. & Kuntz, I. D. (2001). DOCK 4.0: search strategies for automated molecular docking of flexible molecule databases. *J Comput Aided Mol Des* **15**, 411-28.

Ferrara, P., Gohlke, H., Price, D. J., Klebe, G. & Brooks, C. L., 3rd. (2004). Assessing scoring functions for protein-ligand interactions. *J Med Chem* **47**, 3032-47.

Flower, D. R. (1999). Modelling G-protein-coupled receptors for drug design. *Biochim Biophys Acta* **1422**, 207-34.

Fonseca, C. M. & Fleming, P. J. (1998a). Multiobjective optimization and multiple constraint handling with evolutionary algorithms-Part I: A unified formulation. *IEEE T Syst Man Cy A* **28**, 38-47.

Fonseca, C. M. & Fleming, P. J. (1998b). Multiobjective optimization and multiple constraint handling with evolutionary algorithms-Part II: Application example. *IEEE T Syst Man Cy A* **28**, 38-47.

Foster, J. A. (2001). Evolutionary computation. *Nat Rev Genet* **2**, 428-36.

Friesner, R. A., Banks, J. L., Murphy, R. B., Halgren, T. A., Klicic, J. J., Mainz, D. T., Repasky, M. P., Knoll, E. H., Shelley, M., Perry, J. K., Shaw, D. E., Francis, P. & Shenkin, P. S. (2004). Glide: a new approach for rapid, accurate docking and scoring. 1. Method and assessment of docking accuracy. *J Med Chem* **47**, 1739-49.

Garcea, G., Manson, M. M., Neal, C. P., Pattenden, C. J., Sutton, C. D., Dennison, A. R. & Berry, D. P. (2007). Glycogen synthase kinase-3 beta; a new target in pancreatic cancer? *Curr Cancer Drug Targets* **7**, 209-15.

Gillet, V. J., Khatib, W., Willett, P., Fleming, P. J. & Green, D. V. (2002). Combinatorial library design using a multiobjective genetic algorithm. *J Chem Inf Comput Sci* **42**, 375-85.

Gilson, M. K. & Zhou, H. X. (2007). Calculation of protein-ligand binding affinities. *Annu Rev Biophys Biomol Struct* **36**, 21-42.

Gohlke, H., Hendlich, M. & Klebe, G. (2000). Knowledge-based scoring function to predict protein-ligand interactions. *J Mol Biol* **295**, 337-56.

Gohlke, H. & Klebe, G. (2002). Approaches to the description and prediction of the binding affinity of small-molecule ligands to macromolecular receptors. *Angew Chem Int Ed Engl* **41**, 2644-76.

Gohlke, H. & Klebe, G. (2001). Statistical potentials and scoring functions applied to protein-ligand binding. *Curr Opin Struct Biol* **11**, 231-5.

Goldberg, D. (1989). *Genetic Algorithms in Search, Optimization, and Machine Learning*, Addison-Wesley.

Goodford, P. J. (1985). A computational procedure for determining energetically favorable binding sites on biologically important macromolecules. *J Med Chem* **28**, 849-57.

Graves, A. P., Brenk, R. & Shoichet, B. K. (2005). Decoys for docking. *J Med Chem* **48**, 3714-28.

Gschwend, D. A. & Kuntz, I. D. (1996). Orientational sampling and rigid-body minimization in molecular docking revisited: on-the-fly optimization and degeneracy removal. *J Comput Aided Mol Des* **10**, 123-32.

Halperin, I., Ma, B., Wolfson, H. & Nussinov, R. (2002). Principles of docking: An overview of search algorithms and a guide to scoring functions. *Proteins* **47**, 409-43.

Hancock, P. J. B. &. (1994). An Empirical Comparison of Selection Methods in Evolutionary Algorithms. *Lect Notes Comput Sc* **865**, 80 -94.

Handl, J., Knowles, J. & Kell, D. B. (2005). Computational cluster validation in post-genomic data analysis. *Bioinformatics* **21**, 3201-12.

Handschuh, S., Gasteiger, J. (2000). The search for the spatial and electronic requirements of a drug. *J Mol Model* **6**, 358-378.

Handschuh, S., Wagener, M. & Gasteiger, J. (1998). Superposition of three-dimensional chemical structures allowing for conformational flexibility by a hybrid method. *J Chem Inf Comput Sci* **38**, 220-32.

Hindle, S. A., Rarey, M., Buning, C. & Lengaue, T. (2002). Flexible docking under pharmacophore type constraints. *J Comput Aided Mol Des* **16**, 129-49.

Holland, J. H. (1992). *Adaptation in Natural and Artificial Systems: An Introductory Analysis with Applications to Biology, Control, and Artificial Intelligence*, MIT Press, Cambridge.

Horton, N. & Lewis, M. (1992). Calculation of the free energy of association for protein complexes. *Protein Sci* **1**, 169-81.

Huang, D. & Caflisch, A. (2004). Efficient evaluation of binding free energy using continuum electrostatics solvation. *J Med Chem* **47**, 5791-7.

Jackson, R. M., Gabb, H. A. & Sternberg, M. J. (1998). Rapid refinement of protein interfaces incorporating solvation: application to the docking problem. *J Mol Biol* **276**, 265-85.

Jain, A. N. (2006). Scoring functions for protein-ligand docking. *Curr Protein Pept Sci* **7**, 407-20.

Jain, A. N. (2003). Surflex: fully automatic flexible molecular docking using a molecular similarity-based search engine. *J Med Chem* **46**, 499-511.

Jain, T. & Jayaram, B. (2005). An all atom energy based computational protocol for predicting binding affinities of protein-ligand complexes. *FEBS Lett* **579**, 6659-66.

Jalaie, M. & Shanmugasundaram, V. (2006). Virtual screening: are we there yet? *Mini Rev Med Chem* **6**, 1159-67.

Jansen, J. M. & Martin, E. J. (2004). Target-biased scoring approaches and expert systems in structure-based virtual screening. *Curr Opin Chem Biol* **8**, 359-64.

Jiang, F. & Kim, S. H. (1991). "Soft docking": matching of molecular surface cubes. *J Mol Biol* **219**, 79-102.

Jianjun Hu, K. S., Zhun Fan1, Ronald C. Rosenberg3, Erik D. Goodman1. (2003). A Sustainable Multi-objective Evolutionary Optimization Framework. *Lec Notes Comput Sc* **2723**, 1029-1040.

Jones, G., Willett, P. & Glen, R. C. (1995). Molecular recognition of receptor sites using a genetic algorithm with a description of desolvation. *J Mol Biol* **245**, 43-53.

Jones, G., Willett, P., Glen, R. C., Leach, A. R. & Taylor, R. (1997). Development and validation of a genetic algorithm for flexible docking. *J Mol Biol* **267**, 727-48.

Jorgensen, W. L. (2004). The many roles of computation in drug discovery. *Science* **303**, 1813-8.

Jorgensen, W. L., & Tirado-Rives, J. (1988). The OPLS Potential Functions for Proteins. Energy Minimization for Crystals of Cyclic Peptides and Crambin. *J Am Chem Soc* **110**, 1657-1666.

Kellenberger, E., Rodrigo, J., Muller, P. & Rognan, D. (2004). Comparative evaluation of eight docking tools for docking and virtual screening accuracy. *Proteins*

57, 225-42.

Kitchen, D. B., Decornez, H., Furr, J. R. & Bajorath, J. (2004). Docking and scoring in virtual screening for drug discovery: methods and applications. *Nat Rev Drug Discov* **3**, 935-49.

Klabunde, T. & Hessler, G. (2002). Drug design strategies for targeting G-protein-coupled receptors. *Chembiochem* **3**, 928-44.

Klebe, G. (2006). Virtual ligand screening: strategies, perspectives and limitations. *Drug Discov Today* **11**, 580-94.

Knegtel, R. M., Kuntz, I. D. & Oshiro, C. M. (1997). Molecular docking to ensembles of protein structures. *J Mol Biol* **266**, 424-40.

Knowles, J. D. & Corne, D. W. (2000). Approximating the non-dominated front using the Pareto archived evolutionary strategy. *Evolutionary Computation Journal* **8**, 149-172.

Kollman, P. (1993). Free energy calculations - applications to chemical and biological phenomina. *Chem Rev* **7**, 2395.

Kontoyianni, M., McClellan, L. M. & Sokol, G. S. (2004). Evaluation of docking performance: comparative data on docking algorithms. *J Med Chem* **47**, 558-65.

Koshland, D. E., Jr. (2004). Crazy, but correct. *Nature* **432**, 447.

Kramer, B., Rarey, M. & Lengauer, T. (1999). Evaluation of the FLEXX incremental construction algorithm for protein-ligand docking. *Proteins* **37**, 228-41.

Kuntz, I. D., Blaney, J. M., Oatley, S. J., Langridge, R. & Ferrin, T. E. (1982). A geometric approach to macromolecule-ligand interactions. *J Mol Biol* **161**, 269-88.

Laumanns, M., Thiele, L., Deb, K. & Zitzler, E. (2002). Combining convergence and diversity in evolutionary multiobjective optimisation. *Evol Comput* **10**, 263-282.

Laurie, A. T. & Jackson, R. M. (2005). Q-SiteFinder: an energy-based method for the prediction of protein-ligand binding sites. *Bioinformatics* **21**, 1908-16.

Leach, A. R. (1994). Ligand docking to proteins with discrete side-chain flexibility. *J Mol Biol* **235**, 345-56.

Leach, A. R., Shoichet, B. K. & Peishoff, C. E. (2006). Prediction of protein-ligand interactions. Docking and scoring: successes and gaps. *J Med Chem* **49**, 5851-5.

Lee, B. K. & Richards, F. M. (1971). The interpretation of protein structures: estimation of static accessibility. *JMolBiol* **55**, 379–400.

Lee, M. C., Yang, R. & Duan, Y. (2005). Comparison between Generalized-Born and Poisson-Boltzmann methods in physics-based scoring functions for protein structure prediction. *J Mol Model* **12**, 101-10.

Lescot, E., Bureau, R., Sopkova-de Oliveira Santos, J., Rochais, C., Lisowski, V., Lancelot, J. C. & Rault, S. (2005). 3D-QSAR and docking studies of selective GSK-3beta inhibitors. Comparison with a thieno[2,3-b]pyrrolizinone derivative, a new potential lead for GSK-3beta ligands. *J Chem Inf Model* **45**, 708-15.

Liu, M. & Wang, S. (1999). MCDOCK: a Monte Carlo simulation approach to the molecular docking problem. *J Comput Aided Mol Des* **13**, 435-51.

Majeux, N., Scarsi, M., Apostolakis, J. Ehrhardt, C. and Calflisch, A. (1999). Exhaustive docking of molecular fragments with electrostatic solvation. **37**, 88-105.

McConkey, B. J., Sobolev, V. & Edelman, M. (2002). Quantification of protein surfaces, volumes and atom-atom contacts using a constrained Voronoi procedure. *Bioinformatics* **18**, 1365-73.

McLachlan, A. D. (1979). Gene duplications in the structural evolution of chymotrypsin. *J Mol Biol* **128**, 49-79.

Morreale, A., Gil-Redondo, R. & Ortiz, A. R. (2007). A new implicit solvent model for protein-ligand docking. *Proteins* **67**, 606-16.

Morris, G. M., Goodsell, D.S., Halliday, R.S., Huey, R., Hart, W.E., Belew, R.K. and & Olson, A. J. (1998). Automated docking using a Lamarckian genetic algorithm and an empirical binding free energy function. *J. Comp. Chem.* **19**, 1639-1662.

Mudher, A. & Lovestone, S. (2002). Alzheimer's disease-do tauists and baptists finally shake hands? *Trends Neurosci* **25**, 22-6.

Naerum, L., Norskov-Lauritsen, L. & Olesen, P. H. (2002). Scaffold hopping and optimization towards libraries of glycogen synthase kinase-3 inhibitors. *Bioorg Med Chem Lett* **12**, 1525-8.

Nelder, J. A. & Mead, R. (1965). A simplex method for function minimization. *Comput J* **7**, 308-313.

Nicolaou, C. A., Brown, N. & Pattichis, C. S. (2007). Molecular optimization using computational multi-objective methods. *Curr Opin Drug Discov Devel* **10**, 316-24.

Oleinikova, A., Smolin, N., Brovchenko, I., Geiger, A. & Winter, R. (2005). Formation of spanning water networks on protein surfaces via 2D percolation transition. *J Phys Chem B* **109**, 1988-98.

Onodera, K., Satou, K. & Hirota, H. (2007). Evaluations of molecular docking programs for virtual screening. *J Chem Inf Model* **47**, 1609-18.

Oshiro, C. M., Kuntz, I. D. & Dixon, J. S. (1995). Flexible ligand docking using a genetic algorithm. *J Comput Aided Mol Des* **9**, 113-30.

Perola, E., Walters, W. P. & Charifson, P. S. (2004). A detailed comparison of current docking and scoring methods on systems of pharmaceutical relevance. *Proteins* **56**, 235-49.

Pickett, S. D. & Sternberg, M. J. (1993). Empirical scale of side-chain conformational entropy in protein folding. *J Mol Biol* **231**, 825-39.

Polgar, T., Baki, A., Szendrei, G. I. & Keseru, G. M. (2005). Comparative virtual and experimental high-throughput screening for glycogen synthase kinase-3beta inhibitors. *J Med Chem* **48**, 7946-59.

Press, W. H., Flannery, B. P., Teukolsky, S. A. and Vetterling, W. T. (1992). Numerical Recipes in Fortran. 402-406.

Rarey, M., Kramer, B. & Lengauer, T. (1999). Docking of hydrophobic ligands with interaction-based matching algorithms. *Bioinformatics* **15**, 243-50.

Rarey, M., Kramer, B. & Lengauer, T. (1999). The particle concept: placing discrete water molecules during protein-ligand docking predictions. *Proteins* **34**, 17-28.

Rarey, M., Wefing, S. & Lengauer, T. (1996). Placement of medium-sized molecular fragments into active sites of proteins. *J Comput Aided Mol Des* **10**, 41-54.

Rudenko, O. & Schoenauer, M. (2004). A steady performance stopping criterion for Pareto-based evolutionary algorithms. *Proc. 6th Intl Conf. on Multi Objective*

Programming and Goal Programming, 2004.

Schaffer, J. D. (1984). Some experiments in machine learning using vector-evaluated genetic algorithms: PhD thesis, Vanderbilt University.

Shoichet, B. K., McGovern, S. L., Wei, B. & Irwin, J. J. (2002). Lead discovery using molecular docking. *Curr Opin Chem Biol* **6**, 439-46.

Solis, F. J. & Wets, R. J. B. (1981). Minimization by Random Search Techniques. *Math Oper Res* **6**, 19-30.

Sotriffer, C. & Klebe, G. (2002). Identification and mapping of small-molecule binding sites in proteins: computational tools for structure-based drug design. *Farmaco* **57**, 243-51.

Takashima, A. (2006). GSK-3 is essential in the pathogenesis of Alzheimer's disease. *J Alzheimers Dis* **9**, 309-17.

Tame, J. R. (1999). Scoring functions: a view from the bench. *J Comput Aided Mol Des* **13**, 99-108.

Taylor, J. S. & Burnett, R. M. (2000). DARWIN: a program for docking flexible molecules. *Proteins* **41**, 173-91.

Taylor, R. D., Jewsbury, P. J. & Essex, J. W. (2002). A review of protein-small molecule docking methods. *J Comput Aided Mol Des* **16**, 151-66.

Venkatachalam, C. M., Jiang, X., Oldfield, T. & Waldman, M. (2003). LigandFit: a novel method for the shape-directed rapid docking of ligands to protein active sites. *J Mol Graph Model* **21**, 289-307.

Verdonk, M. L., Chessari, G., Cole, J. C., Hartshorn, M. J., Murray, C. W., Nissink, J. W., Taylor, R. D. & Taylor, R. (2005). Modeling water molecules in protein-ligand docking using GOLD. *J Med Chem* **48**, 6504-15.

Verkhivker, G., Appelt, K., Freer, S. T. & Villafranca, J. E. (1995). Empirical free energy calculations of ligand-protein crystallographic complexes. I. Knowledge-based ligand-protein interaction potentials applied to the prediction of human immunodeficiency virus 1 protease binding affinity. *Protein Eng* **8**, 677-91.

Wang, J., Wolf, R. M., Caldwell, J. W., Kollman, P. A. & Case, D. A. (2004). Development and testing of a general AMBER force field. *J Comput Chem* **25**, 1157-74.

Wang, R., Lu, Y., Fang, X. & Wang, S. (2004). An extensive test of 14 scoring functions using the PDBbind refined set of 800 protein-ligand complexes. *J Chem Inf Comput Sci* **44**, 2114-25.

Wang, R., Lu, Y. & Wang, S. (2003). Comparative evaluation of 11 scoring functions for molecular docking. *J Med Chem* **46**, 2287-303.

Wang, R. & Wang, S. (2001). How does consensus scoring work for virtual library screening? An idealized computer experiment. *J Chem Inf Comput Sci* **41**, 1422-6.

Wei, B. Q., Baase, W. A., Weaver, L. H., Matthews, B. W. & Shoichet, B. K. (2002). A model binding site for testing scoring functions in molecular docking. *J Mol Biol* **322**, 339-55.

Wei, B. Q., Weaver, L. H., Ferrari, A. M., Matthews, B. W. & Shoichet, B. K. (2004). Testing a flexible-receptor docking algorithm in a model binding site. *J Mol Biol* **337**, 1161-82.

Westhead, D. R., Clark, D. E. & Murray, C. W. (1997). A comparison of heuristic search algorithms for molecular docking. *J Comput Aided Mol Des* **11**, 209-28.

Wright, T., Gillet, V. J., Green, D. V. & Pickett, S. D. (2003). Optimizing the size and configuration of combinatorial libraries. *J Chem Inf Comput Sci* **43**, 381-90.

Yang, J. M., Chen, Y. F., Shen, T. W., Kristal, B. S. & Hsu, D. F. (2005). Consensus scoring criteria for improving enrichment in virtual screening. *J Chem Inf Model* **45**, 1134-46.

Zhang, C., Liu, S., Zhu, Q. & Zhou, Y. (2005). A knowledge-based energy function for protein-ligand, protein-protein, and protein-DNA complexes. *J Med Chem* **48**, 2325-35.

Zhou, Z., Felts, A. K., Friesner, R. A. & Levy, R. M. (2007). Comparative performance of several flexible docking programs and scoring functions: enrichment studies for a diverse set of pharmaceutically relevant targets. *J Chem Inf Model* **47**, 1599-608.