Understanding Atmospheric Chemistry Using Graph-Theory, Visualisation and Machine Learning

Daniel Ellis

Doctor of Philosophy University of York Chemistry March 2020 Veritatem inquirenti, semel in vita de omnibus, quantum fieri potest, esse dubitandu:

In order to seek truth, it is necessary once in the course of our life, to doubt, as far as possible, of all things.

- Descartes, Rene, Principles of Philosophy

Abstract

Atmospheric chemistry mechanisms play a pivotal role in our understanding of societal problems such as air pollution, climate change and stratospheric ozone loss. This thesis explores the benefits of representing these mechanisms in terms of a mathematical graph (or network) which connects species (nodes) through reactions (edges). We use the Dynamically Simple Model of Atmospheric Chemical Complexity and the Master Chemical Mechanism to explore the number of real-world scenarios - using graph theory and machine learning to visualise, understand and analyse the underlying chemistry of the lower atmosphere.

We begin by exploring different visualisation techniques to depict chemistry within the atmosphere. It is found that the sociograph framework provides the most (visually) intuitive delineation of the species and their reactions. For large, complex systems, this type of quasi-qualitative analysis has its limitations - physical and cognitive. Instead, the relationships between species in the network are quantified using graph centrality metrics and then compared against well-established methods such as the Jacobian and Rate Of Production Analysis. Further development of graph theory allows us to couple natural language processing, network decomposition, and clustering to identify species with similar lifetimes, reaction styles, or temporal profiles.

Having explored aspects of mechanism analysis, visualisation and reduction, we examine how varying representations of species structure can affect the patterns highlighted by unsupervised machine learning models. This is done by visualising them in 2D space and serves as a precursor to potential future work involving Graph Convoluted Neural Networks - thus consolidating the contents of this thesis.

Ultimately it is found that using a graph-theory approach can prove highly beneficial in the understanding and explanation of chemical mechanisms, but should not (as of yet) be used in substitution of existing investigation and reduction methods.

Acknowlegements

First and foremost, I would like to thank my family, for which I wish I had spent more time with rather than being squirrelled away in an office. I could have gotten here only with your support for which I am grateful.

Next, are my supervisors Andrew and Mat - mainly putting up with me but also being highly supportive throughout the PhD.

A special mention also needs to go out to all my friends and colleagues at (Westminster Boating Base) who helped keep me sane irrespective to the severe watersport withdrawal, kept me employed and invited me to the annual France whitewater trip each year.

On a similar note, thanks to all my university friends and collegues (Rosie, Tomas, Ben, Mike, Killian, Pete, Peter etc.) with whom I could bounce ideas and discuss random work adjacent projects when needed.

Additionally the Earth0 HPC cluster, for not failing on me, even after it was no longer officially in use (or supported).

Finally, I am not sure whether to thank or apologise to those unfortunate enough to have to read any part of this work, but I am grateful anyway. viii

Declaration

I declare that this thesis is a presentation of original work and I am the sole author. This work has not previously been presented for an award at this, or any other, University. All sources are acknowledged as References. <u>x</u>_____

Contents

		List of	f Abbreviations	xviii
1	Int	roduct	tion	1
	1.1	Backg	round	4
		1.1.1	A Preface On Humanity And The Climate	4
		1.1.2	Formation Of The Atmosphere	4
		1.1.3	Rise Of The Homo Sapiens ('Wise Man')	4
	1.2	Motiva	ation (How The Atmosphere Affects Us)	5
		1.2.1	Air Quality - It Is The Air We Breathe	5
		1.2.2	Stratospheric Ozone - The Protective Barrier	6
		1.2.3	Changing Climate	7
	1.3	Tropos	spheric Chemistry	7
		1.3.1	Ozone Production/Loss $\ldots \ldots \ldots$	7
		1.3.2	The NOx Cycle	8
		1.3.3	HOx Cycle	10
			1.3.3.1 The Hydroperoxide Radical	10
	1.4	Model	ling The Earth	12
		1.4.1	Earth System Models (ESM)	12
	1.5	The 0	D Chemical Box Model	13
		1.5.1	Chemical Mechanisms	14
		1.5.2	Numerical Integration	14
			1.5.2.1 Non-Stiff Equations $\ldots \ldots \ldots$	14
			1.5.2.2 Numerically Stiff Equations (Atmospheric Chemistry) $\ldots \ldots \ldots$	15
		1.5.3	The Model Development Cycle	15
		1.5.4	The Dynamically Simple Model Of Atmospheric Chemical Complexity	16
	1.6	Thesis	3 Layout	16
2	Vis	sualisa	tion and its use in understanding Complex Data	21
	2.1	Introd	$uction \ldots \ldots$	24
		2.1.1	Communicatory Practices Of Early Humans	24
		2.1.2	The Origin Of Big Data	25
	2.2	Visual	lisation Design	26
		2.2.1	Storytelling	26

		2.2.2	Metaph	or Selection	27
			2.2.2.1	Nature-Inspired	27
			2.2.2.2	Man-Made	28
			2.2.2.3	Composite	29
			2.2.2.4	Domain Specific	30
	2.3	Visual	lising Rela	ationships	33
		2.3.1	The Dat	taset	33
		2.3.2	The Soc	ciograph	34
			2.3.2.1	The Chord Diagram	34
			2.3.2.2	Direct Representation Of The Relational Matrix	38
			2.3.2.3	Arc Diagrams	39
			2.3.2.4	The Traditional Network Graph	43
	2.4	Conclu	usion .		46
3	App	olying	Visual A	Analytics to the Atmospheric Chemistry Network	53
	3.1	Introd	uction .		56
		3.1.1	Network	s And Their Role In Visual Analytics	56
		3.1.2	Graphs	In Chemistry	56
			3.1.2.1	Using Sociograms To Describe Reactions	57
		3.1.3	Modelin	g Chemistry As A Directed Graph.	57
	3.2	Graph	Syntacti	CS	60
		3.2.1	Selecting	g The Correct Evaluation Criteria	61
			3.2.1.1	Edge Crossing	62
			3.2.1.2	Node Distribution and Overlap	63
		3.2.2	Automa	ted Graph Drawing Layouts	63
			3.2.2.1	Replication Of Hand-Drawing Methods	63
			3.2.2.2	Projection Based	64
			3.2.2.3	Force-Directed	66
	3.3	Select	ing The I	Best Graph Drawing Layout.	69
		3.3.1	Graph-1	Node Distribution	69
			3.3.1.1	Evaluating Node Distribution For The Beijing Mechanism $\ \ . \ . \ .$.	70
			3.3.1.2	Distribution Of Primary Emitted VOCs	72
			3.3.1.3	Calculation Of Spatial Clustering	72
	3.4	Graph	Semanti	cs	76
		3.4.1	Limitati	ions	76
		3.4.2	Node Er	ncoding	78
		3.4.3	Edge Pr	operties	82
			3.4.3.1	Muti-Variate Edges	82
			3.4.3.2	Edge Direction	82
			3.4.3.3	Edge Shape	82
			3.4.3.4	Edge Bundling	85

		3.4.3	B.5 Power, Routing And Confluence	86
		3.4.3	3.6 Angle / Continuity	89
		3.4.4 Tem	poral Projection	90
		3.4.5 Add	itional Dimensions	92
		3.4.6 Sum	mary Of Semantic Representation	92
	3.5	A Chemistr	y Case Study	93
		3.5.1 Sem	antic And Syntatic Considerations.	93
		3.5.1	.1 Syntatic Representation	93
		3.5.1	.2 Example Semantic Representation Using A Methane Mechanism	93
		3.5.2 A M	odel Of Beijing	94
		3.5.2	2.1 Similarity Between Graph Shape	96
		3.5.2	2.2 Network Branch Classification	98
	3.6	Conclusion		100
4	\mathbf{Ch}	emical mod	lel diagnostics using graph theory and metrics.	.09
	4.1	Introduction	n	112
	4.2	Graph Metr	rics	113
		4.2.1 Cent	trality Metrics And Academic Publishing.	113
		4.2.2 The	Master Chemical Mechanism (MCM)	114
		4.2.3 Data	a Collection	114
		4.2.4 Visu	alising The Data.	116
		4.2.5 Filte	ring The Data	116
		4.2.6 The	Co-Citation Network	119
		4.2.7 The	Co-Authorship Network	119
	4.3	Metric Anal	lysis	120
		4.3.1 Degr	cee Centrality	121
		4.3.2 Clos	eness Centrality	123
		4.3.3 Betw	veenness	125
		4.3.4 Spec	tral Methods And Matrix Analysis	126
		4.3.5 Page	e Rank	127
		4.3.5	5.1 The Google Matrix	127
		4.3.5	5.2 Solving The Algebra	128
		4.3.5	5.3 Prediction	129
		4.3.6 Cone	clusions	130
	4.4	Classifying	The Master Chemical Mechanism Network	131
		4.4.1 Netv	vork Density	131
		4.4.2 Sma	ll World Phenomena	132
		4.4.3 Powe	er Law And Scale-Free Graphs	133
		4.4.4 Desc	ribing The MCM Network	135
	4.5	Graph Cons	struction Methodology	135
		4.5.0	0.1 Concentration Time Series	136

			4.5.0.2	Rate Of Production And Loss	137
			4.5.0.3	The Jacobian	139
		4.5.1	Graph (Construction Methodology For Simulated Data	140
		4.5.2	A Pract	ical Example Using The MCM	141
	4.6	Case S	Study		144
		4.6.1	Establis	hing Initial Conditions From Observational Data	144
			4.6.1.1	The Origin Of Artificial Neural Networks	144
			4.6.1.2	The Multi-Layer Perceptron	145
			4.6.1.3	Applying The Mlpregressor To Observational Data	147
			4.6.1.4	Model Initialisation Procedure	153
			4.6.1.5	Extracting The Required Results	153
			4.6.1.6	Unifying The Results	158
		4.6.2	Compar	ing Results	158
			4.6.2.1	What Is TF-IDF	158
			4.6.2.2	Metric Comparison	160
			4.6.2.3	Individual Categories	161
		4.6.3	Scenario	Analysis	162
		4.6.4	Providir	ng An Overall Overview Using The TF-IDF And The Metric Sum	167
	4.7	Causa	lity Analy	ysis using the Jacobian and Pagerank	168
		4.7.1	Source A	Analysis using PageRank	168
		4.7.2	Mathem	natical Equivalent of Edge Reversal	168
		4.7.3	A Calcu	lated Case Study	169
			4.7.3.1	Personalised PageRank	170
			4.7.3.2	Iterative Analysis of the Jacobian	171
		4.7.4	Verdict		172
	4.8	Conclu	usions .		172
5	\mathbf{Us}	ing Gr	aph Clu	ustering And Natural Language Processing To Aid Mechanisn	n
	Red	luction	•		183
	5.1	Introd	uction .		186
	5.2	Mecha	nism Rec	luction	187
		5.2.1	Species	Categories	187
		5.2.2	Reaction	n Removal	187
		5.2.3	Species	Removal	188
		5.2.4	Lumping	g	189
			5.2.4.1	Chemical Lumping	189
	5.3	Data S	Setup .		189
		5.3.1	The Me	chanism	190
		5.3.2	The Box	x-Model	190
		5.3.3	Model I	nputs	190
	5.4	Graph	Based R	eduction	191

		5.4.1	Graph Parallels	191
		5.4.2	Types Of Graph Clustering	191
		5.4.3	Walk/Flow-Based Clustering	192
		5.4.4	Louvain Clustering	193
		5.4.5	Infomap Clustering	193
	5.5	Select	ion Criteria For Graph Clustering	194
	5.6	Evalua	ation Of Infomap On A Real Simulation.	195
		5.6.1	Species Type And Clustering	197
		5.6.2	Number Of Clusters	200
	5.7	Reduc	tion Through Lifetime	201
			5.7.0.1 Calculating The Lifetime	202
		5.7.1	Comparing Magnitude And Direction	202
			5.7.1.1 Euclidian Distance	202
			5.7.1.2 Cosine Distance	203
		5.7.2	Temporal Lifetime Vector Comparison	203
		5.7.3	A Quick Comparison	205
	5.8	Result	s2	206
		5.8.1	The Co-Grouping Network	206
		5.8.2	Comparing Daytime And Nighttime Groups	207
		5.8.3	Determining Cluster Suitabiltiy	208
	5.9	Concl	usions	211
6	5.9 Co	Concle mputa	usions	211
6	5.9 Co teri	Concla mputa ng	usions	211 ? 17
6	5.9 Co teri 6.1	Concla mputa ng Introd	usions 2 tional Learning of Species Structure using Visualisation and Vector Clus- 2 uction 2	211 211 217 220
6	5.9 Cor teri 6.1	Concle mputa ng Introd 6.1.1	usions 2 tional Learning of Species Structure using Visualisation and Vector Clus- 2 luction 2 Historical Significance 2	211 217 220 220
6	5.9 Con teri 6.1	Conche mputa ng Introd 6.1.1 6.1.2	usions 2 tional Learning of Species Structure using Visualisation and Vector Clus- 2 huction 2 Historical Significance 2 Theory And Simulation In Science 2	211 217 220 220 220
6	5.9 Co: teri 6.1	Conche mputa ng Introd 6.1.1 6.1.2 6.1.3	usions 2 tional Learning of Species Structure using Visualisation and Vector Clus- 2 luction 2 historical Significance 2 Theory And Simulation In Science 2 Chapter Aims 2	211 217 220 220 220 220
6	 5.9 Conternit 6.1 6.2 	Conche mputa ng Introd 6.1.1 6.1.2 6.1.3 Specie	usions 2 tional Learning of Species Structure using Visualisation and Vector Clus- 2 luction 2 historical Significance 2 Theory And Simulation In Science 2 Chapter Aims 2 es Of The MCM And Ways To Represent Them. 2	211 217 220 220 220 220 221 221
6	 5.9 Conterning 6.1 6.2 	Conche mputa ng Introd 6.1.1 6.1.2 6.1.3 Specie 6.2.1	usions 2 tional Learning of Species Structure using Visualisation and Vector Clus- 2 luction 2 Historical Significance 2 Theory And Simulation In Science 2 Chapter Aims 2 so Of The MCM And Ways To Represent Them. 2 Input Generation 2	2111 217 220 220 220 221 221 221 221
6	 5.9 Conternio 6.1 6.2 	Conche mputa ng Introd 6.1.1 6.1.2 6.1.3 Specie 6.2.1 6.2.2	usions 2 tional Learning of Species Structure using Visualisation and Vector Clus- 2 luction 2 historical Significance 2 Theory And Simulation In Science 2 Chapter Aims 2 so Of The MCM And Ways To Represent Them. 2 Input Generation 2 Manual Categorisation 2	2111 217 2200 2200 2201 221 221 221 221
6	 5.9 Conternio 6.1 6.2 	Conche mputa ng Introd 6.1.1 6.1.2 6.1.3 Specie 6.2.1 6.2.2 6.2.3	usions 2 tional Learning of Species Structure using Visualisation and Vector Clus- 2 uction 2 historical Significance 2 Theory And Simulation In Science 2 Chapter Aims 2 so Of The MCM And Ways To Represent Them. 2 Input Generation 2 Manual Categorisation 2 Tokenization 2	211 217 220 220 220 221 221 221 221 222 222 223
6	 5.9 Conternit 6.1 6.2 	Conche mputa ng Introd 6.1.1 6.1.2 6.1.3 Specie 6.2.1 6.2.2 6.2.3	usions 2 tional Learning of Species Structure using Visualisation and Vector Clus- 2 uction 2 Historical Significance 2 Theory And Simulation In Science 2 Chapter Aims 2 us Of The MCM And Ways To Represent Them. 2 Input Generation 2 Manual Categorisation 2 6.2.3.1 Species Names 2	2111 217 220 220 220 221 221 221 222 222 223 224
6	 5.9 Conternio 6.1 6.2 	Conche mputa ng Introd 6.1.1 6.1.2 6.1.3 Specie 6.2.1 6.2.2 6.2.3	usions 2 tional Learning of Species Structure using Visualisation and Vector Clus- 2 uction 2 Historical Significance 2 Theory And Simulation In Science 2 Chapter Aims 2 Input Generation 2 Manual Categorisation 2 Coloration 2 6.2.3.2 SMILES Strings	2111 217 220 220 220 221 221 221 222 223 224 224
6	5.9 Conteri 6.1 6.2	Conche mputa ng Introd 6.1.1 6.1.2 6.1.3 Specie 6.2.1 6.2.2 6.2.3 6.2.3	usions 2 tional Learning of Species Structure using Visualisation and Vector Clus- 2 luction 2 Historical Significance 2 Theory And Simulation In Science 2 Chapter Aims 2 ss Of The MCM And Ways To Represent Them. 2 Input Generation 2 Manual Categorisation 2 6.2.3.1 Species Names 2 6.2.3.2 SMILES Strings 2 Graph Inspired 2 2	2111 217 220 220 221 221 221 222 223 224 224 224 224
6	 5.9 Conternio 6.1 6.2 	Conche mputa ng Introd 6.1.1 6.1.2 6.1.3 Specie 6.2.1 6.2.2 6.2.3 6.2.4	usions 2 tional Learning of Species Structure using Visualisation and Vector Clus- 2 uction 2 Historical Significance 2 Theory And Simulation In Science 2 Chapter Aims 2 so Of The MCM And Ways To Represent Them. 2 Input Generation 2 Manual Categorisation 2 6.2.3.1 Species Names 6.2.3.2 SMILES Strings Graph Inspired 2 6.2.4.1 The Species Graph (Fingerprint)	2111 217 2200 2200 2211 2211 2221 2223 2223 2224 2224 2225 2225
6	 5.9 Conternio 6.1 6.2 	Conchemputang Introd 6.1.1 6.1.2 6.1.3 Specie 6.2.1 6.2.2 6.2.3 6.2.4	usions 2 tional Learning of Species Structure using Visualisation and Vector Clus- 2 huction 2 Historical Significance 2 Theory And Simulation In Science 2 Chapter Aims 2 ss Of The MCM And Ways To Represent Them. 2 Input Generation 2 Tokenization 2 6.2.3.1 Species Names 6.2.3.2 SMILES Strings Graph Inspired 2 6.2.4.1 The Species Graph (Fingerprint) 2.4.2 Node Embeddings (Node2Vec)	2111 217 220 220 220 221 221 221 221 222 223 224 224 224 225 225 226
6	5.9 Conteri 6.1 6.2	Conche mputa ng Introd 6.1.1 6.1.2 6.1.3 Specie 6.2.1 6.2.2 6.2.3 6.2.4 6.2.4	usions 2 tional Learning of Species Structure using Visualisation and Vector Clus- 2 huction 2 Historical Significance 2 Theory And Simulation In Science 2 Chapter Aims 2 ss Of The MCM And Ways To Represent Them. 2 Input Generation 2 Manual Categorisation 2 6.2.3.1 Species Names 2 Graph Inspired 2 6.2.4.1 The Species Graph (Fingerprint) 2 6.2.4.2 Node Embeddings (Node2Vec) 2 Molecular Fingerprints 2 2	2111 217 220 220 221 221 221 222 223 224 224 225 225 225 225 226 227
6	 5.9 Conternio 6.1 6.2 	Conche mputa ng Introd 6.1.1 6.1.2 6.1.3 Specie 6.2.1 6.2.2 6.2.3 6.2.4 6.2.5	usions 2 tional Learning of Species Structure using Visualisation and Vector Clus- 2 uction 2 Historical Significance 2 Theory And Simulation In Science 2 Chapter Aims 2 so Of The MCM And Ways To Represent Them. 2 Input Generation 2 Manual Categorisation 2 6.2.3.1 Species Names 2 6.2.3.2 SMILES Strings 2 Graph Inspired 2 2 6.2.4.1 The Species Graph (Fingerprint) 2 6.2.4.2 Node Embeddings (Node2Vec) 2 Molecular Fingerprints 2 2	211 217 220 220 221 221 221 221 222 223 224 224 225 225 225 225 226 227 228
6	 5.9 Conternio 6.1 6.2 	Conche mputa ng Introd 6.1.1 6.1.2 6.1.3 Specie 6.2.1 6.2.2 6.2.3 6.2.4 6.2.4	usions 2 tional Learning of Species Structure using Visualisation and Vector Clus- 2 uction 2 Historical Significance 2 Theory And Simulation In Science 2 Chapter Aims 2 as Of The MCM And Ways To Represent Them. 2 Input Generation 2 Manual Categorisation 2 6.2.3.1 Species Names 2 Graph Inspired 2 6.2.4.1 The Species Graph (Fingerprint) 2 6.2.4.2 Node Embeddings (Node2Vec) 2 Molecular Fingerprints 2 6.2.5.1 Molecular Quantum Numbers (MQN) 2 6.2.5.2 Molecular Access System (MACCS) 2	211 217 220 220 221 221 221 222 223 224 224 224 225 225 225 226 227 228 228

		6.3.1	Preparation Of The Data	229
		6.3.2	Principle Component Analysis (PCA)	229
			6.3.2.1 Mathematical Explanation Of PCA	230
		6.3.3	t-Distributed Stochastic Neighbor Embedding (t-SNE)	231
			6.3.3.1 Mathematical Explanation Of t-SNE	232
		6.3.4	PCA vs t-SNE, A Quick Comparison.	233
		6.3.5	The Auto-Encoder (AE)	235
			6.3.5.1 Demonstration Of Non-Linear Activation Functions	236
		6.3.6	Node2Vec	237
			6.3.6.1 Sentence Construction By Sampling Of A Network	238
			6.3.6.2 Word2Vec	239
		6.3.7	Summary Of Dimensionality Reduction Methods	239
	6.4	Visual	isation Of Clustering	239
		6.4.1	Viewing The 2D Species Embeddings	239
		6.4.2	Exposing Overlapping Data	240
		6.4.3	Gooey Effect (Gaussian Blur)	240
		6.4.4	Four Colours Theorem	240
	6.5	Cluste	r Evaluation	241
		6.5.1	Automated Selection Of Clusters	241
			6.5.1.1 Clustering (Silhouette) Coefficient	243
		6.5.2	Feature Extraction	243
			6.5.2.1 Random Forests	243
			6.5.2.2 Calculating Importance Using Random Forests	244
	6.6	Result	s	245
		6.6.1	Visual Overview	245
		6.6.2	Mathematical Cluster Analysis	249
		6.6.3	Feature Selection Comparison	251
		6.6.4	Cluster Comparison	256
		6.6.5	Bi / Tri Cluster Groups	256
			6.6.5.1 Multicluster Groups	259
	6.7	Conclu	isions	262
7	Co	nclusic	ons and Future Work	269
	7.1	Conclu	isions	272
	7.2	Result	s	272
	7.3	Genera	al Overview	274
	7.4	Future	e Work	274
		7.4.1	Policy and Communication	274
		7.4.2	Dynamic Box Model Emulation	274
		Repro	ducability	275
		1	· · · · · · · · · · · · · · · · · · ·	

Aj	Appendices 2								
\mathbf{A}	Supplementary Mathematics								
	A.1	PCA	280						
		A.1.1 Statistics	280						
		A.1.2 Matrices and Eigenvectors	280						
	A.2	t-SNE	281						
		A.2.1 Student t distribution	281						
		A.2.1.1 t-Score	281						
в	Neu	ral Network Activation Functions	282						
	B.1	Binary Step	282						
	B.2	Linear	282						
	B.3	Sigmoid / Logistic	283						
	B.4	Hyperbolic Tangent	283						
	B.5	Rectified Linear Unit	284						
	B.6	Swish	284						
	B.7	A note on backpropagation	285						
С	Gra	phs and Networks	286						
	C.1	Heavily Labeled Citation Graph	286						
	C.2	Centrality on the UK rail network.	287						
D	Miscellaneous								
	D.1	Correspondence with Mike Jenkin	289						
	D.2	Functional Groups	292						
\mathbf{E}	Cha	pter Keywords	293						
	E.1	Introduction	293						
	E.2	Visualisation and its use in understanding Complex Data	293						
	E.3	Applying Visual Analytics to the Atmospheric Chemistry Network	294						
	E.4	Chemical model diagnostics using graph theory and metrics	294						
	E.5	Using Graph Clustering And Natural Language Processing To Aid Mechanism Reduc- tion.	294						
	E.6	Computational Learning of Species Structure using Visualisation and Vector Clusterin	g295						
	E.7	Conclusions and Future Work	295						

List of Abbreviations

Atmosphere

OH + HO
$0\Pi + \Pi O_2$
$\mathrm{NO} + \mathrm{NO2}$
Σ oxidized atmospheric odd-nitrogen species
NOy - NOx
PeroxyAcyl Nitrate
parts per {million, billion, trillion} by volume

Modelling

DSMACC GEOSChem	Dynamically Simple Model of Atmospheric Chemical Complexity Chemistry component of NASA's Goddard Farth Observing Sys-
GEOSCHEIII	tem
KPP	Kinetic Pre Processor
ROPA	Rate of Production (and Loss) Analysis
TUV	Tropospheric, Ultraviolet and Visible Radiation Model

Artificial Intelligence

Common Representative Intermediates
International Chemical Identifier (developed by IUPAC)
International Union of Pure and Applied Chemistry
Molecular ACCess System
Master Chemical Mechanism
Molecular Quantum Number
SMILES arbitrary target specification
Simplified Molecular-Input Line-Entry System

Artificial Intelligence

\mathbf{AE}	Auto Encoder
DBSCAN	Density-Based Spatial Clustering of Applications with Noise
DR	Dimensionality Reduction
GMM	Gaussian Mixture Model
GNN	Graph Neural Network
\mathbf{ML}	Machine Learning
OPTICS	Ordering Points To Identify the Clustering Structure
PCA	Principle Component Analysis
t-SNE	t-distributed Stochastic Neighbor Embedding

List of Figures

1.1	Reported deaths attributed to air pollution by country (2016) A cartogram (and cloropleth) showing the number of premature deaths attributed to ambient air pollution per 100,000. The colour bar range is from 9 in Canada (light blue) to 170 in eastern Europe (navy) people. Data Source: [WHO, 2016]	6
1.2	Changes in NOx concentrations due to anthropogenic emissions. A reduction in activity and trasport produces a notable decrease of Nitrogen dioxide concentrations in the troposphere. Source: [Stevens, 2020]	9
1.3	Spatial and temporal scales of variability of atmospheric species. This shows that the longer lived a species, the further it is likely transported through the atmosphere. Source: [Seinfeld and Pandis, 2016]	11
1.4	The HOx cycle. The OH aids in the oxidation of VOCs, which makes them more water soluble - this allowing for their removal from the atmosphere. In a high NOx evironment the RO_2 radicals can then reactive with NO to produce NO_2 , and consequently more errors	19
15	A diagram showing the longitudinal lateral and vertical decomposition of	12
1.0	a 3D global model. (Diagram not of GeosChem.) Source: [Henderson-Sellers, 2015]	13
1.6	The scientific development cycle. This shows the iterative nature between modelling, observation and laboratory experimentation	15
2.1	An example pictograph: a 2000 year old petroglyth in Utah titled 'Tse Hane' (rock that tells a story). Source: [ugc, 2019]	25
2.2	Data size and complexity increases with time due to (availability or improve- ments in scientific understanding). The red line shows the change in the number of websites (\log_{10} normalised) within the world wide web domain for years 1996-2015. Similarly, we compare the various releases of the master chemical mechanism (a dis- crete process), where more and more complex reaction schemes are iteratively appended over time. These are introduced later on. <i>NOTE: The two lines are here to illustrate</i> the growing trends between the subjects and are not incomparable due to their different scales and continuous/discrete natures. Source: [InternetLiveStats, 2020; Jenkins, 2002]	26
2.3	Two tree-inspired visualisations. (a) shows the decisions made on a single decision tree within a random Forrest. Hear each branch split corresponds to a decision and the node/leaf colour represents the category of the decision. Stronger and more important decisions correspond to larger leaves and thicker branches. (b) shows a radial plot in the shape of a tree trunk. Here time is shown radiating outwards from the centre. This allows us to spot any changes in events - much like the rings of a tree can be used to identify when natural disasters (such as tsunamis or avalanches) have struck them. This specific visualisation shows the net flux of species from a chemical simulation. These are coloured from low fluxes (blue) to high fluxes (red). The abrupt changes here show the diurnal cycle where photochemical reactions stop and then start up again.	28
2.4	The 1933 tube map design for London. Source: [Beck, 2017]	29
2.5	Caption for LOF	30

2.6	Two T-ϕ-gram plots. An example of the T- ϕ -gram data (a), and instructions on how to interpret it (b). The full scale figures are found within the source. Source: [Brooks, 2019]	31
2.7	Examples of connected scatterplots (a) A connected scatterplot comparing the number of vehicle-related fatalities with the number miles driven per capita [Fairfield, 2012] (b) a plot showing how the correlation of two spatial variables are transformed from a classical scatter/line plot into a connected scatterplot [Haroz et al., 2016]. Here the blue and green lines represent different values changing over time (orange) across the x-axis. The figures to the right show how different correlations are transformed into a connected scatter plot, where the variables (blue and green) are plotted across each axis, and time (orange) is shown as a line joining each datapoint. The data points are plotted in equal lengths of time, meaning that the distance between consecutive points informs us of the amount each variable has changed.	32
2.8	The generation flowchart used within the MCM. This shows the process under- gone by a new species to generate its products. Any unseen products are then fed back into the flowchart until the entire mechanism has been produced. Source: [Saunders et al., 2003]	34
2.9	The generation flowchart used within Gecko. This is very similar to the MCM protocol (Figure 2.8), but provides a clearer representation of the machine read desicions for the mechanism generation. Source: [Aumont et al., 2005]	35
2.10	A chord diagram of the protocol structure. This shows the proportional prob- ability a species from the MCM will follow one or more of the paths presented in Figure 2.9. The outside ring represents the radical (red), and non-radical (gold) split between groups. The inside ring splits these into individual groups, providing a finer level of detail.	37
2.11	The adjacency or relational matrix is showing species from a propane subset of the MCM. Here species are sorted by the number of carbons they contain (red= 3, orange= 1). Black cells represent reactions between species containing different numbers of carbons. If looking at clusters in a network (Chapter 5) sorting elements in an adjacency matrix aid in the visual identification of these. Squares features along the diagonal indicate a larger number of links between grouped items colocated in the axis, and less to those in other locations (clusters).	38
2.12	An arc diagram from the MCM database derived by the protocol (Fig- ure 2.9). All possible pathways for each species are extracted. These are then grouped and sorted by the number of functional group/pathways available. Species are shown along the blue line and sorted by increasing number groups to the right. Reaction pairs (reactant-product) are then depicted through the use of arcs. If a species reacts to pro- duce a species with more functional groups, it is represented by a forwards facing arc. If the product contains a smaller number of functional groups, the drawn arc moves downwards and backwards.	39
2.13	Arc diagram features for the Hydroperoxyl and Hydroxide radicals. The HO_x cycle chemistry (b) can ge seen within certain groups in the network. The main ones of these are hilighted in (a).	41
2.14	Arc diagram features for photolysis and hydroxide reactions. Photolysis results in species with a reduced number of functional groups, and therefore longer arcs. OH reactions for the same species do not produce such a drastic change on group number, and therefore have a smaller arc length.	41
2.15	Arc diagram features for the PAN reactions.	42
2.16	Comparing a range of MCM and CRI mechanims using their graph shape and structure. Source: Ellis [2020]	44

2.17	Voronoi cells of each node from the graph layout - used to identify changes in mechanisms. A difference plot between the different graphs in Figure 2.16. These use colours to show us species that are added or taken away between different versions. Subplots (a) and (b) show the increases in mechanism size of the MCM whilst (c) and (d) show the reduction from MCM v3.2 to CRI v2.0(r1), and followed by the fith reduction to CRI v2.0(r5). Figure colouring: purple cells only exist within the first mechanism, pink only exist within the second, and blue are present in both. Source: Ellis [2020]	45
3.1	The molecule C141CO3 (MCM name) shown in both 2D and 3D node- link structures. This is a the result of a series of inorganic species reactions and a desocciation from BCARY - the only sesqueterpine in the MCM. 3D visualisation by [Bergwerf, 2019].	57
3.2	A systematic representation of the degregation of butane. Using this we are able to see the process C_4H_{10} undergoes before its ultimate demise as carbon dioxide and water. Source: [Jenkin et al., 1997]	58
3.3	The Roche Metabolic Pathways of the human body. This example demonstrates the ability to manually represent the complex chemistry of the body using a graph structure. (Original A0 version is available at the source). Source: [Michal, 1965]	59
3.4	Comparison of different representations of flight data by [Martin Grandjean, 2016]. The top figure shows the data represented by a force-directed graph layout (described below) and a Geo-layout showing each point at its location on the Earth.	61
3.5	An orthogonal circuit schematic of the Model 3 240 portable cathode ray tube television. The circuit schematic of the television (top left) shows a much simpler representation of how different components within the television are connected. Source: [JVC, 2020]	62
3.6	A transit map showing all the possible routes from methane to carbon diox- ide. This was drawn using MemoryMap [Foo, 2019] and uses a version of the MCM methane subset, where carbon dioxide has been introduced	64
3.7	A selection of map projections. These have been created using DataDrivenDocuments [Bostock, 2012] and show a range of methods for mapping the spheroid shape of the Earth onto a 2D plane.	65
3.8	The Mercator Projection. (a) represents the output from the Mercator graph layout algorithm. (b) provides a kernel density analysis of the node distribution within this. Here (a) shows graph structure by revealing the density of connections between different nodes, while (b) reveals the density of nodes at a specific location.	65
3.9	Demonstration of the formation for a quadtree from a force directed graph of Methane (including inorganics) . (a) shows the force directed graph of Methane from which the quadtree has been constructed- edge colours represent the flux between species. Here we partition the area into 4 and start at the top-leftmost cell. This is then partitioned into 4 itself in a recursive process until there is only one point per cell. We repeat the process to any remaining cells in a clockwise manner (b). The hierarchical tree (b right) shows the containing structure for each node. Here the colours represent the order in which nodes have selected (starting at pink and ending in blue).	67
3.10	A graph of the full MCM - a hairball. The high number of nodes and edges (especially those to inorganic species), causes a high degree of obfuscation, rendering the graph unusable. Species with a large number of reactions (links) are labelled	70
3.11	Contour and kernel density plots showing the node distribution for different graph layouts. Line charts show the distribution of nodes in the x and y directions, while the contours represent density with respect to the location of each node (the crosses). Primary emitted species are coloured orange, and the darker contour polygons show areas of higher density.	71
3.12	A visual analysis of node-cluster density using Voronoi tesselation. Each polygon is centred on a node - its area represents the space between the node and its	

nearest neighbours. Colours follow the normalised size of the Voronoi cells/polygons. . 74

3.13	Voronoi $\log_{10}(\text{Area})$ BoxPlot for all plots in Figure 3.12. This provides a mathematical analysis for the areas around each node within a graph.	75
3.14	Important acuities in visualisation. Here a double prime " represents an arc minute which equates to an angle of 1/60 of a degree. A single prime ' is that of an arc second with is 1/60th of an arc minute (or 1/120 of a degree). For comparison the maximum angular resulution of the human eye is stated as 28 arc seconds [Deering, 1998] - this means we can only ever see up to 28 nodes or 2 verneers (disjoint lines) at any one time. Source: [Jankun-Kelly et al., 2014; Ware, 2013c]	77
3.15	A graph showing 5 different node encoding methods. These are Circle Atrributes (red), Chemical structure (blue), Species Name (green), External Labels (maroon) and interactive selection (orange). The network shows the Common Representative Intermediate species [Jenkin et al., 2008] mechanism. Node colours represent primary emitted VOCs (red), MCM species (orange) and lumped CRI-only species (blue).	79
3.16	Using mouseover edge-selection to highlight all links related to a node. This figure shows how in using interactivity it is possible to reduce clutter and filter the information presented by a densely populated graph. In this case, the Mercator projection (Subsubsection 3.2.2.2) is used, with reactions relating to Carbon Monoxide (centre) highlighted. Orange lines represent reactions producing CO whist the red (some of which may be hidden) are of reactions with CO.	81
3.17	A selection of edge shapes for the butane network. These show linear (a), quadratic (b) and bezier (c) edge shapes for the same network. In general, the bezier curves appear to provide the shapes of the most aesthetically pleasing graphs	84
3.18	How the compatibility threshold affects edge bundling using the Mercator graph from Subsubsection 3.2.2.2. In increasing the amount, edges are attracted (θ) it is possible to improve the clarity of a graph. However, there reaches a point where this distortion can worsen the result, confusing the reader, or creating a false positive. For this reason, I generally use only a slight bundling value $> 0.7.$	85
3.19	An example of confluent bundling. A traditional network (a), Edge bundling (b), Power Graph (c) and Confluent graph (d) representations. Source: [Bach, 2020] .	86
3.20	The routing graph of the butane mechanism. Here paths which contain two or more bundles have an extra 'routing' node introduced (orange stroke)	87
3.21	Confluent graph with crossing artifacts. The routing graph with the addition of basis-splines using the orange routing nodes in Figure 3.20 as control points.	88
3.22	Confluent graphs without crossing artifacts. The remaining confluent graph with crossing edges removed.	89
3.23	Film style representation of temporal changes in a network. Showing the temporal changes from a model simulation of the Beijing atmosphere. (a) shows a weighted graph at midnight. With the addition of daylight, the chemistry speeds up, causing the force graph to contract, changing the overall network shape (the faster reactions have a stronger attractive force). The animation of this can be found at https://github.com/wolfiex/DanEllisThesis/blob/master/daynight_26mb.gif	91
3.24	A 3D representation of a graph to hilight certain features. The first, second and third generation species of isoprene shown as an interactive 3D anaglyph.	92
3.25	A weighted and unweighted force diagram of the methane mechanism. Here it is seen that upon weighting, edges with a larger flux (pink) are drawn closer than those of a weaker one (blue).	94
3.26	A flow chart of the process performed by the custom gephi script used to generate the data set	95
3.27	A sample of 224 (out of the 2000) graphs generated using the ForceAtlas2 algorithm. These represent the conditions of a spun up simulation of Beijing at noon. The shapes of each graph, and general shapes are discussed in Subsubsection 3.5.2.1 and Subsubsection 3.5.2.2	06
		30

2	2	TTL
	ı	xxı

3.28	A normalised scatter plot of 2D space produced by the t-SNE algorithm. Each triangle represents a different arrangement of the MCM nodes shown in Fig- ure 3.27, and the colours/density contours show the regions in which we find similar images/graphs. Cluster numbers correspond to the groups in Figure 3.29	97
3.29	A selection of graphs for each of the labeled groubs in Figure 3.28. These reveal that symmetric similarity between like-positioned points within the t-SNE output.	98
3.30	Highlighting the groups of species, and their products within one of the MCM network graphs from Figure 3.27 These are Aromatics (gold), Terpenes (turquoise) and Alkane/Alkene carbon chains (red/blue) $\ldots \ldots \ldots \ldots \ldots \ldots$	99
4.1	The human family tree. This is a visual depiction of the human lineage, starting with our common ancestorial roots. In Chapter 2 it was shown that trees / graphs ¹ are useful in showing relationships between items. Source: [Wood, 2014] $\ldots \ldots \ldots$	112
4.2	150 years of letters to Nature. A visualisation showing how previous research is used to inspire future studies. Important discoveries (DNA, Cloning(frogs), Bio-Currents, Ozone Hole, Molecular Sieves and Exoplanets) are split into research which contributed to their formation (below), and the consequent papers produced from each discovery. Use of colour is used to emphasise the multi-disciplinary nature of prolific scientific discovery. Source: [Barabási, 2019]	115
4.3	Initial 3D graph representation of the scraped MCM citation graph. (a) shows the 'classic' graph representation of the network. (b) shows a size representation using an orthographic perspective. Here time is shown across the x axis, with yellow being the most recent. (c) uses a perspective camera, which emphasises the time component of the data. Still captures of 2D and 3D visualisations of the dataset. Node size corresponds to the number of citations, and colour (and z-axis) corresponds to the publication year for each paper.	117
4.4	The co-author network. In representing the authorship network as a force-directed graph, we see cliques or clusters of people who publish together. It is noted that this often occurs when they have a similar geographical location. Node sizes and colour represent author rankings using the PageRank algorithm (Subsection 4.3.5) \ldots .	120
4.5	Degree Centrality. In applying the degree centrality to the co-authorship network, it is possible to pick the authors with the greatest number of papers, of which the top 10 have been listed	122
4.6	Closeness centrality within the co-Author network. Here a colour/size gradient is seen, with the nodes that are more central (in location) and better connected having a higher closeness than those in the peripheries - which are harder to get to	124
4.7	Betweenness centrality within the co-Author network. Nodes which lie on a pivotal position (connecting/bottleneck) tend to have a high betweenness value due to their crutial role within the network. The colour represents the betweenness centrality	126
4.8	Page Rank centrality within the co-Author network . Node size and colour represent the ranking of each node from the page rank algorithm. Larger, lighter coloured nodes are more important.	130
4.9	How the MCM graph density scales with number of species. A figure showing that an increasing number of species within a mechanism subset results in an increased model sparsity (decreasing density).	132
4.10	A figure showing the small worldness for many Monte-Carlo selected MCM subsets. The network structure of these is then assessed using the omega coefficient, with [-1,0,1] corresponding to the perfect lattice, small-world and random network structure. Here Node size and colour represents the number of reactions in the mechanism subset and the number of primary VOCs (blue=small, green=large)	133
4.11	The different network structures. A visual depiction of the different graph structures. Source: Needham and Hodler [2019]	134

4.12	Comparing the MCM subsets against a power law, logarithmic and expo- nential distribution. The fit for different cumulative probability distributions of nodes in the MCM network is compared to determine the type of network hierarchy the chemistry follows. This is done by comparing the distance of the calculated distribution of data against a perfect one using the Kolmogorov-Smirnov test. The closer the two distributions, the lower the KS distance, and the better the fit.	135
4.13	A concentration (mixing ratios) time series from a simple methane-only simulation. This is the simplest method for identifying changes in species within a model simulation. This multi-plot shows the changes in concentration profiles for all initialised species (NOx:10ppb; $CH_4:20ppb$; $O_3:30ppb$) following an initial 3 day spin-up to steady state.	136
4.14	Rate of production and loss analysis plot for CH_3CO_3 exhibiting a net loss (daytime). An example ROPA plot from a simulation representing the chemistry within Beijing. This is used to identify the usefulness and weaknesses of using such a method. DUMMY represents the deposition term for any species.	138
4.15	A graphical representation of Equation 4.17 derrived from the Equation 4.10 \ldots .	143
4.16	Reversing the directions on negatively weighted edges from Figure 4.15	143
4.17	Simplifying Figure 4.16	144
4.18	The Mark 1 perceptron Both software and hardware are different manifestations of a flow chart. The perceptron hardware accomplished what is now done using software. Source: Cornell [2020]	145
4.19	The Human Cortex - A biological neural network. A vertical cross section of the human cortex between an adult (top) and 1.5 month old infant (bottom) showing a layer like structure with a change in depth (left to right). Source: Cajal [2020]	146
4.20	Cape Verde MLP predicted and observational data of Ozone (mixing ratio). Each segment represents data from a different month. Within each month segment exists 24 hour segments to create a diurnal. Observational concentrations are plotted in the form of a translucent scatterplot and summarised using the boxplot on the right of the month segment. MLP predicted results are shown using the solid lines. Concentration in mixing ratio.	149
4.21	Cape Verde MLP predicted and observational data of NO (mixing ratio). Each segment represents data from a different month. Within each month segment exists 24 hour segments to create a diurnal. Observational concentrations are plotted in the form of a translucent scatterplot and summarised using the boxplot on the right of the month segment. MLP predicted results are shown using the solid lines. Concentration in mixing ratio.	150
4.22	Cape Verde MLP predicted and observational data of NO ₂ (mixing ratio). Each segment represents data from a different month. Within each month segment exists 24 hour segments to create a diurnal. Observational concentrations are plotted in the form of a translucent scatterplot and summarised using the boxplot on the right of the month segment. MLP predicted results are shown using the solid lines. Concentration in mixing ratio.	151
4.23	Cape Verde MLP predicted and observational data of iso-Pentane. Each segment represents data from a different month. Within each month segment exists 24 hour segments to create a diurnal. Observational concentrations are plotted in the form of a translucent scatterplot and summarised using the boxplot on the right of the month segment. MLP predicted results are shown using the solid lines. Concentration in mixing ratio.	152
4.24	The mixing ratio profile for London. This shows a the change in mixing ratio over time for HO_x, NO_x , HCHO, Ozone and RO_2 species for a simulation run generated by the mlpregressor. Left of the dashed line shows the last 6 days of spinup, where the values are reset at noon each day until the species fractional difference is less than 0.001	.156
	v 1 i i i i i i i i i i i i i i i i i i	-

4.25	The mixing ratio profile for Beijing. This shows the change in mixing ratio over time for HO_x , NO_x , HCHO, Ozone and RO_2 species for a simulation run generated by the MLPRegressor. Left of the dashed line shows the last six days of spinup, where the initial values are reset at noon each day until the species fractional difference is less than $0.001 \dots $	157
4.26	The different IDF outputs. A plot showing Inverse Document Frequency profiles against Document Frequency. This shows that the probabilistic IDF highlights words that are more important across all items, while the smooth IDF shows files which are more important individually. The general IDF (which is used) produces a result starting at two and tending to zero. This provides the best response and can easily be scaled between the range of $[0,1]$ by dividing the output by 2. Source: [Mquantin, 2020]	160
4.27	The bivariate colourplot key.	161
4.28	An example force graph showing the complex chemistry of London. (Weightings not from initial conditions described.) Source: [Lewis, 2018]	163
4.29	An example force graph showing the complex chemistry of Beijing. (Weightings not from initial conditions described.)	164
4.30	A bivariate heatmap comparison of London.	165
4.31	A bivariate heatmap comparison of Beijing.	166
4.32	Link reversal of the Jacobian Sensitivity matrix graph results in a graph of the adjoint. Showing changing the direction of links in a graph is equivalent to applying the transpose to an adjacency matrix (right). In the case of a Jacobian based graph, this is analogous to using the adjoint to propagate the model back in time - something that can be used to identify the influence upon a species with a model	169
4.33	The reversed subgraph between α -pinene, and NC101CO. This is a subgraph showing the production of NC ₁₀₁ CO. Arrows point towards a species precursor	170
4.34	Showing total the influence from each species on HCHO for a sample MCM subset of Butane. Species importance (node size) is determined using the reverse PageRank algorithm starting at HCHO (middle). It is calculated by taking a snapshot of a chemical simulation and rendered using the transpose of the Jacobian relational matrix. Link width is representative of the cumulative sum of the weights contributing to a concentration change of HCHO.	173
5.1	The proposed plans for the change of the UK National Watersports Centre Whitewater Course (Holme Pierrepont). Walk based clustering is analogous to the movement of a river. Clusters (or modules) are identified as areas where the 'flow' becomes trapped, much like water in the pools immediately following a hydraulic jump. Source: [Cornes, 2008]	193
5.2	A Huffman tree, and the Huffman code generated from it. A Huffman tree is created using the frequency of occurrence of an item. The more often it appears in the source, the shorter the path to it. Source: [Sad CRUD Developer, 2016]	194
5.3	A graph of CRI v2.2 showing the hulls of the first level of hierarchical clustering. Nodes are coloured by the splits in branches, and the hulls enclose the nodes which lie within 95% of the (median) centre of the cluster.	196
5.4	Species structure within each cluster. A nested bubble chart is used to show the full hierarchical structure of the mechanism. This allows us to evaluate the species structure/type that has been extracted in each level of the hierarchical split. Node sizes are representative of the \log_{10} number of walkers that have become trapped by the flow algorithm at a location.	197
5.5	A radial treemap showing the hierarchical clustering of the CRI mechanism. The simulation results used are representative of the chemistry within London at Noon local time and generated using DSMACC and the InfoMap algorithm.	198

5.6	A radial tree of the InfoMap algorithm with a forced number of groups. Here a loss of hierarchical structure can be seen when compared to Figure 5.5. By setting a high number of required clusters, many species are grouped by themselves, which does not provide a useful output for mechanism lumping	201
5.7	Cosine distance against Euclidian Distance . Normalised version of the two distance metrics are plotted in an $x - y$ scatterplot. Each dot represents a different species pair plotted at the location of their value for each metrics. Species pairs with similar values and profiles have small values for each metric and are located towards the upper left hand corner. (b) shows the same results as (a) but has a no-overlapping (collision detection) algorithm applied show all points, and therefore aid interactive selection.	204
5.8	Gaussian Kernel Density Estimate plot showing the distributions present for the {0,1} scaled euclidean and cosine distances. This graph shows the density profiles for each metric in Figure 5.7a - these show 1-(normalised metric distance), and therefore the peak at 1 corresponds to that in the top left of Figure 5.7a. Peaks here correspond to the regions of high densitry within the scatterplot Figure 5.7b	205
5.9	Comparing the best (a-b) and worst (c-d) species combinations using the combined similarity metrics. Here species which only undergo a simple decay seem to be the easiest to group together. Species pairs between an photolytic and non photolytic species produce different profiles at differing magnitudes and are therefore difficult to match.	206
5.10	Filtering the infomap clustering relationship matrix/graph How the clustering relationship network changes as weak links (links between species which do not appear in many of the infomap groupings) are removed.	207
5.11	An alluvial diagram showing the changes in clusters between noon and mid- night. On the left are all groups that appear in $>45\%$ of the midnight simulation results. On the right are groups which appear $>45\%$ of the midday results. In the mid- dle exist the clusters extracted which appear in $>45\%$ of all runs. Here it is seen that there exist a series of species which may exist in the daytime or nighttime chemistry, but do not persist between both. Sizes represent the number of species and colours (greys) has no purpose other than to differentiate different items	209
5.12	Comparing the best (a-b) and worst (c-d) species pairs from Table 5.1 . Species which make a good candidate for reduction have a similar diurnal profile and production/loss patterns as well as ranges of magnitude in which the concentration lies. This is seen in subplots (a) and (b). Bad pairings either cover very different magnitude ranges (d) or have dice different temporal profiles (c and d). Time is in the format DD-MM HH.	210
6.1	The multifunctionality of the MCM. A chord diagram showing the functionalisa- tion of all species within the MCM v3.3.1. Arc sizes represent what percentage of all functional groups in the MCM mechanism a group contains. Translucent areas of no outwards links represent species with multiples of a certain functional group, of which Alcohols and Ketones have the most. Source: [Ellis, 2019]	223
6.2	Construction process of a SMILES string. The example compound is Melatonin. Although this does not exist within the atmosphere, it provides a clear example of the SMILES string methodology. Figure 6.2a is made using SMILES drawer: [Probst and Reymond, 2018]	225
6.3	Constructing a graph from species structure. (a) shows the maximum number of times an atom occurs for any single species in the MCM. (b) depicts the graph-like chemical structure of INB1NBCO3(a product from isoprene). This is a highly processed species stemming from Isoprene, and this makes for a good example of the bond matrix. Finally, a matrix representing the bonds in INB_1NBCO_3 is created from the maximum possible occurrence matrix from (a). For simplicity, empty row/column pairs have been removed to produce (c). This matrix will always be symmetrical as the bonds do not have a direction.	226

6.4	A graph of an MCM subset representing the chemistry within Beijing. Here colours show the increase of $O-C$ ratio as species are oxidised (lighter). All emitted species ultimately tend towards carbon monoxide, which is at the centre of the graph. Node clusters symbolise groups of species which react more between themselves and less with others (This graph only represents the mechanism structure)	227
6.5	Determining the Principal Compnent of a sample dataset. It can be seen that in a change in axis to follow the first principal component (right), it is possible to explain most of the variation in the samle dataset (left). Source: [Powell, 2020]	230
6.6	Representing the t-SNE algorithm as a fully connected force graph. Here each node is attached to every other node. Nodes with a strong relationship are pulled closer together than those with a weaker one	231
6.7	An example of how the curse of dimensionality affects the mapping of points a certian distance from eachother.	232
6.8	Showing the difference between PCA and t-SNE clustering. These figures show the clustering of a set of standardized species concentration profiles (c) across two styles of dimensionality reduction: PCA (a) and t-SNE (b)	235
6.9	An example autoencoder structure which reduces a 16 dimensional input to 2. Draw with the aid of [Krizhevsky et al., 2012]	236
6.10	Comparing the result of the 2D encoding and decoding of an Ozone-NOx-Methane isopleth. The original data (a) is reduced to two dimensions and then reconstructed back into 3D. This is done with Principal Component Analysis (b) and an AutoEncoder (c). The original isopleth is created using 300 simulations of different intial conditions: NOx (variable), Methane (variable) and Ozone (constant). These were designed using a latin hypercube and converted into a surface plot using Delaunay triangulation.	237
6.11	The process of converting a graph into a vector using Node2Vec. Source:[Cohen, 2018]	237
6.12	Calculation of the random walk path. Source: [Grover and Leskovec, 2019] $\ . \ . \ . \ .$	238
6.13	An example 4 colour matching This uses the first implementation of the algorithm mentioned in Subsection 6.4.4. The greedy approach does not often find the optimum solution, which may result in 5 colours instead. Observable Notebook : Daniel Ellis [2019]	241
6.14	A comparison of different clustering methods on a toy dataset. The plot shows the performance of several vector clustering algorithms in Scikit-Learn. Cluster algorithms are represented across the horizontal axis, and several types of datasets are across the vertical. Clustered groups are coloured. Source: [sklearn, 2019]	242
6.15	A decision tree aggregate from a random forest plotted with the Epiphyte version of the TreeSurgeon program [Ellis and Sherwen, 2019]. The data originates from Sherwen et al. [2019] and the importance of Temperature (blue), Depth (orange) and Chlorophyll a (green). It is shown that all models create their first split based on the temperature (is it >21 degrees). In the case it is (right branch) the sea depth is seen as the most important variable to test (is it deeper than 26m). This sort of split allows us to get a feel for which (if any) properties are dominant in partitioning the data.	244
6.16	Comparing clusters for all inputs after a reduction to 2 dimensions using Principal Component analysis. Each graph has undergone several clustering al- gorithms under a range of parameters. The result with the best silhouette coefficient has been chosen. Colours follow the greedy four colour theorem and are there only to indicate the contrast between cluster boundaries	246
6.17	Comparing clusters for all inputs after a reduction to 2 dimensions using an AutoEncoder. Each graph has undergone several clustering algorithms under a range of parameters. The result with the best silhouette coefficient has been chosen. Colours follow the greedy four colour theorem and are there only to indicate the contrast between cluster boundaries.	247

6.18	Comparing clusters for all inputs after a reduction to 2 dimensions using t-SNE. Each graph has undergone several clustering algorithms under a range of parameters. The result with the best silhouette coefficient has been chosen. Colours follow the greedy four colour theorem and are there only to indicate the contrast between cluster boundaries.	248
6.19	The individual decomposition of clusters in Figure 6.16(c) This shows that the main difference between the two clusters is the existence of Nitrogen elements within Nitrate and Peroxyacetyl Nitrate (PAN) groups. The table on the right acts as a key for the colours and shows the overall importance of each feature in separating an item into the various clusters (using an ensemble of decision trees).	250
6.20	Comparing feature importance for PCA clusters. Importance ranges are trimmed at 40% for comparison. Some categories may contain values greater than this. All bars sum to 100%.	253
6.21	Comparing feature importance for AE clusters. Importance ranges are trimmed at 40% for comparison. Some categories may contain values greater than this. All bars sum to 100%.	254
6.22	Comparing feature importance for t-SNE clusters. Importance ranges are trimmed at 40% for comparison. Some categories may contain values greater than this. All bars sum to 100%	255
6.23	Comparing individual clusters between MACCS for PCA and t-SNE algo- rithm output. The bar chart to the right is the cumelative chart which represents the splits in deciding the cluster a species falls into from Subsection 6.6.3. Unlabeled bar charts to the left represent the partitioning of species within an individual cluster.	257
6.24	Comparing individual clusters between Node2Vec for PCA and t-SNE al- gorithm output. The bar chart to the right is the cumulative chart which represents the splits in deciding the cluster a species falls into from Subsection 6.6.3. Unlabeled bar charts to the left represent the partitioning of species within an individual cluster.	258
6.25	Case Study 1: PCA graph-fingerprint We compare the functional group distribu- tion for individual clusters within the PCA 2D representation of the graph-fingerprint input.	259
6.26	Case Study 1: AE SMILES We compare the functional group distribution for individual clusters within the AE 2D representation of the SMILES input	260
6.27	Case Study 1: t-SNE MQN. We compare the functional group distribution for individual clusters within the t-SNE 2D representation of the Mollecular Quantum Number fingerprint.	261
B.1	Binary Step activation function.	282
B.2	Linear activation function.	283
B.3	Sigmoid activation function.	283
B.4	Tanh activation function.	284
B.5	ReLU activation function.	284
B.6	Swish activation function.	284
C.1	The labelled co-author network - as referenced in Chapter 4	286

List of Tables

4.1	A selection of research papers not directly connected to the field of atmospheric modelling.	118
4.2	In-Degree of the citation network: The top 3 most cited papers	123
4.3	Out-Degree of the citation network: The top 3 most citing papers	123
4.4	(2-page split) The initial conditions created from the MLPRegressor prediction of ob- servational data. Although not specified the mixing ratios for methane is set by the model at 1770ppb, the temperature is 298K, and water vapour is at 2%. * Starred values are of the wrong units and should be multiplied by 1000. As there was no time to rerun these, their results have been omitted from this chapter.	155
4.5	A table of the top 10 ranked species for each simulation. Only species that exist within at least 3 out of the four simulations are used. The Nan-Mean takes the mean of all available data, ignoring runs where a species is not present. Species presented within the table follow the MCM naming convention.	167
4.6	A reversed graph Page Rank test with $NC_{101}CO$	171
4.7	The net flux of the three species: NAPINBOOH, NAPINBO, NAPINBO ₂ , taken from a simulation of Borneo at 2020-06-24 12:09:56	171
4.8	The influence on $NC_{101}CO$ from other species, taken from a simulation of Borneo at 2020-06-24 12:09:56	171
5.1	A table of the normalised similarity values for the lumped species. Numbers closest to 1 show the worst possible paring in the mechanism, and numbers approaching 0 show the best.	210
6.1	A set of regular expressions that are used to determine the number of occurrences of a functional group within a SMILES string. These were written to scan the SMILES string a match specific patterns corresponding to each functional group. A similar process is used within [rdkit, 2019] to construct MACCS keys (discussed later)	222
6.2	The inputs to the PCA dimensionality reduction algorithm sorted by the best obtained silhoette coefficient.	251
6.3	The inputs to the AutoEncoder dimensionality reduction algorithm sorted by the best obtained silhoette coefficient.	251
6.4	The inputs to the t-SNE dimensionality reduction algorithm sorted by the best obtained silhoette coefficient.	251

Chapter 1

Introduction

"In the beginning the Universe was created. This has made a lot of people very angry and been widely regarded as a bad move"

- Douglas Adams, The Restaurant at the End of the Universe

1.1 Background

1.1.1 A Preface On Humanity And The Climate

The development of humanity is not unlike the chirography of an Aristotelian tragedy. It starts with a simple/primitive species cradling a noble cause - to improve their chances of survival. Here the protagonist (humankind) develops a fatal flaw: insecurity and latent destruction of their home due to a sudden rise to power. Having acknowledged this flaw, we now strive to improve our understanding of the universe, correct past mistakes and stem the tide of inevitable change.

With tragedy being an imitation not of humanity, but of action and life, happiness and misery, it is only expected that such a comparison to our current affairs should stir feelings of catharsis when exploring our need for research and scientific advancement. It is with that I begin this thesis with the beginning of the planet, its atmosphere and consequently the start of humankind.

1.1.2 Formation Of The Atmosphere

4.5 billion years ago the Earth began as a disk of dust and gas orbiting our sun. The movement of such gasses produces a resonant drag instability, which causes them to clump together [Hopkins and Squire, 2018; Woo, 2018]. As these 'clumps' become denser, other forces come in to play and further increase their size. These eventually produced the hot mix of gas and solid, which was to become Earth. As the Earth cooled, the volatile components of the primordial gas cloud surrounding it begins to form an atmosphere. At this point, oxygen was not only absent in the atmosphere but also had many sinks within the Earths anoxidised crust. It was not until oxygenic photosynthesis ([Peretó, 2011]) that the concentrations of oxygen in the atmosphere started to increase. Eventually the development of multicellular cyanobacteria¹ resulted in biologically induced oxygen accumulating in the atmosphere, [University of Zurich, 2013]. This led to the most significant climate event in the history of the planet: the Great Oxygenation Event (2.5 billion years ago), [Planavsky et al., 2014]. This increase in oxygen allowed organisms to become larger and more active, eventually resulting in the human race.

1.1.3 Rise Of The Homo Sapiens ('Wise Man')

2-6 million years ago there were many varieties of the 'homo' genus (Figure 4.1,Wood [2014]). 70,000 years ago homo sapiens came into existence and started the cognitive revolution. Here again in brain

 $^{^{1}}$ The phylum of photosynthetic prokaryotic (cells not containing a distinct nucleus) bacteria - e.g. blue-green algae

size increased communication, tool development and analysis capabilities. However, the evolutionary brain enlargement required an increase in net energy intake [Navarrete et al., 2011] (the brain makes up for 2-3% of human body mass but consumes 25% of the body's energy at rest [Harari, 2015]).

A change of diet [Aiello and Wheeler, 1995] soon addressed this energy imbalance, provisioning and sharing (cooperative breeding) and tool-assisted processing such as cooking [Wrangham, 2009] - the first known case of anthropogenic indoor air pollution. The increase of cerebral power eventually led to the agricultural revolution² (12,000 years ago) and the scientific revolution³ (500 years ago), [Harari, 2015].

Air pollution and climate have always been a concern for the human race. Such disquietude was first documented 6000 years ago with the ancient greeks (lead in the air) [Lomborg et al., 2001] and the Romans (Rome was reported to have a 'stink of soot and heavy air') [Miller, 2010]. In 1285 the smell of burning jet⁴ drove the Queen of England to leave Nottingham and 22 years later King Edward released the first air pollution act [Brimblecombe, 1977]. In the 18th century, the United Kingdom entered the Industrial age, here combustion was used to power machines and replace hand tools with mechanical ones. With this started the age of technology and automation - a process requiring energy, and thus increasing emissions to the atmosphere. In the present day, technology is ever increasing in efficiency - however, the rate of this is not yet sufficient to mitigate any damage already caused.

1.2 Motivation (How The Atmosphere Affects Us)

The atmosphere constitutes an integral part of the Earth system. It is responsible for shielding the planetary surface from harmful radiation; allowing the transport of energy (weather and climate forcing), and interacting with the biosphere. This section explores the many roles of the atmosphere, and consequently, the interests and motivation of climate and atmospheric science. We start with the composition of the atmosphere and air quality (Subsection 1.2.1), and then relate this to the different roles of ozone (Subsection 1.2.2), concluding on changing climate and radiative forcing, for with OH plays a vital role (Subsection 1.2.3).

²Domestication of plants and animals.

³humankind admits ignorance and gain unprecedented control

 $^{^4\}mathrm{The}$ lowest rank of coal and very common at the time.

1.2.1 Air Quality - It Is The Air We Breathe

The atmosphere consists mainly of nitrogen (N_2) and oxygen $(O_2)^5$, in addition to a vast range of other species [Pryor et al., 2015]. Human beings rely on oxygen to convert sugars and fatty acids into energy. The procurement of this lies through the breathing of the air surrounding us - the composition of which can have dire effects on our respiration system. Pollutants such as particulate matter (PM), ozone (O_3) , nitrogen dioxide (NO_2) and sulphur (SO_2) dioxide can cause respiratory problems, heart disease, strokes, cancer and chronic obstructive pulmonary disease WHO [2018]. Over 80% of people who live in urban environmets⁶ are exposed to poor air quality levels exceeding the recommended limits by World Health Organisation, air quality poses a significant risk to human life - It is estimated that 4.2 million premature deaths globally are linked to ambient air pollution⁷ (Figure 1.1).



Figure 1.1: Reported deaths attributed to air pollution by country (2016) A cartogram (and cloropleth) showing the number of premature deaths attributed to ambient air pollution per 100,000. The colour bar range is from 9 in Canada (light blue) to 170 in eastern Europe (navy) people. Data Source: [WHO, 2016]

⁵These form 99% of its dry-air total mass

⁶Which measure the levels of air pollution.

 $^{^7\}mathrm{A}$ similar number can also be attributed to indoor air pollution - which also falls under the umbrella term of Air-Quality.
1.2.2 Stratospheric Ozone - The Protective Barrier

Ozone plays a vital role in the stratosphere. This was seen in the 1980s where the use of Cloro Fluro Carbon (CFC) aerosols resulted in the thinning of the atmospheric ozone [Farman et al., 1985]⁸. This resulted in an increase in UV-B radiation, and in consequence skin cancers, immune suppression and disorders of the eye [Bais et al., 2018]. Due to this, the Montreal Protocol on Substances that Deplete the Ozone Layer was put into place to reduce the adverse effects experienced by humans and the Earths surface [UNEP, 1987]. As part of this, CFCs are still being phased out resulting in a gradual decrease in the damage of the ozone hole.

1.2.3 Changing Climate

Over the last 30 years, a large body of scientists has established that humans have a warming effect on the planet [Houghton et al., 1996b,a; IPCC, 2007, 2013; IPBES, 2019]. Here it has been shown that changes in temperature can lead to the melting of glaciers, rise of sea levels, extreme weather events and the extinction of many species.

The impact of this resulted in the nations of the world to develop a 'legally binding' agreement to combat climate change (the Paris agreement) [Phillips, 2015; UNFCCC, 2015]. However, a recent analysis on the current state of the world outlines that the failure to set, implement and without adequate targets and procedures within the last decade mean that we now need to do 'four times the work, or [do it in] one-third of the time' [Höhne et al., 2020].

1.3 Tropospheric Chemistry

The lowest part of the atmosphere (<18km)⁹ is called the troposphere. This contains 75% of the mass of the atmosphere, and comes from the greek $\tau \rho o \pi o \varsigma$ which means 'way' or 'turn towards change'. This describes the turbulent mixing that happens due to friction in the lower 2km of the atmosphere (the boundary layer). As the troposphere is the closest part of the ground, this where most of the complex chemistry which affects us at the surface happens. This section describes the underlying chemical processes which exist in the atmosphere.

 $^{^{8}\}mathrm{Here}$ the chlorine attacks the double bond and 'steals' an oxygen atom from the O_{3} molecule.

⁹18km at the tropics, 17km in the mid-latitudes and 6km at the poles.

1.3.1 Ozone Production/Loss

In the troposphere, the mixing ratio of ozone is controlled by the photostationary state relationship (Equation 1.1-1.3). Since the concentrations of ozone $(20-60 \text{ ppbv})^{10}$ are often much higher than that of the nitrogen oxides, NO (1-60 pptv)¹¹ and NO₂ (5-70 pptv), the rapid rate of reaction between Equation 1.1 does lead to a net change in O₃ concentration ¹² [Jacobson, 2005].

$$NO + O_3 \xrightarrow{k_1} NO_2 + O_2 \tag{1.1}$$

$$NO_2 \xrightarrow{hv(J)} NO + O \quad (\lambda < 420nm)$$
(1.2)

$$O + O_2 + M \xrightarrow{k_2} O_3 + M \tag{1.3}$$

Using Equation 1.1 and Equation 1.2 it is possible to describe the change in NO_2 as:

$$\frac{d[NO_2]}{dt} = k1[NO][O_3] - J[NO_2]$$
(1.4)

If the relative change of NO_2 is small, it can be thought of as being in a steady-state. This means that Equation 1.4 can be simplified to produced a relationship between O_3 , NO and NO_2 in steady-state (Equation 1.5). Here if any two concentrations are known, the third can be calculated.

$$[O_3] = \frac{J[NO_2]}{k1[NO]}$$
(1.5)

As ozone is a secondary pollutant (made not emitted), and its primary reaction produces the null cycle, the production of ozone in the atmosphere requires an increase in nitrogen dioxide concentrations.

1.3.2 The NOx Cycle

Ozone production/loss in the troposphere is directly dependant on the concentration of available Nitrogen Oxides (NOx) (Subsection 1.3.1). These are predominantly emitted by motor vehicles and power stations and can are known to cause respiratory problems in children and asthmatics as well as disrupting terrestrial and aquatic ecosystems [EEA, 2018]. Although NOx may be released naturally, the anthropogenic influence on their emissions was highlighted in early 2020 where the COVID-

¹⁰ppbv: parts per billion volume

¹¹pptv: parts per trillion volume

 $^{^{12}}$ In urban areas NO concentrations may rise to be be greater than those of O₃ during the night. This leads to a decrease in from Equation 1.1



19 coronavirus disrupted travel across mainland China, causing a significant drop in anthropogenic emissions - Figure 1.2.

Figure 1.2: Changes in NOx concentrations due to anthropogenic emissions. A reduction in activity and trasport produces a notable decrease of Nitrogen dioxide concentrations in the troposphere. Source: [Stevens, 2020]

During the day nitrate (NO_3) radicals can be formed through the reaction with O_3 : Equation 1.4 and Equation 1.6, however this is quickly destroyed through rapid photolysis (Equation 1.7) [Ng et al., 2017]. At night photolysis reactions such as Equation 1.7 and Equation 1.2 are no longer possible and the ozone production process shuts down.

$$NO_2 + O_3 \xrightarrow{k_3} NO_3 + O_2$$
 (1.6)

$$NO_3 \xrightarrow{hv} NO_2 + O({}^3P) \tag{1.7}$$

The increased amount of NO₃ can now react with NO₂ to produce dinitric pentoxide (N₂O₅) and (in solution) nitric acid (HNO₃) - Equation 1.8 and Equation 1.9. Equation 1.8 is a three-body forwards pressure dependant reaction and a reverse temperature dependant reaction. During the day at the lower troposphere, it is warm, and the reverse reaction can occur within seconds, however, at night or high altitudes it can take anywhere from hours to months [Jacobson, 2005].

$$NO_3 + NO_2 \rightleftharpoons N_2O_5 + O$$
 (1.8)

$$N_2O_5 + H_2O \xrightarrow{k_4} 2 HNO_3$$
(1.9)

1.3.3 HOx Cycle

The hydroxyl (OH) radical is central to tropospheric chemistry and a major sink for many of the greenhouse gasses (including ozone - see Equation 1.12) [Olson et al., 1997]. Its primary source of production is through the action of UV in sunlight to photolyse ozone [Jacobson, 2005]:

$$O_3 \xrightarrow{hv} O^1 D$$
 (1.10)

$$O^1 D + H_2 O \longrightarrow 2 OH$$
 (1.11)

$$OH + O_3 \longrightarrow HO_2 + O_2$$
 (1.12)

As OH is highly reactive, with a lifetime of < 1 seconds - Figure 1.3, it is not transported a long distance and only exists during daytime (when it is still being produced). In reacting with a VOC, the hydroxyl radical scavenges hydrogen to form a radical species and water (H₂O). This produced radical species can then move on to react with O₂ to produce a RO₂ species Equation 1.14.

$$OH + RH \longrightarrow R. + H_2O$$
 (1.13)

$$R. + O_2 \longrightarrow RO_2 \tag{1.14}$$

$$\operatorname{RO}_2 + NO \longrightarrow RO + \operatorname{NO}_2$$
 (1.15)

The created RO_2 can then convert NO to NO_2 producing an RO (Equation 1.16) which also does the same via the hydroperoxide radical HO₂ (Equation 1.17 - 1.18). This NOx conversion is able to drive Ozone formation in the conventional way: Equation 1.1-1.3.

$$\mathrm{RO}_2 + NO \longrightarrow RO + \mathrm{NO}_2$$
 (1.16)

$$RO + O_2 \longrightarrow HO_2$$
 (1.17)

$$HO_2 + NO \longrightarrow NO_2 + OH$$
 (1.18)

1.3.3.1 The Hydroperoxide Radical

Unlike OH, HO₂ can exist both during daytime and night. It can further react with ozone to reproduce the hydroxyl radical and create two O_2 molecules - Equation 1.19.

$$HO_2 + O_3 \longrightarrow OH + O_2 + O_2$$
 (1.19)

The loss of ozone loss depends on the NO mixing ratio, where if NO > 10 pptv, HO₂ will react predominantly with the NOx species. At lower concentrations (3-10 pptv) HO₂ reacts mainly with ozone, and at deficient concentrations, it reacts mostly with itself [Finlayson-Pitts and Pitts, 2000]. Combined OH and HO₂ form the HOx species and the cycle in Figure 1.4.



Figure 1.3: Spatial and temporal scales of variability of atmospheric species. This shows that the longer lived a species, the further it is likely transported through the atmosphere. Source: [Seinfeld and Pandis, 2016]



Figure 1.4: The HOx cycle. The OH aids in the oxidation of VOCs, which makes them more water soluble - this allowing for their removal from the atmosphere. In a high NOx evironment the RO_2 radicals can then reactive with NO to produce NO_2 , and consequently more ozone.

1.4 Modelling The Earth

In the previous section, the air quality and its detrimental effects on human health were seen to influence policy for cities and industry. For a policy to be passed there needs to not only evidence of the problem but a strong suggestion that any proposed changes will have the desired effect. As it is not possible to perform experiments on complex, and often unknown, chemistry at every location on the planet, we are forced to rely on the numerical simulation of the Earth System, and the constituent parts within it.

1.4.1 Earth System Models (ESM)

ESMs are models capable of predicting past or future interactions of the planetary system. They represent our foremost understanding of the complex interplay between land-surface (geosphere), ocean (hydrosphere), ice (cryosphere) and the air (atmosphere), and act as a surrogate to manual experimentation - which is just not possible on the global scale. ESMs can be split into individual parts. One example of this is the Chemistry section of the Goddard Earth Observing System (an integrated ESM and data assimilation model hosted by NASA's Goddard space flight centre [Community, 2020]) - GEOS Chem. GEOS-Chem is a global 3D model of atmospheric chemistry which is driven by the meteorology provided by NASA [GEOS-Chem, 2020]. Here the Earth is split up into cubic cells



Figure 1.5: A diagram showing the longitudinal, lateral and vertical decomposition of a **3D global model.** (Diagram not of GeosChem.) Source: [Henderson-Sellers, 2015]

longitudinally, latitudinally, and vertically (Figure 1.5)¹³. Each one of these cells performs several perturbations of the chemistry within them before any long-lived species are transported, and the process is repeated. If extracted separately, a single one of these cells may be used to explore the sensitivity of different species for a range of input conditions. This is the bases of the atmospheric box model.

1.5 The 0D Chemical Box Model

In exploring the sensitivities of individual species within a simulation, it is possible to use a zerodimensional box model. This is, in essence, a single cell within the global structure which has been constrained in location and height (pressure). A box model allows for better in-depth analysis of the chemistry within a model, without any of the overhead of having to run it for the entire planet. Such studies make zero-dimensional models perfectly suited for studying the sensitivity of chemical schemes under a range of conditions, for example, [Emmerson and Evans, 2009].

In general, a box model consists of two main parts - a mathematical representation of the reactions in the atmosphere and the rate they occur (this is known as a mechanism); and a method to propagate this chemistry forwards in time (the integrator).

¹³This image is not from GEOS-Chem.

1.5.1 Chemical Mechanisms

Mechanisms are at the heart of every chemistry simulation. They are a mathematical representation of the possible reactions (and the rates at which these may occur) which describe the evolution of the atmosphere within a numerical model. Different models contain varying levels of chemical complexity depending on their foci. However, there is a need for a 'gold standard' or 'benchmark mechanism' which contains a comprehensive representation of the current 'state of the science'. For the last decade, the benchmarking mechanism for both the UK and internationally has been the Master Chemical Mechanism [Rickard, 2020] (this shall be described throughout the rest of this thesis).

1.5.2 Numerical Integration

Using a mechanism it is possible to determine how quickly each species in a reaction is changing at a certain set of conditions using its a slope or derivative. In this way, integration allows us to find the change in concentration over time as well as the rate at which this is happening. Taking the reaction of N_2O_5 (Equation 1.20), we can write the rate of change for each species over time (Equation 1.21)¹⁴. In integrating this equation, we can calculate the actual change in concentration (Equation 1.22) - this is the foundation of atmospheric models.

$$N_2O_5 \longrightarrow NO_2 + NO_3$$
 (1.20)

$$-d[N_2O_5]/dt = d[NO_2]/dt + d[NO_3]/dt$$
(1.21)

$$-\int d[N_2O_5]/dt = \int d[NO_2]/dt + \int d[NO_3]/dt$$
(1.22)

1.5.2.1 Non-Stiff Equations

Non-stiff ordinary differential equations are made up components which all evolve at similar timescales. They can easuly be solved by explicit (calculate the next step using the current time only) numerical methods which can be solved with the forward Euler¹⁵, Runge Kutta ¹⁶ or Verlet methods¹⁷ [C.J.Budd, 2019; Wild, 2015].

¹⁴This is also known as the flux.

¹⁵Euler:Evaluate gradient at the time point, move forward in time and repeat.

 $^{{}^{16}}$ **RK4:**The next value is determined by the present value combigned with the weighted avarage of 4 increments across an interval h.

 $^{^{17}}$ Verlet: A central difference method to find the gradient using the avarage of a forwards and backwards timestep.

1.5.2.2 Numerically Stiff Equations (Atmospheric Chemistry)

Unfortunately, chemical lifetimes within the atmosphere range several orders of magnitude (Figure 1.3). This creates a numerically stiff system that requires the step size to be chosen based on the most rapid component, which can lead to inefficiency in the computation of the entire system. Solvers for stiff equations usually create and evaluate the Jacobian matrix (a matrix of second-order partial derivatives describing the effect species have on each other).

Standard solvers include Backward Differentiation Formulas (known as the 'gear' method and implemented as the LSODE solver in [Sandu and Sander, 2006]), implicit Runge Kutta and the Rosenbrock methods. Gear methods are multistep methods with high stability and have been applied to a range of atmospheric chemistry models [Hairer et al., 2002; GEOS-Chem, 2020; Jacobson, 2005]. The Rosenbrock method can be likened to a linearly-implicit Runge Kutta method which uses the Jacobian matrix directly within the integration formula (solved at each stage) to avoid solving for a non-linear system [Zhang et al., 2011]. This provides an efficient solver with modest accuracy (less than 10^{-5}) which is more than suitable for use within atmospheric chemistry calculations [Zhang et al., 2011; Sandu et al., 1997].

1.5.3 The Model Development Cycle

Scientific understanding is the product of many cycles of trial and error, Figure 1.6. In atmospheric chemistry, we start with a hypothesis or a question, e.g. will change X has a negative response on Y. We then construct a theoretical model to represent the chemistry within. This chemistry is updated to reflect the rates and reactions that have been recorded in laboratory/chamber experiments. This cycle is then repeated until the model, and real-world observations produce a comparable result. It is essential to point out that improving the predictive capabilities of a model are iterative tasks which both feedback and respond to changes in our understanding of atmospheric reactions.



Figure 1.6: The scientific development cycle. This shows the iterative nature between modelling, observation and laboratory experimentation

1.5.4 The Dynamically Simple Model Of Atmospheric Chemical Complexity

Within this thesis, the Dynamically Simple Model of Atmospheric Chemical Complexity (DSMACC) was used to run model simulations. This a simple box model designed for the comparison of a range of gas-phase chemical schemes under different conditions [Emmerson and Evans, 2009]. The DSMACC model uses the Kinetic PreProcessor (KPP) to convert a chemical mechanism into the set of ordinary differential equations which can be solved using a suite of FORTRAN numerical integrators it provides [Sandu and Sander, 2006]. The Tropospheric and Ultraviolet (TUV) model from Bräuer [2020] is used to calculate the strengths of different photolysis reactions for the mechanism. These are determined at the start of a simulation and then predicted using cubic splines [Ellis, 2020]. This is the model setup that will be used to propagate the chemistry forwards in time using the Rosebrock integrator.

1.6 Thesis Layout

This thesis will explore a series of methods for describing and understanding the complex chemistry which may exist as part of an atmospheric chemistry mechanism. The mechanism used is a nearexplicit representation of our foremost understanding of how gas-phase chemistry in the troposphere reacts - the Master Chemical Mechanism, [Rickard, 2020]. We begin by exploring the use of visualisation to convey complex scientific data (Chapter 2). Next, we apply this to the representation of species in a mechanism and the relationships between them. It is found that the node-link style graph format is the most beneficial, the use of which is then explored further (Chapter 3). However, in doing so, sizeable complex networks are shown to reach the limits of human cognition and visual representation. A series of mathematic metrics are used to leverage our understanding of the species in a chemical network using graph theory (Chapter 4). The use of computation to aid in graph analysis is further extended when graph clustering methods are applied as a method to similar group species within a chemical network (Chapter 5). Finally in a bid towards the use of neural graph networks (see future work, Section 7.4), a range of different chemical representations for machine learning are explored using several dimensionality reduction algorithms (Chapter 6).

Bibliography

- Aiello, L. C. and Wheeler, P. (1995). The expensive-tissue hypothesis: The brain and the digestive system in human and primate evolution. *Current anthropology*, 36(2):199–221.
- Bais, A. F., Lucas, R. M., Bornman, J. F., Williamson, C. E., Sulzberger, B., Austin, A. T., Wilson,
 S. R., Andrady, A. L., Bernhard, G., and McKenzie (2018). Environmental Effects Of Ozone
 Depletion, Uv Radiation And Interactions With Climate Change: Unep Environmental Effects
 Assessment Panel, Update 2017. Photochemical & photobiological sciences: Official journal of the
 European Photochemistry Association and the European Society for Photobiology, 17(2):127–179.
 http://dx.doi.org/10.1039/c7pp90043k.
- Bräuer, P. (2020). TUV 5.2X DSMACC. https://github.com/pb866/TUV DSMACC.
- Brimblecombe, P. (1977). London air pollution, 1500–1900. Atmospheric Environment (1967), 11(12):1157 1162. http://www.sciencedirect.com/science/article/pii/0004698177900919.
- C.J.Budd (2019). Advanced Numerical Methods Lectures Part 2 . *online*. https://people.bath.ac. uk/mamamf/chapt6and7.pdf.
- Community, T. I. G.-C. (2020). Geoschem/geos-chem: Geos-chem 12.7.1. Zenodo. https://doi.org/ 10.5281/zenodo.3676008.
- EEA (2018). Air Quality In Europe 2018. https://www.eea.europa.eu/publications/ air-quality-in-europe-2018.
- Ellis, D. (2020). DSMACC-Testing. https://github.com/wolfiex/DSMACC-testing.
- Emmerson, K. M. and Evans, M. J. (2009). Comparison of tropospheric gas-phase chemistry schemes for use within global models. *Atmospheric Chemistry and Physics*, 9(5):1831–1845. https://www. atmos-chem-phys.net/9/1831/2009/.
- Farman, J. C., Gardiner, B. G., and Shanklin, J. D. (1985). Large Losses Of Total Ozone In Antarctica Reveal Seasonal Clox/Nox Interaction. *Nature*, 315(6016):207–210. https://doi.org/10.1038/ 315207a0.
- Finlayson-Pitts, B. J. and Pitts, J. N. (2000). Chemistry of the upper and lower atmosphere. Academic Press, San Diego. http://www.sciencedirect.com/science/article/pii/B9780122570605500034.
- GEOS-Chem (2020). Geos-Chem Publications. *online*. http://acmg.seas.harvard.edu/geos/geos_pub. html.

- Hairer, E., Nørsett, S. P., and Wanner, G. (2002). Solving Ordinary Differential Equations I (2Nd Revised. Ed.): Nonstiff Problems. Springer-Verlag, Berlin, Heidelberg.
- Harari, Y. (2015). Sapiens: A Brief History Of Humankind. Harper. https://books.google.co.uk/ books?id=FmyBAwAAQBAJ.
- Henderson-Sellers (2015). Climate Data Services | Nasa Center For Climate Simulation. https://www.nccs.nasa.gov/services/climate-data-services.
- Höhne, N., den Elzen, M., Rogelj, J., Metz, B., Fransen, T., Kuramochi, T., Olhoff, A., Alcamo, J., Winkler, H., Fu, S., Schaeffer, M., Schaeffer, R., Peters, G. P., Maxwell, S., and Dubash, N. K. (2020). Emissions: World Has Four Times The Work Or One-Third Of The Time. *Nature*, 579(7797):25–28. http://dx.doi.org/10.1038/d41586-020-00571-x.
- Hopkins, P. F. and Squire, J. (2018). The Resonant Drag Instability (Rdi): Acoustic Modes. Monthly notices of the Royal Astronomical Society, 480(2):2813–2838. https://academic.oup.com/mnras/ article-pdf/480/2/2813/25498305/sty1982.pdf.
- Houghton, J., Filho, L. M., Callander, B., Harris, N., Kattenberg, A., and Maskell, K. (1996a). Climate Change 1995 The Science Of Climate Change. The Intergovernmental Panel on Climate Change.
- Houghton, J., Jenkins, G., and Ephraums, J. (1996b). Climate Change 1990 The Science Of Climate Change. The Intergovernmental Panel on Climate Change.
- IPBES (2019). Global Assessment Report On Biodiversity And Ecosystem Services | The Intergovernmental Science-Policy Platform On Biodiversity And Ecosystem Services. https://ipbes.net/ global-assessment.
- IPCC (2007). Fourth Assessment Report: Climate Change 2007: The Ar4 Synthesis Report. Geneva: IPCC. http://www.ipcc.ch/ipccreports/ar4-wg1.htm.
- IPCC (2013). Climate Change 2013: The Physical Science Basis. Contribution Of Working Group I To The Fifth Assessment Report Of The Intergovernmental Panel On Climate Change. Cambridge University Press, Cambridge, United Kingdom and New York, NY, USA. www.climatechange2013. org.
- Jacobson, M. (2005). Fundamentals Of Atmospheric Modelling. Cambridge University Press. https://www.cambridge.org/core/books/fundamentals-of-atmospheric-modeling/ A6B866737D682B17EE46F8449F76FB2C.
- Lomborg, B., Matthews, M., and of Cambridge (Gran Bretaña), U. (2001). The Skeptical Environmentalist: Measuring The Real State Of The World. Cambridge University Press. https://books.google.co.uk/books?id=JuLko8USApwC.

- Miller, B. (2010). Clean Coal Engineering Technology. Elsevier Science. https://books.google.co.uk/ books?id=b2W5S3Lb4fwC.
- Navarrete, A., van Schaik, C. P., and Isler, K. (2011). Energetics And The Evolution Of Human Brain Size. Nature, 480(7375):91–93. http://dx.doi.org/10.1038/nature10629.
- Ng, N. L., Brown, S. S., Archibald, A. T., Atlas, E., Cohen, R. C., Crowley, J. N., Day, D. A., Donahue, N. M., Fry, J. L., Fuchs, H., Griffin, R. J., Guzman, M. I., Herrmann, H., Hodzic, A., Iinuma, Y., Jimenez, J. L., Kiendler-Scharr, A., Lee, B. H., Luecken, D. J., Mao, J., McLaren, R., Mutzel, A., Osthoff, H. D., Ouyang, B., Picquet-Varrault, B., Platt, U., Pye, H. O. T., Rudich, Y., Schwantes, R. H., Shiraiwa, M., Stutz, J., Thornton, J. A., Tilgner, A., Williams, B. J., and Zaveri, R. A. (2017). Nitrate radicals and biogenic volatile organic compounds: Oxidation, mechanisms, and organic aerosol. Atmospheric Chemistry and Physics, 17(3):2103–2162. https://www.atmos-chem-phys. net/17/2103/2017/.
- Olson, J., Prather, M., Berntsen, T., Carmichael, G., Chatfield, R., Connell, P., Derwent, R., Horowitz, L., Jin, S., Kanakidou, M., Kasibhatla, P., Kotamarthi, R., Kuhn, M., Law, K., Penner, J., Perliski, L., Sillman, S., Stordal, F., Thompson, A., and Wild, O. (1997). Results From The Intergovernmental Panel On Climatic Change Photochemical Model Intercomparison (Photocomp). Journal of geophysical research, 102(D5):5979–5991. http://doi.wiley.com/10.1029/96JD03380.
- Peretó, J. (2011). Oxygenic Photosynthesis, pages 1209–1209. Springer Berlin Heidelberg, Berlin, Heidelberg. https://doi.org/10.1007/978-3-642-11274-4 1721.
- Phillips, S. (2015). Cheers As World Adopts Historic Paris Climate Deal. ABC News. https://www. abc.net.au/news/2015-12-12/world-adopts-climate-deal-at-paris-talks/7023712.
- Planavsky, N. J., Asael, D., Hofmann, A., Reinhard, C. T., Lalonde, S. V., Knudsen, A., Wang, X., Ossa Ossa, F., Pecoits, E., Smith, A. J. B., Beukes, N. J., Bekker, A., Johnson, T. M., Konhauser, K. O., Lyons, T. W., and Rouxel, O. J. (2014). Evidence For Oxygenic Photosynthesis Half A Billion Years Before The Great Oxidation Event. *Nature geoscience*, 7(4):283–286. https://doi.org/ 10.1038/ngeo2122.
- Pryor, S., Crippa, P., and Sullivan, R. (2015). Atmospheric chemistry. In Reference Module in Earth Systems and Environmental Sciences. Elsevier. http://www.sciencedirect.com/science/article/pii/ B9780124095489091776.

Rickard, A. (2020). MCM Website. http://mcm.york.ac.uk/.

Sandu, A. and Sander, R. (2006). Technical note: Simulating chemical systems in fortran90 and

matlab with the kinetic preprocessor kpp-2.1. Atmospheric Chemistry and Physics, 6(1):187–195. https://www.atmos-chem-phys.net/6/187/2006/.

- Sandu, A., Verwer, J., Blom, J., Spee, E., Carmichael, G., and Potra, F. (1997). Benchmarking stiff ode solvers for atmospheric chemistry problems ii: Rosenbrock solvers. *Atmospheric Environment*, 31(20):3459 – 3472. http://www.sciencedirect.com/science/article/pii/S1352231097832128.
- Seinfeld, J. and Pandis, S. (2016). Atmospheric Chemistry And Physics: From Air Pollution To Climate Change. Wiley. https://books.google.co.uk/books?id=n RmCgAAQBAJ.
- Stevens, J. (2020). Airborne Nitrogen Dioxide Plummets Over China. NASA Earth Observatory. https://earthobservatory.nasa.gov/images/146362/airborne-nitrogen-dioxide-plummets-over-china.
- UNEP (1987). The Montreal Protocol On Substances That Deplete The Ozone Layer. https://ozone. unep.org/treaties/montreal-protocol.
- UNFCCC (2015). The Paris Agreement | United Nations Climate Change. https://unfccc.int/ process-and-meetings/the-paris-agreement/the-paris-agreement.
- University of Zurich (2013). Great Oxidation Event: More Oxygen Through Multicellularity. Science Daily. https://www.sciencedaily.com/releases/2013/01/130117084856.htm.
- WHO (2016). World Health Organization | Ambient And Household Air Pollution And Health. *online*. https://www.who.int/airpollution/data/en/.
- WHO (2018). World Health Organization | Ambient Air Pollution: Health Impacts. *online*. https://www.who.int/airpollution/ambient/health-impacts/en/.
- Wild, O. (2015). Chemical Solvers A Slide Deck From The Ukca Theory And Practice Workshop. online. https://www.ukca.ac.uk/images/b/b1/Solvers_for_web.pdf.
- Woo, M. (2018). Planet Formation? It'S A Drag. *Scientific American*. https://www.scientificamerican.com/article/planet-formation-its-a-drag/.
- Wood, B. (2014). The Origin Of Humans Is Surprisingly Complicated. *Scientific American*. https://www.scientificamerican.com/article/the-origin-of-humans-is-surprisingly-complicated/.
- Wrangham, R. (2009). Catching Fire: How Cooking Made Us Human. Basic Books.
- Zhang, H., Linford, J. C., Sandu, A., and Sander, R. (2011). Chemical Mechanism Solvers In Air Quality Models. Atmosphere, 2(3):510–532. https://www.mdpi.com/2073-4433/2/3/510.

Chapter 2

Visualisation and its use in understanding Complex Data

"If you really want to understand something, the best way is to try and explain it to someone else. That forces you to sort it out in your mind. ... By the time you've sorted out a complicated idea into little steps that even a stupid machine can deal with, you've learned something about it yourself.¹"

- Douglas Adams, Dirk Gently's Holistic Detective Agency

 $^{^{1}}$ Omitted at ellipsis : "And the more slow and dim-witted your pupil, the more you have to break things down into more and more simple ideas. And that's really the essence of programming."

2.1 Introduction

When representing complex scientific data, we concern ourselves with finding a method which allows the reader to gain maximum insight into the information presented. This chapter begins with the evolution of humanity and explores how the development of the human brain increased our ability to communicate and store information. Next, we look at the use of storytelling (Subsection 2.2.1) and metaphor selection (Subsection 2.2.2) enable us to convey complex tasks and information in a user-intuitive way. In establishing a series of considerations for visualisation design, we apply these to the representation of atmospheric chemistry (a set of species with a production/loss relationship between them) in the form of several relational sociographs (Section 2.3). From this, it is found that the node-link graph framework is the best suited for the task - a concept which will be discussed in Chapter 3.

2.1.1 Communicatory Practices Of Early Humans

In nature, animals rely on the propagation of DNA to encode information critical to their survival. Examples of these are found in hives (where an insects role is defined by its genetic composition), or in Oscines (songbirds) which have an inherent predisposition to learn species-specific songs, [De Smedt, 2013; Hunt and Gadau, 2017; Ackerman, 2016; Wada, 2010; Harari, 2015]. For humans; however, this process is highly impractical due to the vast and varied nature of the information need to process. Instead, we have developed a predisposition to learning language at an early age. In essence, a skill allowing for the effective communication of ideas, conditions and dangers between a large number of people².

The downside to learnt behaviours, such as language, is that communicatory patterns are limited to only the people they have been taught to. Here problems of differing language and dialect significantly reduce the amount of information which may be passed between groups/tribes. Such issues were quickly overcome through the use of visualisation in the form of pictographs (cave paintings - e.g. Figure 2.1). Such methods complement our ability to both detect shapes and spot patterns within nature³ as well as providing an intuitive method of communication between separate groups.

As communities continue to increase in size, problems of accounting and resource management start to emerge. Here the ability to store large amounts of data had not been previously required by a hunter-

²Several studies, exploring the ratio of the neocortex to the rest of the brain, suggest that the number of relationships a human can successfully monitor is limited to 150. It is suggested that ideas of gossip and common metaphysical beliefs are the reason for this [Harari, 2015; Aiello and Dunbar, 1993; Dunbar et al., 1997]. This limit is still seen in social networks today [Hill and Dunbar, 2003].

³It has been found that 10,000 year-old pictographs show hints of a shared cultural background between spatially different groups of humans [M. Chazine, 2005].

gatherer species. This problem was again solved by the Samaritans (3500BC) with the creation of writing - a system for coordinating affairs and storing information external to a humans brain [Nissen et al., 1993; Schmandt-Besserat, 1992]. Using this quantities and items are depicted using a system of signs and shapes (cuneiform⁴) - a practical and intuitive way for us to apply the pattern recognition and analytical parts of our brain while reducing the cognitive load by breaking up the problem into manageable parts. Throughout history, we have continued to apply this system of intertwining data information with visual artefacts to enable people to cope with the complexities of the information provided, [Tufte, 1983]. It is for this reason that visualisation can be used as a means of enhancing the reader's ability to understand the large-scale complexities of scientific data.



Figure 2.1: An example pictograph: a 2000 year old petroglyth in Utah titled 'Tse Hane' (rock that tells a story). Source: [ugc, 2019]

2.1.2 The Origin Of Big Data

The term 'Big Data' originated in the mid-1990s, appearing in several job adverts and a slide deck by Jon Mashey, the chief scientist at Silicon Graphics (SGI), [Mashey, 1998; Diebold, 2012]. This term is used as a way to describe the ever-increasing amount of data we can generate each day. Such phenomenons are seen in anything from the growing number of websites to the number of reactions within an explicit gas-phase chemistry mechanism, Figure 2.2. Coupled with the ability to collect large amounts of data is our requirement to analyse and understand it. This has commonly be done through the use of visualisation, a topic in which careful consideration must be made to ensure the correct information is conveyed [Kirk, 2016]. In a paper on mining big data, [Fan and Bifet, 2013] explains that the main task of Big Data analysis is on deciding how to visualise the results- simply its size introduces complexity in uncovering a user-friendly method to represent the information required.

Although graphical representation has been an integral part of the data comprehension process, it is only relatively recently (1990's) that it is recognised as a research field [Wybrow et al., 2014]. Even though we are not explicitly dealing with 'big' data, the number of species and reactions occurring within the troposphere is still sufficiently large and complex that many of the same problems still exist.

 $^{^{4}}$ This is often mistaken for hieroglyphics. Although both are forms of logographic script, hieroglyphs are restricted to the ancient Egyptian sociolinguistic context.

It is for this reason that this section⁵ we will explore the considerations that need to be made before selecting a visualisation design (Section 2.2) and the methods we can use to represent the complex relationships within the atmospheric chemistry domain (Section 2.3).



Figure 2.2: Data size and complexity increases with time due to (availability or improvements in scientific understanding). The red line shows the change in the number of websites $(\log_{10} \text{ normalised})$ within the world wide web domain for years 1996-2015. Similarly, we compare the various releases of the master chemical mechanism (a discrete process), where more and more complex reaction schemes are iteratively appended over time. These are introduced later on.

NOTE: The two lines are here to illustrate the growing trends between the subjects and are not incomparable due to their different scales and continuous/discrete natures.

Source: [InternetLiveStats, 2020; Jenkins, 2002]

2.2 Visualisation Design

New ideas are developed and refined through an iterative process of cognition and discussion with other researchers [Roberts et al., 2014]. As individuals, we are constrained by our experience and knowledge; novel ideas often consist of an amalgamation of many existing concepts [Descartes et al., 1996; Johnson, 2010]. [Ziemkiewicz and Kosara, 2008] explains that the process of understanding a visualisation depends heavily on the interaction between the user's internal knowledge and the ideas depicted within a visualisation. This means that the careful selection of content and medium (of presentation) can directly influence what a reader takes away from the graphic. In its design, a visualisation must be both relatable (in metaphor) and intuitive (or explained through the use of storytelling).

 $^{^{5}}$ and consequently much of this thesis

2.2.1 Storytelling

Storytelling is commonly used to highlight the process of cause and effect. It has applications in the education for both the explanation of scientific concepts ([Martin and Miller, 1988]) and the dangers of the world (fables such as Little Red Riding Hood and Goldilocks teach the dangers of speaking to strangers and the importance of respecting personal property). As human life is subject to the conditions of the 'arrow of time', the inherent familiarity of linearly consequential events make the narrative a great way to inform the reader of new (unseen) concepts. Such methods are not limited to the education of young children, but also allow for the understanding of real-world events through the use of dreams, media and the news [Gottschall, 2012; Freud and Cronin, 2013].

In entwining narrative and visualisation, it is possible to provide the user with a higher level of understanding and enable them to draw their own (guided) conclusions from the data. Michal and Franconeri [2017] explains that within the visual analysis of a graph, users often create a narrative through the use of visual routine. Here the emphasis is placed on storytelling to 'guide' and educate the reader of any events which led to a conclusion - this uses a question-answer cause-effect structure. It is this process that makes 'storytelling' an effective method of communication to a non-expert audience [Dahlstrom, 2014].

2.2.2 Metaphor Selection

Storytelling often involves metaphors to create content which is relatable and intuitive to the user. Such metaphors have infiltrated many parts of our everyday lives ranging from descriptions in fables to the concept of money and belief that govern everyday life through the inter-subjective⁶ [Harari, 2015]. In general there exist three categories of metaphors which may be used - natural (Subsubsection 2.2.2.1), man-made (Subsubsection 2.2.2.2) and composite (Subsubsection 2.2.2.3).

2.2.2.1 Nature-Inspired

Inspiration for metaphors often comes from objects or events encountered from everyday life - the most effective of which have an inherent familiarity for all readers. As nature is universal to everyone on the planet, a nature-inspired metaphor not only guarantees a basic level of item comprehension but also contain an aesthetically pleasing familiarity to them. Common examples of these can include the use of ice to represent glacial melting or trees to show branches in decisions (decision tree: Figure 2.3a) or temporal changes in conditions with a trunk cross-section style plot (Figure 2.3b). Natural metaphors

 $^{^{6}}$ The inter-subjective is something that exists within the communication network. It allows a fictional idea, such as a limited-liability company, to exist as a real physical entity with a bank account and subject to laws.



are often useful in conveying complex ideas due to their low learning curve.

Figure 2.3: Two tree-inspired visualisations.

(a) shows the decisions made on a single decision tree within a random Forrest. Hear each branch split corresponds to a decision and the node/leaf colour represents the category of the decision. Stronger and more important decisions correspond to larger leaves and thicker branches.

(b) shows a radial plot in the shape of a tree trunk. Here time is shown radiating outwards from the centre. This allows us to spot any changes in events - much like the rings of a tree can be used to identify when natural disasters (such as tsunamis or avalanches) have struck them. This specific visualisation shows the net flux of species from a chemical simulation. These are coloured from low fluxes (blue) to high fluxes (red). The abrupt changes here show the diurnal cycle where photochemical reactions stop and then start up again.

2.2.2.2 Man-Made

Similar to nature, another similarity between many readers would be their familiarity with the urban environment. Metaphor inspiration from human-made objects such as buildings often contain characteristics of symmetry and manual/mechanical design. These allow the user to interpret any features presented using their pre-existing knowledge about an object. The most famous example of this is (Figure 2.4), where stations are positioned at 0,90 and 45-degree angles. This provides a clearer representation, much like a road map than the current space-specific location of each station. Although other designs, such as a concentric one, [Fisher, 2013], have been attempted, adaptations of Beck's design are still being used in the present day owing to their intuitive nature.



Figure 2.4: The 1933 tube map design for London. Source: [Beck, 2017]

2.2.2.3 Composite

Finally, it is possible to combine ideas into a composite metaphor, a concept common in much of greek mythology. An example of this is Pegasus, where the combination of two familiar items (wings and a horse) results in something novel and unseen (a winged horse) which has implications from its existence associated with it. Overcoming the problems presented by these composite designs involves a level of lateral thinking, prototyping (sketching) and redefining to produce a confluent user-visual metaphor interaction [Ziemkiewicz and Kosara, 2008; Roberts, 2011].

The development of new items can also be applied to the creation of xenographics - here the combination of visualisations allows for the representation of complex relationships with a smaller learning curve. An example of this is the amalgamation of a 'pizza/pie chart'⁷ with a bar chart to create a radial plot in Figure 2.5 - A novel representation style which helped explain and establish the methods of modern-day nursing by Florence Nightingale.

⁷There is a good amount of literature that suggests these are far from optimal methods of representing data.



Figure 2.5: A (stacked) Radial bar-chart showing the causes of mortality in the British army^8 , [Nightingale, 1820].

2.2.2.4 Domain Specific

Composite designs are not constrained to the combination of different metaphors. In specialised roles, it is common to find visualisations which draw on pre-existing knowledge either about the content of the figure. Although this involves a new learning curve before information about the topic can be extracted, once this is obtained, the wealth and complexity that is portrayed by a single visualisation drastically increases. Two of the more common examples where prior knowledge is required to read a plot are connected plots⁹ and Tephigrams.

T- ϕ -grams (Tephigrams) are defined from their axis of temperature (*T*), entropy(ϕ) are used in the field of meteorology and weather forecasting. The combine grid lines for constant temperature (isothermic) and pressure (isobaric), as well as allowing the user to calculate the equivalent potential temperature for saturated air parcels and the saturation mixing ratio concerning plane water surface, Figure 2.6b. This provides a good example where both scientific knowledge and ability to read the diagram are required to obtain meaningful information from it.

⁸Also known as the Nightingale Rose or Coxcomb plot.

⁹Occasionally known as snail trail chart (R users seem to like re-inventing the wheel).



(a) A tephi diagram (b) Description of each axis on the tephi diagram.

Figure 2.6: Two T- ϕ -gram plots. An example of the T- ϕ -gram data (a), and instructions on how to interpret it (b). The full scale figures are found within the source. Source: [Brooks, 2019]

Connected scatter-plots are useful in representing the temporal changes of correlation between to items. These are particularly useful in the field of economics due to their ability to highlight the different trends that can occur over time. Figure 2.7 shows how the number of vehicle-related fatalities changes with the number miles driven. In this situation, a simple x - y plot may highlight a decreasing inverse relationship between the number of fatalities and the miles driven per capita over time, but lose some of the many features presented by the additional dimension. Figure 2.7 b is an extract from [Haroz et al., 2016], which shows how the changes between two variables (blue and green) over time are represented within the connected scatterplot (yellow). Examples include unchanging variables (Figure 2.7b.A) which are represented as a single point. Like correlations are shown by a 45 degree angle line (Figure 2.7b.C) and inverse correlations are shown by a like orthogonal to the like concentration one (-45 degree)- (Figure 2.7b.D).



Figure 2.7: **Examples of connected scatterplots** (a) A connected scatterplot comparing the number of vehicle-related fatalities with the number miles driven per capita [Fairfield, 2012]

(b) a plot showing how the correlation of two spatial variables are transformed from a classical scatter/line plot into a connected scatterplot [Haroz et al., 2016]. Here the blue and green lines represent different values changing over time (orange) across the x-axis. The figures to the right show how different correlations are transformed into a connected scatter plot, where the variables (blue and green) are plotted across each axis, and time (orange) is shown as a line joining each datapoint. The data points are plotted in equal lengths of time, meaning that the distance between consecutive points informs us of the amount each variable has changed.

2.3 Visualising Relationships

Visualisation is important in conveying the impact of science and data to both the expert (e.g. complex figures in [IPCC, 2007]) and non-expert (gross over-simplifications, e.g. the climate stripes [Hawkins, 2019]). This section builds on the considerations discussed in the previous chapter and explores the different ways in which the relationship between species in the troposphere can be represented. We start by defining the dataset (Subsection 2.3.1) and then discuss several different visualisation methods (Subsection 2.3.2).

2.3.1 The Dataset

Within an atmospheric chemical model, the chemistry of the troposphere is described using a chemical mechanism. The chemical mechanism used in this thesis is known as *the Master Chemical Mechanism* (MCM) v3.3.1, which contains 5809 species and 17224 reactions [Rickard, 2020]. The MCM is a near explicit representation of our understanding of the gas-phase chemistry within the troposphere. In its mathematical form, it describes how species are related and at what rate they react.

As the aim of this section is to identify essential features within the chemical mechanism we begin by looking at the mechanisms construction protocol (Figure 2.8 and [Figure 2.9 which is the protocol for MCMGecko, but is highly similar to that of the main MCM and provides a better representation of the computational decisions used in mechanism construction]). This construction methodology mimics the reasoning and procedures performed by an analytic chemist and allows for the simultaneous construction of a consistent and compatible chemical mechanism between several areas of research. The procedure is becoming semi-automated and follows several iterations starting from a set of primary emitted VOCs. These are run from the protocol to build a set of degradation reactions in which the selected species may undergo. Any new products are then introduced to the procedure until everything has been oxidised to produced carbon dioxide and water. This produces the set of near explicit equations which are then used to describe the evolution of chemistry within an atmospheric model.



Figure 2.8: The generation flowchart used within the MCM. This shows the process undergone by a new species to generate its products. Any unseen products are then fed back into the flowchart until the entire mechanism has been produced. Source: [Saunders et al., 2003]

2.3.2 The Sociograph

Social network analysis is a type of sociological work (sociometry) that aims to reveal the interweaving and interlocking relations between items or individuals, [Scott, 1988]. [Dunbar et al., 1997] argues that the evolution of language as a result of social grooming (gossip), and therefore its adaption for storytelling. It follows that using the social network construct to represent the underlying patterns between objects makes use of both our inherent ability to discern complex patterns through the use of storytelling. In trying to depict the relationships between a social network, we employ the use of sociograms. These are a class of visualisations which reveal certain properties of a social network. The following section will look at the extraction of useful information from the relationships within the MCM through the use of several sociograms.

2.3.2.1 The Chord Diagram

A chord diagram is a visual sociogram known for its use in summarising the overarching relations within a dense social network [Jalali, 2016]. Here arcs are used to represent groups (or a node from a social network graph), and their length corresponds to the percentage of items they contain. Within Figure 2.10 we represent the different routes a species may react within the MCM protocol flowchart shown in Figure 2.9. The figure contains two sets of arcs. These represent the different level of



Figure 2.9: The generation flowchart used within Gecko. This is very similar to the MCM protocol (Figure 2.8), but provides a clearer representation of the machine read desicions for the mechanism generation. Source: [Aumont et al., 2005]

classification of the chemistry - The first (outer) arc represents the split in channels between radical (red) and non-radical (orange) species which further separated into the finer categories within each branch (the labelled inner arcs). The inner arcs again show the probability a randomly chosen reaction on a species will fall under a particular category - if we convert an arc into a segment we have a pie chart of probability.

It is worth noting that segment sizes do not represent the number of species undergoing a specific reaction pathway, but rather the percentage of all possible pathways which follow that route. This is because species often undergo a range of reactions, each of which counts as an individual weighting. It is for this reason that even though almost all¹⁰ contain a C-H bond, hydrogen abstraction does not consume the whole graph. Many species have multiple possible pathways in which they may react, and the chord diagram presents the likeliness of a rection for all possible methods of reaction for all species.

From this, we see that hydroxy reactions are the most common with C-H bonds being in abundance¹¹. Additionally, we find that when applying the MCM protocol, a third of species contain at least one carbonyl group. Next, we look at the co-occurrence of branches for different species. These are represented using the area of a circle connecting two arcs (a chord). Each chord has two edges connecting two arcs¹². It is possible to discern the percentage of items going between these and other branches by comparing the width of each chord to its parent arc. Here, for example, we see a roughly even split between species with a C-H bond (i.e. all species) and every other group. This suggests an even distribution of reaction types between species. This means that in comparing the arc length of each chord, we can visually determine the percentage of group A which relates to its partner group B. Finally it is also possible to determine the number of items in a group which contain themselves. Chemically these are species with multiples of one functional group that undergo a specific reaction pathway more than one time. Although these reactions will usually be combined within a mechanism (to avoid duplication), their rate would be increased accordingly.

 $^{^{10}\}mathrm{Except}$ for any inorganic species.

 $^{^{11}\}mathrm{This}$ is seen within the graph layout Figure 3.10

¹²except for self-loops, although these are addressed below.



Figure 2.10: A chord diagram of the protocol structure. This shows the proportional probability a species from the MCM will follow one or more of the paths presented in Figure 2.9. The outside ring represents the radical (red), and non-radical (gold) split between groups. The inside ring splits these into individual groups, providing a finer level of detail.

The chord layout provides an easy way to calculate the percentage of items which contain multiple properties. It requires a relatively low learning curve and is intuitive to those with experience using pie charts (namely the Microsoft Office generation), however, this radial format can sometimes also make it more challenging to read. Special attention needs to be paid to the order arcs are drawn, as with some datasets, specific configurations may obfuscate trends. Finally, the chord diagram requires a certain amount of data munging before its employment. It is due to this that finer details about the system may be lost, especially if there is a course scale between chord sizes.

2.3.2.2 Direct Representation Of The Relational Matrix

Relational matrices, also known as the adjacency matrix, are a $n \times n$ square matrix highlighting the relationships between items. These can be constructed (or used to construct) a graph and provide the easy identification of patterns between nodes. Figure 2.11 shows an MCM subset of propane sorted by the number of carbons. Here we see that for this limited sample species tend to have slightly more reactions with other species with the same number of carbons than with different numbers.

The main downside of visualising the adjacency matrix is the sparsity of many real-world graphs (Chapter 4). If we take the complete MCM v3.3.1 network [Saunders et al., 2003] and visualise it in this manner, it will have a density of 5.9^{-4} . This means that for an average 600×600 pixel figure, only 0.36^2 of a pixel would be coloured in. Moreover, since we are not able to colour only parts of pixels, our final plot will remain blank. Methods to circumvent this involve looking at subgraphs, or individual sections interactively or applying composite graph/adjacency techniques such as those presented in NodeTrix, a method for visualising sparse small world networks¹³ [Agarwal et al., 2017]



Figure 2.11: The adjacency or relational matrix is showing species from a propane subset of the MCM. Here species are sorted by the number of carbons they contain (red= 3, orange= 1). Black cells represent reactions between species containing different numbers of carbons. If looking at clusters in a network (Chapter 5) sorting elements in an adjacency matrix aid in the visual identification of these. Squares features along the diagonal indicate a larger number of links between grouped items colocated in the axis, and less to those in other locations (clusters).

 $^{^{13}\}mathrm{Small}$ world networks are discussed in the following chapter.

2.3.2.3 Arc Diagrams

Arc diagrams are a subset of sociographs where items are represented as nodes along the horizontal. Relationships between nodes are then shown through the use of curved links (arcs). It is this that makes them particularly suited for the highlighting of repetition between music or DNA sequences, [Wattenberg, 2002]. Using the MCM database, we can construct an arc diagram to explore how species containing the same combination of functional groups react. This is done by first determining all branch permutations from the protocol flowchart (Figure 2.9). This produces 179 groups which are positioned across the x axis in ascending group size (the more branches matched, the further to the right a group is positioned).



Figure 2.12: An arc diagram from the MCM database derived by the protocol (Figure 2.9). All possible pathways for each species are extracted. These are then grouped and sorted by the number of functional group/pathways available. Species are shown along the blue line and sorted by increasing number groups to the right. Reaction pairs (reactant-product) are then depicted through the use of arcs. If a species reacts to produce a species with more functional groups, it is represented by a forwards facing arc. If the product contains a smaller number of functional groups, the drawn arc moves downwards and backwards.

The width of each group is determined from the number of species within it. Following links are added between the species based on what groups the product species from each reaction contain. Figure 2.12

discretises reactions which produce products with an increased number of reaction pathways (positive arcs - top), from those which result in species with less (positive arcs - below). Here the cyclic nature of tropospheric chemistry can be seen, with many species producing larger, less stable, products, which then go on to react and decompose back into smaller ones. In some cases, a complete circle between two nodes can be indicative of a catalytic reaction. Using interaction and selective shading, it is possible to isolate specific types of reactions and determine which of the functional groups are responsible for the change experienced within a reaction.

Although such a layout may seem daunting at first, with many lines, in all directions, filtering by type of reactions can draw attention to several features from the chemistry. In Chapter 1, the importance of HO_x cycle for the removal of VOCs and greenhouse gasses was discussed. For this reason, we begin by using arc diagrams to explore the relationship between OH and HO₂ (Figure 2.13a).

Figure 2.13c and Figure 2.13d show arc diagrams where the reactions of interest (photolysis and OH reactions respectively) highlighted in both colour and opacity. These enable us to see patterns between the radical cylcing of OH \longrightarrow HO₂ chemistry (Figure 2.13b). Here the cyclic reaction shown between the dashed lines corresponds to the reaction of RO₂ $\stackrel{\text{HO}_2}{\underset{O_2}{\leftarrow}}$ ROOH (Figure 2.13b).

Applying the same methodology to photolysis and hydroxy reactions, the production of species containing fewer functional groups is seen in Figure 2.14a. Within the highlighted reactions, it is seen that a ROOH species undergoes a reaction with OH or photolyses (Figure 2.14b). In the OH reaction, Hydrogen abstraction is performed to produce an RO₂ species and water, ROOH \xrightarrow{OH} RO₂ + H₂O. Photolysis reactions, however, photolyse, ROOH \xrightarrow{HV} RO₂ + HO₂, reducing the number of functional groups - producing a larger arc.

Finally, Peroxy Acetyl Nitrates (PANs), play a vital role in the modelling of photochemical smog (ozone events), [Singh, 2015]. PANs an effective reservoir species with significant importance within the production of ozone in atmospheric chemistry models (especially if transportation is involved) [Finlayson-Pitts and Pitts, 2000]. Although they are very stable at cold temperatures, these can quickly decompose (thermally) to release NO_x if warmed. In the MCM the thermal decomposition of PANS is determined by the KBPAN rate constant. In comparing reactions of Figure 2.15d, with those of Figure 2.15c (at rate KFPAN), we see a cycle between two arcs forming (Figure 2.15a). This can be explained by the reactions in Figure 2.15b which show that $RC(O)OONO_2 \xrightarrow{KBPAN} RC(O)O_2$ (+NO₂) $\xrightarrow{NO_2} RC(O)OONO_2$.

NOTE: A downside to the arc diagrams format that has been chosen is that for reactions between species of the same number of functional groups, there is no set direction.



Figure 2.13: Arc diagram features for the Hydroperoxyl and Hydroxide radicals. The HO_x cycle chemistry (b) can ge seen within certain groups in the network. The main ones of these are hilighted in (a).



Figure 2.14: Arc diagram features for photolysis and hydroxide reactions. Photolysis results in species with a reduced number of functional groups, and therefore longer arcs. OH reactions for the same species do not produce such a drastic change on group number, and therefore have a smaller arc length.



Figure 2.15: Arc diagram features for the PAN reactions.
2.3.2.4 The Traditional Network Graph

Finally, we have the traditional network representation in the form of a mathematical graph. Here species are represented as nodes (circles) and reactions as the links (lines) between them. This analogy has its roots in social representation and can be described using the metaphor of people holding hands - a concept familiar to most people. Graph representations allow for an overview of the structural relationships within the MCM network, and even to compare it against other reduced mechanisms, Figure 2.16 Here we show the growth of the MCM (left) against two versions (three variations) of the reduced Common Representative Intermediates (CRI) [Jenkin et al., 2008] mechanism in the same space. By fixing species which exist in mechanisms groups (generally the primary emitted VOCs) we produce a 'fingerprint'-like structure we can use to visually identify changes in their size, interconnectedness (density) and structure.

Building on this, an interactive visualisation (Figure 2.17) was constructed to better reveal the differences between of each mechanism in (Figure 2.16). The code for this can be found in [Ellis, 2020]. Figure 2.17a shows the expansion from MCM version 3.1 to 3.2 which included new schemes for crotonaldehyde, ethylene oxide and vinyl chloride, the introduction of methacolein and the integration of dimethyl sulphide (DMS), beta-caryophyllene and limonene [Rickard, 2020] - the latter of which is responsible for the additional South-West pointing branch seen within the graph representations. Similarly, Figure 2.17b shows the upgrade from MCM v3.2 to v3.3.1, the main change is the mechanism update to include the complete degradation mechanism for isoprene [Jenkin et al., 2015]. This change results in the addition of 100 species, many of which are mainly related to OH initiated chemistry. However, since the ratio of species to links (reactions) has now increased, these lie closer to the main body of the network - the reason for which is discussed in Chapter 4.

We can use Figure 2.17 to emphasise the amount that has been added (or lost) in reduction or development. Figure 2.17c shows the difference between the MCM v3.2 and its reduced CRI v2.0 form, which focuses on preserving the overall ozone-forming potential of the mechanism. Figure 2.17d shows a comparison of the CRI v2.0 after a further 5 reductions (CRI v2.0 r1). Using these two plots we can identify regions or branches of chemistry which have been removed (namely biogenic and anthropogenic aromatic branches - bottom left and bottom right) and generate an overview of how well the reduced mechanism structure represents all parts of the contained chemistry. We can see that on average the CRI mechanism does a good job at retaining the core network structure, often lumping the more esoteric (or extreme) branches into a single species at their base.

This type of network representation is found not only the simplest and most intuitive but also the most informative about what effects changing the underlying chemistry may have on a simulation. Chapter 3 expands on the sociograph idea and explores the different ways in which we may tune it to



maximise its potential for useful knowledge transfer.

Figure 2.16: Comparing a range of MCM and CRI mechanims using their graph shape and structure. Source: Ellis [2020]



(c) MCM v3.2 vs CRI v2.0(r1)

(d) CRI v2.0(r1) vs CRI v2.0(r5)

Figure 2.17: Voronoi cells of each node from the graph layout - used to identify changes in mechanisms. A difference plot between the different graphs in Figure 2.16. These use colours to show us species that are added or taken away between different versions. Subplots (a) and (b) show the increases mechanism size of the MCM whilst (c) and (d) show the reduction from MCM v3.2 to CRI v2.0(r1), and followed by the fith reduction to CRI v2.0(r5). Figure colouring: purple cells only exist within the first mechanism, pink only exist within the second, and blue are present in both. Source: Ellis [2020]

2.4 Conclusion

Human cognitive capacity is limited in its skill to comprehend complex information. The use of visualisation can alleviate some of this difficulty by employing our inherent pattern recognition ability and exporting the problem to exist outside our brains. Using storytelling and narrative can not only help us understand a problem but also enables us to explain it to other people. This is particularly important when trying to convey an essential subject to a non-expert audience, such as policymakers or the public.

If the data is relational, a node-link style sociograph provides the best visual representation of the data. We have explored several different sociographs (chord, arc and graph) and found that the graph format provides the most straightforward and most practical approach for the representation of the MCM chemical mechanism. This approach appears to highlight the features of the network structure while allowing us to compare and contrast different atmospheric chemical mechanisms. It is for this reason that Chapter 3 will further explore the use of graph-based representation in representing different chemical schemes of the atmosphere.

Bibliography

- Ackerman, J. (2016). The Genius Of Birds: The Intelligent Life Of Birds. Little, Brown Book Group. https://books.google.co.uk/books?id=3z_sCgAAQBAJ.
- Agarwal, S., Tomar, A., and Sreevalsan-Nair, J. (2017). Nodetrix-Multiplex: Visual Analytics Of Multiplex Small World Networks. In *Complex Networks & Their Applications V*, pages 579–591. Springer International Publishing. http://dx.doi.org/10.1007/978-3-319-50901-3_46.
- Aiello, L. C. and Dunbar, R. I. M. (1993). Neocortex size, group size, and the evolution of language. *Current Anthropology*, 34(2):184–193. http://www.jstor.org/stable/2743982.
- Aumont, B., Szopa, S., and Madronich, S. (2005). Modelling The Evolution Of Organic Carbon During Its Gas-Phase Tropospheric Oxidation: Development Of An Explicit Model Based On A Self Generating Approach. Atmospheric Chemistry and Physics, 5:2497–2517. https: //www.atmos-chem-phys.net/5/2497/2005/acp-5-2497-2005.pdf.
- Beck, H. (2017). London Icon: A History Of Harry Beck'S Iconic Tube Map. https://londontopia. net/site-news/featured/london-icon-tube-map/.
- Brooks, I. (2019). Tephigram.pdf. online. http://www.met.reading.ac.uk/~sgs02rpa/TEACHING/ Tephigram.pdf.
- Dahlstrom, M. F. (2014). Using Narratives And Storytelling To Communicate Science With Nonexpert Audiences. Proceedings of the National Academy of Sciences of the United States of America, 111 Suppl 4:13614–13620. http://dx.doi.org/10.1073/pnas.1320645111.
- De Smedt, T. (2013). Modeling Creativity: Case Studies In Python. Proefschriften UA-LW : taalkunde. Universiteit Antwerpen, Faculteit Letteren en Wijsbegeerte, Departement Taalkunde. https:// books.google.co.uk/books?id=Bp7KwpmFBzoC.
- Descartes, R., Cottingham, J., and Williams, B. (1996). Descartes: Meditations On First Philosophy: With Selections From The Objections And Replies. Cambridge Texts in the History of Philosophy. Cambridge University Press. https://books.google.co.uk/books?id=yMwiTTpwasgC.
- Diebold, F. X. (2012). A personal perspective on the origin(s) and development of big data: The phenomenon, the term, and the discipline, second version. PIER Working Paper Archive 13-003, Penn Institute for Economic Research, Department of Economics, University of Pennsylvania. https: //ideas.repec.org/p/pen/papers/13-003.html.

- Dunbar, R. I. M., S. A., Atzwanger, K., Grammer, K., and Schäfer, K. (1997). Groups, Gossip, And The Evolution Of Language, pages 77–89. Springer US, Boston, MA. https://doi.org/10.1007/ 978-0-585-34289-4 5.
- Ellis, D. (2020). Wolfiex/mcm-blueprint: Thesisref. Zenodo. https://doi.org/10.5281/zenodo.4294816.
- Fairfield, H. (2012). Driving Safety, In Fits And Starts. The New York Times. https://www.nytimes. com/interactive/2012/09/17/science/driving-safety-in-fits-and-starts.html.
- Fan, W. and Bifet, A. (2013). Mining big data: Current status, and forecast to the future. SIGKDD Explor. Newsl., 14(2):1–5. https://doi.org/10.1145/2481244.2481246.
- Finlayson-Pitts, B. J. and Pitts, J. N. (2000). Chemistry of the upper and lower atmosphere. Academic Press, San Diego. http://www.sciencedirect.com/science/article/pii/B9780122570605500034.
- Fisher, J. (2013). Alternative Tube Maps: Circles Within Circles. https://londonist.com/2013/01/ alternative-tube-maps-circles-within-circles.
- Freud, S. and Cronin, A. (2013). The Interpretation Of Dreams. Read Books Limited. https://books. google.co.uk/books?id=U0t8CgAAQBAJ.
- Gottschall, J. (2012). The Storytelling Animal: How Stories Make Us Human. Houghton Mifflin Harcourt. https://books.google.co.uk/books?id=Gd3lT5yP3ZQC.
- Harari, Y. (2015). Sapiens: A Brief History Of Humankind. Harper. https://books.google.co.uk/ books?id=FmyBAwAAQBAJ.
- Haroz, S., Kosara, R., and Franconeri, S. L. (2016). The Connected Scatterplot For Presenting Paired Time Series. *IEEE transactions on visualization and computer graphics*, 22(9):2174–2186. http://dx.doi.org/10.1109/TVCG.2015.2502587.
- Hawkins, E. (2019). #Showyourstripes. https://showyourstripes.info/.
- Hill, R. A. and Dunbar, R. I. M. (2003). Social Network Size In Humans. *Human nature*, 14(1):53–72. http://dx.doi.org/10.1007/s12110-003-1016-y.
- Hunt, G. and Gadau, J. (2017). Advances In Genomics And Epigenomics Of Social Insects. Frontiers Research Topics. Frontiers Media SA. https://books.google.co.uk/books?id=lvItDwAAQBAJ.
- InternetLiveStats (2020). Total Number Of Websites Internet Live Stats. https://www. internetlivestats.com/total-number-of-websites/.
- IPCC (2007). Fourth Assessment Report: Climate Change 2007: The Ar4 Synthesis Report. Geneva: IPCC. http://www.ipcc.ch/ipccreports/ar4-wg1.htm.

- Jalali, A. (2016). Reflections on the use of chord diagrams in social network visualization in process mining. In 2016 IEEE Tenth International Conference on Research Challenges in Information Science (RCIS), pages 1–6.
- Jenkin, M., Watson, L., Utembe, S., and Shallcross, D. (2008). A common representative intermediates (cri) mechanism for voc degradation. part 1: Gas phase mechanism development. Atmospheric Environment, 42(31):7185 – 7195. http://www.sciencedirect.com/science/article/pii/ S1352231008006742.
- Jenkin, M. E., Young, J. C., and Rickard, A. R. (2015). The mcm v3.3.1 degradation scheme for isoprene. Atmospheric Chemistry and Physics, 15(20):11433–11459. https://www.atmos-chem-phys. net/15/11433/2015/.
- Jenkins, M. (2002). History Of The Master Chemical Mechanism (MCM) And Its Development Protocols. slide deck. Presentation for the EPSR group, Imperial Collage.
- Johnson, S. (2010). Where Good Ideas Come From. Penguin Publishing Group. https://books.google. co.uk/books?id=3H2Xg5qxz-8C.
- Kirk, A. (2016). Data Visualisation: A Handbook For Data Driven Design. SAGE Publications. https://books.google.co.uk/books?id=wNpsDAAAQBAJ.
- M. Chazine, J. (2005). Rock Art, Burials, And Habitations: Caves In East Kalimantan. Asian Perspectives, 44(1):219–230. https://muse.jhu.edu/article/185391.
- Martin, K. and Miller, E. (1988). Storytelling And Science. Language Arts. https://www.jstor.org/ stable/41411379.
- Mashey, J. (1998). Big Data And The Next Wave Of Infrastress . *online*. https://static.usenix.org/ event/usenix99/invited_talks/mashey.pdf.
- Michal, A. L. and Franconeri, S. L. (2017). Visual Routines Are Associated With Specific Graph Interpretations. Cognitive research: principles and implications, 2(1):20. http://dx.doi.org/10. 1186/s41235-017-0059-2.
- Nightingale, F. (1820). Notes On Matters Affecting The Health, Efficiency And Hospital Administration Of The British Army 1820-1910. https://www.rct.uk/collection/1075240/ notes-on-matters-affecting-the-health-efficiency-and-hospital-administration-of.
- Nissen, H., Damerow, P., Englund, R., Englund, R., Larsen, P., and Larsen, R. (1993). Archaic Bookkeeping: Early Writing And Techniques Of Economic Administration In The Ancient Near East. University of Chicago Press. https://books.google.co.uk/books?id=YBAzXV4YtQ8C.

Rickard, A. (2020). MCM Website. http://mcm.york.ac.uk/.

- Roberts, J. C. (2011). The Five Design-Sheet (Fds) Approach For Sketching Information Visualization Designs. In Maddock, S. and Jorge, J., editors, *Eurographics 2011 - Education Papers*. The Eurographics Association.
- Roberts, J. C., Yang, J., Kohlbacher, O., Ward, M. O., and Zhou, M. X. (2014). Novel Visual Metaphors For Multivariate Networks, pages 127–150. Springer International Publishing, Cham. http://dx.doi.org/10.1007/978-3-319-06793-3_7.
- Saunders, S. M., Jenkin, M. E., Derwent, R. G., and Pilling, M. J. (2003). Protocol For The Development Of The Master Chemical Mechanism, MCM V3 (Part A): Tropospheric Degradation Of Non-Aromatic Volatile Organic Compounds. *Atmospheric Chemistry and Physics*, 3(1):161–180. https://hal.archives-ouvertes.fr/hal-00295229.
- Schmandt-Besserat, D. (1992). Before Writing, Vol. I: From Counting To Cuneiform. Before Writing. University of Texas Press. https://books.google.co.uk/books?id=_G74dDQO8gUC.
- Scott, J. (1988). Social network analysis. *Sociology*, 22(1):109–127. https://doi.org/10.1177/0038038588022001007.
- Singh, H. (2015). Tropospheric chemistry and composition | peroxyacetyl nitrate. In North, G. R., Pyle, J., and Zhang, F., editors, *Encyclopedia of Atmospheric Sciences (Second Edition)*, pages 251 – 254. Academic Press, Oxford, second edition edition. http://www.sciencedirect.com/science/ article/pii/B9780123822253004333.
- Tufte, E. (1983). The Visual Display Of Quantitative Information. Number v. 914 in The Visual Display of Quantitative Information. Graphics Press. https://books.google.co.uk/books?id= tWpHAAAAMAAJ.
- ugc (2019). Newspaper Rock. http://www.atlasobscura.com/places/newspaper-rock.
- Wada, H. (2010). The Development Of Birdsong | Learn Science At Scitable. https://www.nature. com/scitable/knowledge/library/the-development-of-birdsong-16133266.
- Wattenberg, M. (2002). Arc diagrams: Visualizing structure in strings. In IEEE Symposium on Information Visualization, 2002. INFOVIS 2002., pages 110–116.
- Wybrow, M., Elmqvist, N., Fekete, J.-D., von Landesberger, T., van Wijk, J. J., and Zimmer, B. (2014). Interaction In The Visualization Of Multivariate Networks, pages 97–125. Springer International Publishing, Cham. http://dx.doi.org/10.1007/978-3-319-06793-3 6.

Ziemkiewicz, C. and Kosara, R. (2008). The shaping of information by visual metaphors. *IEEE Transactions on Visualization and Computer Graphics*, 14(6):1269–1276. http://dx.doi.org/10. 1109/TVCG.2008.171.

Chapter 3

Applying Visual Analytics to the Atmospheric Chemistry Network

" I have a notion that when the mind is thinking, it is simply talking to itself, asking questions and answering them."

- Socrates, The collected dialogues of Plato

3.1 Introduction

Chapter 2 viewed the importance of a carefully selected visualisation/metaphor in the representation of scientific data. One such category is that of relational data, where we have a set of items, joined by a chosen relationship. Historically this type of problem was solved using sociographs to present a set of items and the links between them. This chapter begins by looking at the use of sociograms in chemistry (Subsection 3.1.2) and the different ways in which these can help convey information to the reader (Section 3.2, Section 3.4). These sections find the force-directed graph to be the most suited for representing the chemical reactions within a mechanism, and therefore this shall be applied to the network of reactions representing the chemistry within an urban environment - Beijing (Section 3.5).

3.1.1 Networks And Their Role In Visual Analytics

Networks are present everywhere - this ranges from interactions within social media to bank transactions, internet routing, genetics to epidemiology [Martin Grandjean, 2016; Staples et al., 2013; Needham and Hodler, 2019; Baronchelli et al., 2013; Sangers et al., 2019; Kohlbacher et al., 2014; Archambault et al., 2014; Schreiber et al., 2014]. The sociogram (or graph) structure can be applied to any set of items which contain one or more relationships between them. In visualisation, these 'items' are referred to as nodes/vertices, and their relationships as edges/links [Kerren et al., 2014] terms that will be used interchangeably throughout this thesis.

3.1.2 Graphs In Chemistry

Node-link representations have been at the core of chemistry for many years. They have been used to show the bonds between atoms and are integral to the representation of molecules - both physically (with the aid of molecular model kits) or pictorially to show various structural properties (Figure 3.1). These graph-like analogies provide a pseudo-physical representation of the molecules and their reactions in a way that is intuitive to the user. Subsubsection 3.1.2.1 shows the use of a sociograph structure to represent reactions within the troposphere; however, this method of representation is not limited to atmospheric science - for example, Figure 3.3 depicts the biochemical metabolic pathways of the human body. This is an example of another complex chemical network that benefits from this method of representation.



Figure 3.1: The molecule C141CO3 (MCM name) shown in both 2D and 3D node-link structures. This is a the result of a series of inorganic species reactions and a desocciation from BCARY - the only sesqueterpine in the MCM. 3D visualisation by [Bergwerf, 2019].

3.1.2.1 Using Sociograms To Describe Reactions

A collection of reactions representing the chemistry of a region is called a mechanism. The Master Chemical Mechanism [Rickard, 2020] provides a collection of equations describing the gas-phase chemistry which exists within the troposphere (Subsection 2.3.1). In its use in policy, and the evaluation of Air Quality Models ([Dick Derwent, Andrea Fraser, John Abbott and Mike Jenkin , 2010]), it is often useful to understand the degradation process different VOCs undergo. In general, this can be achieved through a series of interconnected reactions in the form of a reaction cycle (Figure 3.2). This type of sociograph shows the directional nature of chemical reactions and the relationships between different species. This has many similarities to a conventional directed graph, except that species (nodes) are sometimes duplicated (for example OH, HO_2 , O_2 in Figure 3.2) to aid in the clarity of the figure.

This provides an excellent example of how the flow-like nature of a sociogram aids in the understanding of a potentially complex chemical system of 171 organic species and 600 reactions. Evolutionary traits, including the genetic predisposition to interpret shapes faster than text ([Harari, 2015]) make the graph structure a much better method for representing such a system.

3.1.3 Modeling Chemistry As A Directed Graph.

Historically it is shown that the graph format has proven to be an efficient means of understanding the reactions within a mechanism. Traditionally these are constructed manually, with the designer making a series of choices on how best to place, and simplify the chemistry based on their application. As our understanding of chemistry improves and we have started to progress into automated and semi-automated mechanism construction. This makes the construction of mechanisms with tens of millions of species and billions of reaction possible ([Aumont et al., 2005]) and is the point where the manual design/simplification of reaction networks becomes infeasible.



Figure 3.2: A systematic representation of the degregation of butane. Using this we are able to see the process C_4H_{10} undergoes before its ultimate demise as carbon dioxide and water. Source: [Jenkin et al., 1997]



VisualAnalyticstotheAtmosphericChemistryNetwork Today automatic graph layouts allow us to generate multivariate and complex graphs quickly [Muelder et al., 2014]. This means that, much like in the construction of a mechanism, we can rely on computeraided design to generate a directed graph representation of the chemistry. Montañez [2016] states that "The beauty of a good information graphic is that it can tell a whole story in a single unit of visual content". This is particularly true for the use of directed graphs in chemistry where we can compare different mechanism structures.

However, several problems emerge from the complete automation of a task. Firstly real-world data very rarely reacts how it is expected to. Here networks of high edge density often obfuscate the graph data and produce what is only described as a 'birds nest', 'hairball' or 'ball of yarn' within the literature [Roberts et al., 2014]. Although such problems can be shown as moments of turbulence, they encourage a greater understanding of the graphic design process and can catalyze to merge unique ideas into an effective visualisation [Johnson, 2010] - much like the composite metaphors in Chapter 2. Having established that a graph network ties in both modern and historical methods for representing relational data, we now look at how to present the graph, both in syntax (Section 3.2) and semantics (Section 3.4).

3.2 Graph Syntactics

Syntactic representation considers how best to distribute information on a page for maximum impact. This can be seen between the force-directed graph (top) and geographical location (bottom) layouts in Figure 3.4. Although the geographical layout gives a more accurate representation of the distances between unconnected nodes (airports), a force-directed graph provides greater insight into the relationships (flights) between each airport. This highlights the importance of choosing a suitable syntactic representation to highlight the features of interest. The remainder of this section discusses the syntactic choices required for the visualisation of a complex chemical mechanism.



Figure 3.4: Comparison of different representations of flight data by [Martin Grandjean, 2016]. The top figure shows the data represented by a force-directed graph layout (described below) and a Geo-layout showing each point at its location on the Earth.

3.2.1 Selecting The Correct Evaluation Criteria.

As chemical networks provide a wealth of information on the reactions within a system, this can prove challenging to user cognition and computational resources [Kerren et al., 2014]. In selecting the best possible graph layout, there are many metrics designed around the improving of visualisations aesthetics [Purchase, 2002]; however, these have often only been evaluated with a handful of criteria in mind. Such metrics can make it difficult to accurately quantify the changes in user-readability, especially if they are not treated as originally intended [Pohl et al., 2009].

3.2.1.1 Edge Crossing

One of the greatest limitations to understanding a graph is the number of overlapping (crossing) edges [Purchase, 1997], especially since users often spend most of their time looking at the edges of a graph to understand it [Pohl et al., 2009].

There exist several types of graph layout algorithms which aim to reduce the number of overlapping edges in a graph. The two most common ones are force-directed and orthogonal. Orthogonal designs are those of straight edges at 90-degree angles, such as in architectural or circuit schematics (Figure 3.5). Force-directed graphs (Subsubsection 3.2.2.3) are a graph layout is designed to simulate a physical system, where node positions are the result of the push and pull of the edges between them. In the task selecting nodes from a specific path, users were twice as more accurate using this layout than the orthogonal one [Pohl et al., 2009].



Figure 3.5: An orthogonal circuit schematic of the Model 3 240 portable cathode ray tube television. The circuit schematic of the television (top left) shows a much simpler representation of how different components within the television are connected. Source: [JVC, 2020]

3.2.1.2 Node Distribution and Overlap

The distribution of nodes across the page can both hinder or increase the readability of a graph especially since larger nodes may obscure smaller ones at the same location. Purchase et al. [2003] found that graphs with an equal node distribution across space, at a medium edge length, greatly improved graph readability - with node distribution and graph-symmetry ranking second in a study on user preference on graphs.

In addition to selecting the best graph layout, there are several methods in which overlapping nodes may be removed - an issue that is sometimes difficult by the treating of nodes as 'point masses' within an algorithm [Dwyer et al., 2006c]. Dwyer et al. [2006b] explains that there are usually two methods for reducing the number of overlapping nodes in a graph; these are:

- Create a layout design capable of taking node size (e.g. [Friedrich and Schreiber, 2004]) into consideration. These designs tend to be layout specific and not absolute in removing all overlap between nodes.
- 2. This requires a level of post-processing in the form of a 'layout adjustment'. Here we reposition nodes after a chosen layout has finished computing. The drawback of this method is that information contained in the graph's shape may be degraded. This can be done through the use of collision detection, or moving nodes to the centre of the vornouli cells [Lyons, 1992].

3.2.2 Automated Graph Drawing Layouts

In their design and evaluation, automatic graph drawing algorithms are created to minimise a specific criterion. This subsection will compare several graph layouts and make a verdict on which one is most suited for the representation of tropospheric chemistry. This task shall use the mechanism extracted in Table 4.4 to represent the VOC's within the Beijing city - a real-world case study using the MCM.

To do this we begin by exploring hand-drawn / map inspired graph layouts (Subsubsection 3.2.2.1,Subsubsection 3.2.2.2), eventually ending at a number of automated forcedirected graphs (Subsubsection 3.2.2.3).

3.2.2.1 Replication Of Hand-Drawing Methods

With the rise of computation, many traditional visualisations adapted for the computer-aided generation. Fields of architecture and circuit design adopted computational software to alleviate some of the difficulties presented by large or complex designs. Similar ideas such as the use of automatically generated transit maps can be used to link chronological or topological items such as ideas [Foo, 2019]. Figure 3.6 shows all the possible paths for the oxidation of methane to produce carbon dioxide (and water), using the MemoryMap algorithm Foo [2019]. Although such methods can be useful in showing isolated pathways, they provide a convoluted representation of large interconnected systems and require some manual intervention.



Figure 3.6: A transit map showing all the possible routes from methane to carbon dioxide. This was drawn using MemoryMap [Foo, 2019] and uses a version of the MCM methane subset, where carbon dioxide has been introduced.

3.2.2.2 Projection Based

One of the oldest fields of data visualisation fall in the realm of cartography. Here the shapes and distances between points on the surface of the earth (an oblate spheroid) are mathematically mapped onto a 1D plane for graphing purposes [Thomas, 1952]. Since the process of dimensionality reduction will produce inherent distortions within the final product, we end up with a range of map projections, with each striving to achieve a different aim (Figure 3.7). The Pierce Quincuncial, for example, is a conformal mapping technique mapping the surface of a sphere to a square with minimal deviation in scale and the ability to be tessellated in all directions. The Mercator, on the other hand, is a cylindrical projection which grew in popularity due to its unique ability to represent any course of constant bearing¹ as a linear segment within the shipping and navigation industry. Finally, the waterman butterfly presents the globe as a truncated octahedron. This allows for the reconstruction of a three-dimensional world from a 2D plane (i.e. a printed sheet).

 $^{^{1}}$ Also known as a 'rhumb', or 'loxodrome', and consists of an arc crossing all meridians of longitude at the same angle.



(a) Pierce Quincucial

(b) Waterman Butterfly

(c) Mercator

Figure 3.7: A selection of map projections. These have been created using DataDrivenDocuments [Bostock, 2012] and show a range of methods for mapping the spheroid shape of the Earth onto a 2D plane.



(a) The Mercator graph.

(b) Mercator species distribution

Figure 3.8: **The Mercator Projection.** (a) represents the output from the Mercator graph layout algorithm. (b) provides a kernel density analysis of the node distribution within this. Here (a) shows graph structure by revealing the density of connections between different nodes, while (b) reveals the density of nodes at a specific location.

More recently, the mathematics of mapping a large dimension onto a simpler one has been applied to the problem of graph representation. [García-Pérez et al., 2019] uses the latent hyperbolic geometry of the Mercator layout to provide a 2D embedding for complex real-world networks. This produces a polar representation $(r \text{ and } \theta)$ of the system, where relationships of related species are of the same angle (θ) , with nodes of a high degree are closer to the centre (low r value, where r is the radius from the centre). Using the chemical mechanism from the APHH Beijing campaign (described above), this produces a layout, (Figure 3.8) where (a) shows the graph-based representation including links, and (b) shows the density distribution for all nodes. Figure 3.8b shows that primary emitted species (orange dots) are uniformly (radially) distributed for angles and Figure 3.8a reveals that influential nodes with a high degree (highly connected) are located close to the centre of the graph. Although the Mercator embedding does reduce the 'hairball' problem experienced by other layouts, it does not take edge weight/direction or self-loops. This means that it works well for the representation of the general network layout, but cannot be used for advanced data exploration concerning simulation results.

3.2.2.3 Force-Directed

Force-directed graph layouts are the results of the Spring-Electrical model. This was first introduced by [Eades, 1984] and further improved by [Fruchterman and Reingold, 1991]. Force-directed layouts are, in essence, a simple physics simulation of like-charged particles representing the nodes. These particles act similarly to protons which experience Coulomb repulsion and try to get away from each other. If there is a relationship between two nodes, an attractive spring-like force is introduced, drawing the nodes back together.

In the case of a weighted graph, where each link (or relationship) has a value associated with it, we can adjust the spring coefficient of the attractive force to reflect this. This results in a layout where strongly connected objects are drawn together, and weakly connected ones further away. Uses for this type of representation have been shown biology, social networks, and with this thesis atmospheric chemistry [Muelder et al., 2014; Kohlbacher et al., 2014].

Next, we describe the Barnes-Hut algorithm, a mapping algorithm which builds a hierarchical tree of the data by splitting a plane into quartiles. This is used within the many force-directed graph layouts, including those of Force Atlas 2 and Yifan Hu, described shortly. Once this has been done a selection of four different layout algorithms shall be discussed.

Barnes Hut Algorithm

Since calculating the attractive/repulsive forces for each node of a large graph can be computationally intensive, many force-directed layouts rely on the Barnes-Hut approximation. This solves the Nbody problem of pairwise reactions between nodes, $O(n^2)$, by approximating long-range reactions by grouping such nodes and applying a single action on their centre of mass- reducing the computational time to $n \log n$.

To do this, first, a spatial index of each node is constructed (see below). This can either be done using a quadtree (2D) or octree (3D). Following this we calculate the centre(s) of mass, allowing us to approximate the repulsive forces of a force-directed graph.

Quadtree Construction: A quadtree is the recursive partitioning of two-dimensional space into a set of quadrants (a set of 4 squares). This process is repeated, with each square then being divided into

four itself, until there is only a single point within a cell. This converts a network, into a hierarchical tree representation of the nested quadrants in which each point resides (a quadtree), Figure 3.9.



(a) Methane Graph (b) Quadtree constructed from Figure 3.9a

Figure 3.9: **Demonstration of the formation for a quadtree from a force directed graph of Methane (including inorganics)**. (a) shows the force directed graph of Methane from which the quadtree has been constructed- edge colours represent the flux between species. Here we partition the area into 4 and start at the top-leftmost cell. This is then partitioned into 4 itself in a recursive process until there is only one point per cell. We repeat the process to any remaining cells in a clockwise manner (b). The hierarchical tree (b right) shows the containing structure for each node. Here the colours represent the order in which nodes have selected (starting at pink and ending in blue).

Having defined this, we move on to looking at the graph layouts.

Force Atlas 2

The force atlas two [Jacomy et al., 2014] algorithms is a force-directed layout designed primarily for scale-free² network spatialization. It is primarily designed for the use of networks consisting of 10 to 10,000 nodes and uses barns-hut approximation for the calculation of forces. Attractive forces are derived from the spring-electric model ($F_a = -k.d$), where k is the spring constant and d is the distance between the two nodes. Optional features for the graph include dissuasion by degree (separating nodes with a high number of total links/reactions), logarithmic attraction forces, adjustable gravity (attraction the centre of mass of the system to prevent disconnected components from drifting away) and collision detection to prevent overlapping nodes. Finally, an adaptive cooling scheme is applied, where the overall energy of a system is gradually decreased, allowing the nodes to settle into a low energy state.

 $^{^{2}}$ A network whose degree distribution follows a power law (7 degrees of separation). This is described in Chapter 4.

Yifan Hu

The Yifan Hu graph layout [Hu, 2004] is a multi-level graph drawing algorithm which uses the Barneshut algorithm with an octree layout. As with the force atlas algorithm, Yifan Hu also has an adaptive cooling aspect to it - meaning that as the algorithm is run its energy is progressively reduced, allowing the system to settle within a low energy state. The Yifan Hu algorithm uses a multilevel approach, running on first a course algorithm, and then refining the results - an efficient process which is unfortunately constrained to only working on undirected networks.

OpenOrd

A force-directed graph algorithm capable of scaling to large graphs [Martin et al., 2011]. OpenOrd uses simulated annealing (see below), which has five distinct phases. These are each run for a fraction of the total number of iterations and mimic the different states experienced when heating/cooling a physical object (liquid, expansion, cool-down, crunch and simmer) - hear each state describes the amount of energy assigned to the nodes within the force simulation. In addition to this, the OpenOrd algorithm applies a degree of edge-cutting to remove a percentage of edges experiencing the most stress within the physical system. This allows the network to open out into a more aesthetically pleasing layout.

Simulated Annealing

Most iterative layouts are updated interactively from some initial configuration in an attempt to reach the lowest energy state of the system. In most cases this results in a minimum configuration; however, this is generally a local minimum rather than the desired global minimum (the optimum low energy state of the entire system) [Davidson and Harel, 1996]. To overcome this, the work of Metropolis et al. [1953], which was later formulated in general terms by [Kirkpatrick et al., 1983], was used to lay the foundation for simulated annealing algorithms.

Annealing is usually used to describe the slow cooling applied to liquids for them to reach a crystalline (totally ordered, minimum energy) form. It can be shown that if the atoms(nodes) are cooled too rapidly (losing energy quickly and coming to a quick stop), they will form amorphous structures representing the local minima, as opposed to the desired global one. If cooled slowly, our graph is allowed to find a thermal equilibrium at every temperature. A slow cooling constant is applied, whilst occasionally supplying the system with short bursts of energy, that may allow it to overcome local minima.

tsNET

t-Distributed Stochastic Neighbor Embedding (t-SNE) is a dimensionality reduction technique which mimics the style of a force-directed graph (this is discussed in Subsection 6.3.3). tsNET³ is a graph drawing algorithm which leverages the non-linear dimensionality reduction capabilities of the t-SNE algorithm [Maaten and Hinton, 2008a]. This works by first computing the shortest-path distances between all nodes to produce a distance matrix. This distance matrix is then used to construct a cost matrix which consists of the sum of three terms:

- 1. A measure of the divergence between picking pairs of low- and high-dimensional data points.
- 2. A compression factor, known to reduce the t-SNE optimisation time, taken from [Maaten and Hinton, 2008b].
- 3. A repulsion term to prevent nodes clumping together.

Node positions are then determined by the minimisation of the cost matrix using gradient descent an optimisation algorithm used to minimise a function by iteratively moving in the direction of the steepest descent.

Note: Although tsNET makes for an excellent alternative to classical graph layouts, it does not take link direction into account.

3.3 Selecting The Best Graph Drawing Layout.

Subsubsection 3.2.1.1 explained the importance of removing overlapping edges and Subsubsection 3.2.1.2, the desire of having a well-distributed graph layout. This subsubsection builds on those criteria, assessing all the graph layouts described within this section (Mercator, Force Atlas 2, Yifan Hu, OpenOrd and tsNET). These all use the chemical mechanism representing species within the APHH campaign in Beijing [Fleming et al., 2017]. Here we look at the distribution (Subsection 3.3.1) and density (Subsubsection 3.3.1.3) as they affect a users ability to isolate the shortest path (fastest flux).

Force-directed graphs place a greater emphasis on node positions,

Criteria, such as the ability to isolate the shortest path (in this case the fastest flux), are essential in determining the usefulness of a graph. Comparing different layouts [Pohl et al., 2009] found 68% of user-chosen routes to reflect the shortest path between them.

 $^{^{3}\}mathrm{A}$ play on t-SNE and network.

This is due to the force-directed layout placing a greater emphasis on node positions and distance than other layouts. For comparison, the same study found this to be 40% for hierarchical layouts and only 2% for orthogonal ones. This section compares some graph layouts and their effect on user readability.

3.3.1 Graph-Node Distribution

Subsubsection 3.2.1.2 explained the importance of node distribution within a graph visualisation. Additionally, Purchase et al. [2003] explains that if we partition the viewing medium into quartiles, and populate each quadrant with an equal number of nodes (homogeneity), this drastically improved the usability and symmetry of a graph.

The problem is that for complex real-world graphs is that they often contain nodes with many reactions between them (Figure 3.10). Such regions of dense, indecipherable links (often referred to as hairballs [Ma and Muelder, 2013]), obscure nodes and edges within a region, making it impossible to read.

Methods such as edge pruning [Dianati, 2016] (removing unimportant links) can be used to reduce complexity. This process may be done either post computation (syntactic representation) - resulting in a loss of information, or during the algorithmic approximation (e.g. within the OpenOrd algorithm) - where any removed edges are then re-introduced in after the final layout has been generated.



Figure 3.10: A graph of the full MCM - a hairball. The high number of nodes and edges (especially those to inorganic species), causes a high degree of obfuscation, rendering the graph unusable. Species with a large number of reactions (links) are labelled.

3.3.1.1 Evaluating Node Distribution For The Beijing Mechanism

In deciding which layout algorithm produces the best graph-node homogeneity, a kernel density approach is used to compare node distributions across 2D space in Figure 3.11, and the Mercator density plot from earlier (Figure 3.8b).



Figure 3.11: Contour and kernel density plots showing the node distribution for different graph layouts. Line charts show the distribution of nodes in the x and y directions, while the contours represent density with respect to the location of each node (the crosses). Primary emitted species are coloured orange, and the darker contour polygons show areas of higher density.

This style of plotting allows for the easy location of areas containing a large number of nodes (high density) through the use of the contour-colour gradient. In an ideal graph, we would have groups (clusters) of high density, all of which would be evenly dispersed around the 2D plane. Additionally, the

use of x/y kernel density can show us the homogeneity of a graph - a perfectly homogeneous (lattice) graph will have a uniform distribution across both axes. For a modular graph of evenly dispersed groups, we would expect an oscillatory distribution of similar amplitudes. Using this criterion, the Mercator (Figure 3.8b), tsNet (Figure 3.11d) and Force Atlas (Figure 3.11b) score the highest, where the OpenOrd and Yifan Hu graphs containing a gaussian-esque distribution across both axes - a distribution conducive to the production of a hairball.

3.3.1.2 Distribution Of Primary Emitted VOCs

Within the construction of an atmospheric chemical mechanism, a chemist first begins with a primary emitted species. This is then broken down to produce other species, depending on its structure and functional groups (Figure 2.9). This process suggests that in constructing a network from such a mechanism, this structure will be prominent. Knowledge dictates that a chemical graph should start from a large emitted species, and aim towards carbon monoxide (and ultimately CO_2 although this is not included in the MCM). To show such a structure, we expect any primary emitted species to be evenly distributed and the chemistry to tend towards the location of CO (the centre). In searching for a layout that satisfies this requirement, the tsNET graph (Figure 3.11d) is found to be the best, followed by the OpenOrd and ForceAtlas2. Yifan Hu (Figure 3.11a) and Mercator (Figure 3.8b) both contain areas where many of the primary emitted (orange) species are grouped and are therefore unsuitable for the representation of the MCM structure.

3.3.1.3 Calculation Of Spatial Clustering

Subsubsection 3.3.1.1 explains that the ideal (modular) graph consists of many groupings of like chemistry evenly scattered across the graph. This requires a degree of regular anisotropy to produce 'clusters' of densely connected nodes, sparsely separated in space. To calculate this, we can rely on Voronoi tesselation.

Voronoi Tesselation

Voronoi Tesselation is the process of finding the largest area closest to a specific point. It can be thought of as a container with a bubble at the location of each node, where each bubble collapses to fill the largest area possible.

Here we begin by partitioning the graph plane into the same number of cells as our nodes. Next, each cell polygon boundary is calculated such that all the points on it lie closer to its seed (origin) node than any other. Mathematically these are referred to as the perpendicular bisectors of the lines between all points.

Finally, we calculate the areas of each polygon and use it to represent the density distribution between neighbouring nodes of our graph. To simplify the visual analysis of each graph, these are also coloured in Figure 3.12.

Visual study of node clustering

The method of using Vornouli tessellation for the calculation of density has been used in the study of neurones [Duyckaerts and Godefroy, 2000] and areas of fixation when viewing images [Over et al., 2006]. We apply it to the nodes of a set of force-directed graphs to determine the graph layout, which provided the best high-low density ratio for the atmospheric chemical mechanism of a Beijing environment.

Using Vornouli tessellation in Figure 3.12 we find that the OpenOrd and tsNET (Figure 3.12d, Figure 3.12c) layouts have the highest isotropy - containing cells of similar sizes and consequently a small colour gradient. This suggests that these are spatially efficient layouts as they do not reveal any additional information about nodes of similar chemistry.

In contrast, the Mercator layout, despite having a high x - y node distribution, contains large areas of unoccupied space due to its non-linear density distribution. Using this measure of spatial modularity, we find that the ForceAtlas2 and YifanHu (Figure 3.12b,Figure 3.12a) graphs have distinct modules of high density distributed across the entire graph, which is what we expected from the MCM network.



Figure 3.12: A visual analysis of node-cluster density using Voronoi tesselation. Each polygon is centred on a node - its area represents the space between the node and its nearest neighbours. Colours follow the normalised size of the Voronoi cells/polygons.

Mathematical Analysis And Layout Selection

In contrast to the qualitative analysis of the visualisation, it is also possible to calculate the distribution of polygon areas (and this node dispersion) of each graph through the use of a boxplot (Figure 3.13), in addition to the minimum, maximum and outlier properties, a boxplot allows for the comparison for the interquartile range (IQR) between graphs. If these values are large, they signify a more significant distribution between sparsely and densely grouped nodes (a commodity that is desired). The medians location within the IQR can also be used to indicate the dominant size of the polygon within the graph. Here a higher number of smaller polygons is desired and can be seen by a median approaching the lower box boundary (25^{th} quartile).



Figure 3.13: Voronoi $\log_{10}(\text{Area})$ BoxPlot for all plots in Figure 3.12. This provides a mathematical analysis for the areas around each node within a graph.

Figure 3.13 shows Mercator to provide the best result. However, it is noted that although the boxplot contains the best ratio, its radial shape is not conducive to the representation of modularity within a chemical network. tsNET contains the largest IQR, and since it is not able to handle the directed edges of an atmospheric chemistry graph, this again has to be ignored. Finally, although OpenOrd can reduce the number of hairballs within a graph by using simulated annealing and edge-cutting, its homogeneous isotropic node distribution (small IQR with a sizeable median value) make it not most effective at highlighting the structure of the MCM. Yfan Hu layout fares better with regards to the box plot, yielding an overall lower box, with a similar IQR and median ratio. Here its lower median suggests more high-density nodes, with a similar distribution to the ForceAtlas2. This makes

sense since the two algorithms share many similarities; however, the inability to handle directed edges makes it unsuitable for our application.

By process of elimination, this leaves the ForceAtlas2 algorithm as the best candidate for representing a chemical mechanism. Its directed nature, coupled with intuitive design make it applicable and easy to explain while maintaining the ability to produce a clear representation of any underlying structure. In addition to this, its uniform spatial distribution (Subsection 3.3.1) makes it a better candidate than the Yifan Hu graph (which fared slightly better in the boxplot).

3.4 Graph Semantics

Deciding the correct semantic (relating to meaning) representation for visualisation is often just as important as selecting the correct syntactic (structure) style. Semantic features are often applied post generation [Bennett et al., 2007] and are used to encode additional information and clarify the data. As a means of achieving both an aesthetically pleasing outcome, and an easy to understand visualisation, we must first consider what features we, or the reader, are most interested. Once this has been decided, we begin to explore various methods for representing them.

3.4.1 Limitations

Before selecting any semantic features, we must inform ourselves of the visual, cognitive and technological limitations of the visualisation, medium or user.

Visual

In visual analytics, the most significant bottleneck falls on the resolving power of the eye - this is known as an acuity (sharpness or clarity of vision at a distance). Acuities are a measure of the angle of an observed object with the viewer's eye using arcs (one arc minute equates to $\frac{1}{60}^{th}$ of a degree). This provides a unit of measurement for the total amount of information density we can feasibly perceive [Ware, 2013c].

In ophthalmology, there exist four types of acuities:

- detection: The smallest size an object can be whilst still being shown
- recognition: The smallest size an object can be to be recognised
- resolution: The smallest distance between two objects before they begin to merge

- localization: The smallest amount of visual change that can be measured between two objects

These provide a set of considerations which may be used to assess a visualisation. Depending on what encoding we use, it is possible to improve/hinder the reader's ability to perceive information, (Figure 3.14). An example of this would be that for a Macbook Pro retina screen⁴, where at 87 pixels/cm we can resolve at most 2 million resolvable nodes (at 57 cm from the screen). If we wished to add links between nodes, the total items identified is reduced to one million [Jankun-Kelly et al., 2014].



Figure 3.14: **Important acuities in visualisation.** Here a double prime " represents an arc minute which equates to an angle of 1/60 of a degree. A single prime ' is that of an arc second with is 1/60th of an arc minute (or 1/120 of a degree). For comparison the maximum angular resulution of the human eye is stated as 28 arc seconds [Deering, 1998] - this means we can only ever see up to 28 nodes or 2 verneers (disjoint lines) at any one time. Source: [Jankun-Kelly et al., 2014; Ware, 2013c]

Cognitive

Although it may be possible to distinguish 1 million nodes and links visually, interpreting and understanding these presents another problem. The visual thinking laboratories [VTL, 2019], have a range of publications exploring how presentation can improve though, cognition and communication between info-graphic and reader. Steven Franconeri [2018], explains that the time required to interpret a visualisation is directly related to the encoding used to highlight the data within it. Also 'intentional blindness'⁵ and misinterpretation are problems which are often occurred with poorly thought out encodings.

In considering the cognitive load of a visualisation Norman [2005] provides a list of three categories which should be explored:

1. Firstly, we have the visceral level, a subconscious process where decisions are made rapidly based on sensory inputs to the body. This is usually due to our inherent ability to locate patterns and

 $^{^{4}}$ A retina screen, is half the maximum possible resolution of the human eye at a 30cm distance. Additionally, the operating system interpolates in sets of 4 pixels, such that the image displayed may not be at full resolution.

 $^{{}^{5}}$ The failure of a user or audience to notice a fully visible feature because their attention was engaged by something else - e.g. misdirection in magic tricks.

changes due to semantic properties which shift the focus of the user.

- 2. Next follows the behavioural level (mostly subconscious). These are often learned reaction to changes noted as part of the visceral level. Here reactions may be honed on and influenced by past experiences and events.
- 3. Finally, we reach the reflective level. Here the user collates all sensory input from the previous two levels and makes an informed conclusion about the underlying data. Conclusions drawn here can be used to bias the methods used within the behavioural level in future events.

Technological

In addition to human limitations, there may be restrictions due to the medium a visualisation is created/presented. To monitor resolution, much scientific research is constrained by the size, resolution and colour quality of the presentation mediums used for talks, printing or posters. Ware [2013c] explains that a printer capable of producing 1200 dots per inch squared, can only do this for black/white binary images. If for instance, 256-greyscale is used, the resulting resolution is then at-least ten times smaller. This is because printers a Monet (dot matrix) style approach to create shading and colour. It follows that at full colour⁶, the output resolution will be worse.

It is also essential to have a graph fitting the same overall shape of the canvas on which it is presented [Taylor and Rodgers, 2005]. This not only makes optimal use of any space available but also reduces the visual complexity as it minimises the number of distinct shapes available to the user.

3.4.2 Node Encoding

Within a graph, the nodes or vertexes are a representation for the set of items we have an interest. In addition to the relationships between them, items often contain a multitude of features which describe them. Examples of these may be useful information for a person, categories of characters or the chemical composition/concentration for a species in the MCM.

Each of these additional properties can contain valuable information for the interpretation of the graph, and the interactions between nodes. It is, for this reason, that graph convoluted neural networks [Klicpera et al., 2018] require a 'feature matrix' describing each node, in addition to the network structure and edge weightings.

This subsection addresses several ways in which additional information can be encoded in the representation of a node, Figure 3.15. Each different category colour matches the title description in the

⁶CYMK (Cyan Magenta Yellow Black)
text.



Figure 3.15: A graph showing 5 different node encoding methods. These are Circle Attributes (red), Chemical structure (blue), Species Name (green), External Labels (maroon) and interactive selection (orange). The network shows the Common Representative Intermediate species [Jenkin et al., 2008] mechanism. Node colours represent primary emitted VOCs (red), MCM species (orange) and lumped CRI-only species (blue).

Circle Attributes

The simplest of these range from the use of colour and stroke (outline) to shape/size as a way of indicating a group. Here it is possible to provide information such as a species concentration based on its size, its importance with its colour, its degree with its opacity and its category with its stroke colour [Ware, 2013a,b]. Such decisions depend on what properties the user is trying to show. For instance, red species in Figure 3.15 are primary emitted VOCs, orange species exist between both the MCM and the CRI (see figure caption) mechanism, and blue ones are lumped species which do not appear as part of the MCM.

Chemical Structure

Traditional chemical diagrams use the chemical structure to depict the node (Figure 3.2). This makes it intuitive to extract information about the functional group and potential bond changes within species. Such a method of representation, is indeed useful, however when visualising hundreds, or thousands of nodes on a page, it results in occlusion or labels being too small to resolve visually.

Species Name

Much like the chemical structure, a species name is proven useful in explaining to the user its chemical structure or properties (often due to prior knowledge, or the ability to look this up). Unfortunately, since names have differing lengths, this can cause problems, especially with large numbers of closely located nodes. A solution to this may be to adjust the font size to fit in within the circle radius of the node. However this does come with its problems - for instance, tiny nodes may have text smaller than a pixel, or the misleading notion that longer names are less important since they are represented by a smaller font.

Interactivity

Ben Shneiderman, one of the first, and most prominent, researchers in human-computer interaction coined the phrase 'overview first, zoom and filter, details on demand' [Shneiderman, 1996]. This is the philosophy behind most data orientated interaction design and can be applied to graphs. One example is the selection only of reactions relative to a node of interest, as is shown in Figure 3.16.

For complicated systems, interactivity plays a vital role in unravelling complexity and reducing clutter [Shneiderman, 1997]. This is a method which lessens the cognitive load (Figure 3.4.1) on the user, allowing them to query only items of interest, while still displaying all the information in a single location [Görg et al., 2007]. This fits in Ben Shneiderman [1985]s 8th rule of interface design, which explains that people only ever remember *'seven plus or minus two chunks'* of information.

A comprehensive list of all available interaction types and styles are provided by Wybrow et al. [2014]. Some examples of interaction are:

81

Hi-lighting

- Hovering
- Brushing and Linking
- Magic Lenses (see hidden objects)

Navigation

- Pan / Zoom
- View Distortion (fisheye)

Visual Structure-Level Interaction

- Selection
- Changing layout/mapping attributes
- Changing representation

Data Level Interactions

- Adding / Filtering
- Search / Query



Figure 3.16: Using mouseover edge-selection to highlight all links related to a node. This figure shows how in using interactivity it is possible to reduce clutter and filter the information presented by a densely populated graph. In this case, the Mercator projection (Subsubsection 3.2.2.2) is used, with reactions relating to Carbon Monoxide (centre) highlighted. Orange lines represent reactions producing CO whist the red (some of which may be hidden) are of reactions with CO.

External Labeling

In cases where interactivity is not possible, such as papers, books and this thesis, an alternative approach to data selection has to be employed. Here nodes which are central to the explanation of a certain point are filtered by the author and displayed through the use of external labels. It is found that having links at 45 and 90-degree angles (such as in transport maps) lead to a clearer layout and better distinction from the links already within the graph. Automatically generated labels within the thesis are made using Lu [2019].

3.4.3 Edge Properties

Defining the purpose of force-directed (graph-energy) models as a means for creating a visualisation from which the viewer can infer properties of the data [Noack, 2004], it can be shown that this criterion is easily met in small and sparse graphs. However, non-planar examples with high edge density (lots of links) can easily result in tangled results with impractical running times [Kumar and Garland, 2006]. In most cases attaining an optimal solution here seems to be computationally infeasible [Davidson and Harel, 1996]. This is generally because graphs primarily focus on hi-lighting a specific purpose or following a set of aesthetic heuristics [Pohl et al., 2009].

3.4.3.1 Muti-Variate Edges

Since there are multiple relationships between species, it is important to decide if simplifying the network would be of benefit. Although it is possible, multiple edges may cause unnecessary clutter for larger networks. Instead, it is often useful to simplify the number of edges in the network and encode the edge properties within the vector object. This allows the user to retrieve any additional information by hovering over the edge or connecting nodes, as required. Should the topic of interest require a specific property, then it would also be possible to remove, or hide, all edges which do not contain it. This produces an interactive graphic containing all the required information, as and when needed, without the unnecessary clutter of having every reaction shown.

3.4.3.2 Edge Direction

When using a directional graph, it is the convention to use arrowheads to represent the direction of flow. However, in high-density regions, it is often found that arrowheads take up precious real estate in the drawing area [Dwyer et al., 2006a]. As an alternative, colour and line-type can be used to represent the direction instead - this was seen in Figure 3.9a (the quadtree example). This example can be shown in the routing networks presented by [Di Battista et al., 2004]. One example applicable for chemistry would be the use of dashed lines to represent mono-directional relationships and continuous lines for bi-directional ones.

3.4.3.3 Edge Shape

Edge shape is essential, as it is the medium we use to represent relationships within a graph. For orthogonal graphs (Subsubsection 3.2.1.1), multiple lines are used to simplify the complexity of a graph and reduce edge crossings [Di Battista et al., 1994]. In increasing the number of lines within an edge, this multi (poly)-line graphs can be modified to give drawing with nicely curved edges. Similarly, it is possible to replace the edges in a straight-line graph with Lombardi-style curves or cubic beziers [Chernobelskiy et al., 2012; Goodrich and Wagner, 1998].

Using a mechanism describing the reactions of Butane in the MCM, we compare many edge shapes (Figure 3.17). In this graph, each node has multiple edges between nodes (multi-link). Each edge represents a different reaction between the species. Figure 3.17 shows the straight-line representation of each graph. Since each edge represents the shortest distance between two nodes, any additional reactions pairs between similar nodes are hidden from view. If using this type of representation, it is advised to take the net edge direction and weight between the values. An improvement to the straight-line graph comes from the use of quadratic curves (Figure 3.17b). This allows for the symmetric distribution of edges within the graph. If instead an asymmetric representation for each edge is required, it is possible to use a bezier curve (described below) instead of a quadratic edge (Figure 3.17c). These can provide additional information through the use of control points, which can alter the shape, steepness and asymmetry of each link.

Bezier Curves

Bezier curves are named after Pierre Bezier who used them in the bodywork design of Renault cars in the 1960s [Hazewinkel, 1997]. Since then they have been widely used in graphs, computer graphics, font design and animation/interactivity response [Goodrich and Wagner, 1998; Hazewinkel, 1997; Mortenson, 1999]. Bezier curves come in a range of possible dimensions; cubic beziers are the most commonly used within network visualisation. These contain four control points, respectively, which can be used to determine the shallowness of the curve through design. In general, relatively shallow curves are prefered, as these do not introduce unnecessary edge crossing or abrupt changes, which have been shown to hinder a users ability to isolate items of interest [Purchase et al., 2003].



(c) Bezier (multi-edge)

Figure 3.17: A selection of edge shapes for the butane network. These show linear (a), quadratic (b) and bezier (c) edge shapes for the same network. In general, the bezier curves appear to provide the shapes of the most aesthetically pleasing graphs.

3.4.3.4 Edge Bundling

Pioneered by Holten [2006], edge bundling techniques are an effective way to reduce visual clutter. Much like a force graph, edges are represented as a string of lined points. This allows for edges to be pulled together (attracted to one another) and produces a visualisation akin to moving water droplets on a hydrophobic surface. Figure 3.18 shows how in changing the amount of attraction between edges, it is possible to reduce clutter in a visualisation.









Figure 3.18: How the compatibility threshold affects edge bundling using the Mercator graph from Subsubsection 3.2.2.2. In increasing the amount, edges are attracted (θ) it is possible to improve the clarity of a graph. However, there reaches a point where this distortion can worsen the result, confusing the reader, or creating a false positive. For this reason, I generally use only a slight bundling value > 0.7.

3.4.3.5 Power, Routing And Confluence.

Confluent graphs use a graph drawing method in which edges are not drawn as individual distinguishable geometric objects, but rather as a crossing free system of arcs and junctions. [Förster et al., 2019]. Their design is similar to that of the edge bundling algorithm, except that rather than bundling edges spatially (a design which may introduce ambiguity), the bundling is done based on connectivity and can help reduce clutter by grouping multiple edges where the all target nodes are also connected to all the source nodes (Figure 3.19) [Bach, 2020].



Figure 3.19: An example of confluent bundling. A traditional network (a), Edge bundling (b), Power Graph (c) and Confluent graph (d) representations. Source: [Bach, 2020]

Using the oxidation of butane as an example, we shall explore the construction process of a confluent. Starting with the network presented in Figure 3.17, we create a power graph - power graphs are a representation of complex networks where sets of items identical source and target links are lumped or grouped within a single item. Next multiple edges which follow the same path are bundled through a 'routing' node to create the routing graph in Figure 3.20. The newly created routing nodes are now used as control points for the mapping of the graph using basis-splines⁷ (Figure 3.21). Finally, any crossing links are removed, leaving the confluent graph in Figure 3.22.

Confluent drawings have been found to have many applications (e.g. the ego-centric author network

⁷These are similar to be irr curves but require a degree (p), n + 1 control points, and a knot vector of m + 1 points. Knots are the things that make the curve continuous

and social interaction graph), they generally perform best in sparse networks with locally dense clusters of a tree-like structure [Bach et al., 2017]. Although sparse, the cyclic nature of atmospheric chemistry does not allow for a sufficient reduction in complexity to make them a suitable improvement over traditional graphs. The use of very close-fitting basis-splines in addition to a routing graph (confluent graph with crossing artefacts), may, however, help to simplify specific layouts or mechanism subsets with a certain amount of tweaking.



Figure 3.20: The routing graph of the butane mechanism. Here paths which contain two or more bundles have an extra 'routing' node introduced (orange stroke)



Figure 3.21: Confluent graph with crossing artifacts. The routing graph with the addition of basis-splines using the orange routing nodes in Figure 3.20 as control points.



Figure 3.22: Confluent graphs without crossing artifacts. The remaining confluent graph with crossing edges removed.

3.4.3.6 Angle / Continuity

Visual representation utilises our conscious and unconscious pattern recognition and intuition abilities [Dixon, 2012]. To avoid apophenia (finding patterns where they do not exist), careful consideration has to be placed in the design of a graph layout. Although edge crossing is often thought of as the most import aesthetic metric, finding continuity between incoming and outbound edges of a node was found to be if equal importance [Ware et al., 2002].

Reducing the angle between related edges increases readability and allows the behavioural process to infer information about a graph correctly. This process can be compared to predicting the direction of turbulent vs laminar flow. In addition to this edges should be spaced evenly around the node, maximising the minimum-edge-angle between all edges of a node [Bennett et al., 2007].

3.4.4 Temporal Projection

Story-telling has been an effective method to convey information, experience and cultural values for almost as long as people have been around. Many real-life physical processes occur over time and thus allow the use of a story-telling analogy. Gershon and Page [2001] provides a generic structure which begins with creating a general overview of the subject. Events are then animated in order of occurrence and defined as we go along. Finally, any remaining conflicts and uncertainty are addressed, and these are rectified. Using this as a template for our graphs, we find that the content is usually given in the form of a title or figure description, the evolution as the visualisation, and finally the reflection and resolution through the use of user interaction (e.g., node hi-lighting, zoom or animation).

Since very few graph layouts support dynamic time-varying graphs [Kumar and Garland, 2006], several methods of visualising temporal events have been developed. Although storylines can be useful for drawing the evolution of simple systems, these break down when dealing with large numbers of dependant variables. Force-directed layouts may be adapted, to suit these better, after which the initial positions of the previous node endpoints are used as the initial positions for consequential simulations. Three methods of representing these are shown in [Ellis, 2018] - screenshots of which are shown in Figure 3.23.



Figure 3.23: Film style representation of temporal changes in a network. Showing the temporal changes from a model simulation of the Beijing atmosphere. (a) shows a weighted graph at midnight. With the addition of daylight, the chemistry speeds up, causing the force graph to contract, changing the overall network shape (the faster reactions have a stronger attractive force). The animation of this can be found at https://github.com/wolfiex/DanEllisThesis/blob/master/daynight_26mb.gif

Finally, user-interaction such as hi-lighting key nodes/links, zoom and animation⁸ may be used to clarify information at the reflection stage.

 $^{^{8}}$ [Archambault et al., 2014] notes that animation poses high demands on the user's visual memory and that snapshots are likely to miss underlying patterns. For this reason, interactive techniques that can allow retrospective selection of timesteps allows for a good compromise between these.

3.4.5 Additional Dimensions

Additional dimensions can be used to emphasise certain aspects of our graphs. For instance, multiple layers may be used in a directional graph to separate the importance of the nodes [Dwyer et al., 2006b]. Figure 3.24 shows the first, second and third-generation species of a mechanism containing isoprene in three dimensions, where each layer in the third direction represents a different generation of species. Such visualisation may be explored interactively, with the aid of a computational input device (a mouse, keyboard or device gyroscope), or with the aid of red-cyan 3D glasses (for non-interactive mediums such as print).

Different layers can be used to separate primary VOCs, from species which result in their production (+1 layers) and loss (-1 layers). Temporal data (e.g. Figure 3.23) can also be presented in this format. The only drawback is the high possibility of obfuscation which may result from many layers of overlapping information.



Figure 3.24: A 3D representation of a graph to hilight certain features. The first, second and third generation species of isoprene shown as an interactive 3D anaglyph.

3.4.6 Summary Of Semantic Representation

In this section, the different semantic methods have been explored. It is found that additional information such as concentration or functional groups may be represented as node size/colour and that there are a range of edge plotting styles that may be used to reduce clutter.

In the next section, we combine both semantic and syntactic representation and apply it to an atmospheric chemical mechanism. Using the graph visualisation tools described above, we access the quality of information which may be inferred through representing a mechanism in this way.

3.5 A Chemistry Case Study

To conclude, we apply many of the tools described above to a single case study. We select an MCM subset representing the VOCs measured as part of the APHH campaign in Beijing [Fleming et al., 2017]. The chemistry of the mechanism is initiated using the conditions in Table 4.4, and propagated by the Dynamically simple model of atmospheric chemical complexity (DSMACC) [Emmerson and Evans, 2009; Ellis, 2020]. We run this forwards to steady-state and extract the flux between species on noon. The edge weight is the net flux (product of the species concentration multiplied by the rate of reaction for all reactions between those species and products), normalised to a value between 1 and zero.

3.5.1 Semantic And Syntatic Considerations.

3.5.1.1 Syntatic Representation

Since we shall be using simulation data, we require a layout which deals with both direction and edge weights. The spring-like description of the ForceAtlas2 is chosen from Section 3.2. This feature highlights fast reactions by bringing nodes together. Such a property has been observed to help users select the shortest path within a network [Pohl et al., 2009]. Here users picked the shortest path an average of 68% for force-directed graphs, compared to 40% for hierarchical and 2% for orthogonal layouts. Such properties can help us locate any trends in fast reactions which may control the chemistry within a system.

3.5.1.2 Example Semantic Representation Using A Methane Mechanism.

Since the graph generated by the methane mechanism contains only a handful of species, our screen real-estate allows the listing of names for each node. Node sizes are scaled to represent the concentration of each species at that time point, and edges are coloured to represent the strength of each relationship between them. Here pink edges represent a fast-flux and blue ones a slow one. In comparing the change in graph shape between a weighted Figure 3.25b and unweighted Figure 3.25a graph, we can see that nodes connected by a high edge weight (fast-flux) are drawn closer to each other than those with a slow flux.



Figure 3.25: A weighted and unweighted force diagram of the methane mechanism. Here it is seen that upon weighting, edges with a larger flux (pink) are drawn closer than those of a weaker one (blue).

3.5.2 A Model Of Beijing

To perform a sensitivity study on the initial positions of nodes within the force atlas algorithm a graph consisting of links and weightings is constructed using a box model simulation of the Beijing summer environment (mid-day) and feed it the gephi software [Bastian et al., 2009] - an open-source software designed for the exploration of networks. We then script the java code to perform the functions in Figure 3.26. As part of this, nodes are initiated with a random position; the ForceAtlas2 layout is then run and then the graph is rotated and translated such that it is centred around carbon monoxide and has a 45-degree angle between this and formaldehyde. This step constrains the general orientation of the graph, allowing us to analyse the generated graphs for global and local minima. The final step is to save a copy of the generated graph layout and repeat to generate a data set, a subset of which is shown in Figure 3.27. These are discussed further in Subsubsection 3.5.2.1.



Repeat x 2000

Figure 3.26: A flow chart of the process performed by the custom gephi script used to generate the data set

*	×	*	: *	*	×.	*	*	*	*	*	*	*	×	*	*
×	*	×	***	×.	*	*	*	*	*	*	*	*	-	**	×
*	*	*	*	*	°₩¢	*	*	*	*	*	*	*	*	. *	*
*	*	*	*	×	*	**	*	*	*	*	*	×	*	*	*
*	*	*	*	*	*	*	*	*	`₩	*	*	*	*	*	*
*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*
*	*	*	*	*	*	*	*	*	· **	*	*	*	*	*	*
*	*	×	लंग		*	*	*	*	*	*	*	*	**	**	*
-	**	*	-	*	*	×	*	*	*	*	*	*	*	*	*
*	¥	*	*	*	*	*	*	***	***	*	*	*	*	*	*
*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*
*	*	*	*	*	*	*	×	*	*	×	*	*	*	*	*
*	*	*	*	*	*	*	*	*	*	×	*	*	*	*	*
*	*	*	*	*	*	*	*	-	×	*	*	*	*	*	*

Figure 3.27: A sample of 224 (out of the 2000) graphs generated using the ForceAtlas2 algorithm. These represent the conditions of a spun up simulation of Beijing at noon. The shapes of each graph, and general shapes are discussed in Subsubsection 3.5.2.1 and Subsubsection 3.5.2.2.

3.5.2.1 Similarity Between Graph Shape

Although through the use of manual intervention, it is possible to perform a superficial level of shape analysis, our cognitive capabilities do not allow us to perform this task for all the simulations of Figure 3.27- less so the entire 2000 graphs in the dataset. To overcome this problem, we rely on a method of machine learning called t-Distributed Stochastic Neighbor Embedding (t-SNE) - described in Subsection 6.3.3 and is the foundation of the tsNET layout algorithm. This is a dimensionality reduction technique used in the automatic categorisation of images or photographs [Stefaner, 2020; Sangkloy et al., 2016].

The input for the t-SNE for each dataset is a flattened (1 dimensional) representation of the pixels in the image - we start and by taking a binary matrix representing each image, split it up into rows, and glue these together. The pixelmap for each image is then fed into the t-SNE algorithm from the Scikit Learn package [Pedregosa et al., 2011]. This reduces the logical list of pixels for each image into a two-dimensional representation of their similarity. We plot each file, for its (x, y) coordinate, and isolate clusters of similarity using density contours in Figure 4.9.



Figure 3.28: A normalised scatter plot of 2D space produced by the t-SNE algorithm. Each triangle represents a different arrangement of the MCM nodes shown in Figure 3.27, and the colours/density contours show the regions in which we find similar images/graphs. Cluster numbers correspond to the groups in Figure 3.29.

Using interactivity and/or vector cluster detection techniques, it is possible to examine which files contribute to an area of high density. Figure 3.29 shows a sample of four graphs from each of the four corresponding clusters. Although individual node locations may vary, patterns on the macro scale start to emerge, with similar groups exhibiting symmetrical symmetry, e.g. groups 1/2 and 3/4. A constraint in the overall degree of freedom of the network can be attributed solely to its structure, and consequently the chemistry which forms this. The non-random nature of the produced graph layouts means that it would be possible to juxtapose a variety of mechanisms using the ForceAtlas2 layout.



Figure 3.29: A selection of graphs for each of the labeled groubs in Figure 3.28. These reveal that symmetric similarity between like-positioned points within the t-SNE output.

3.5.2.2 Network Branch Classification

In Subsubsection 3.5.2.1 it was seen that there exist a certain branch pattern that emerges from the structure of the MCM (Figure 3.29). Upon manual inspection of the simulations (Figure 3.27) many graphs appear to contain three branches for each graph - using this it may be hypothesized that these are a result of the mechanism, and by consequence the chemistry it describes.

To test for this, we categorise all primary emitted species into Alkanes, Alkenes, Aromatics and Terpenes. All nodes and links in close proximity are regarded as products of these species and are placed within the same group. Using a randomly selected graph from the dataset, the network is separated spatially, and nodes within the Voronoi cell (These are described in Subsubsection 3.3.1.3) of a primary emitted species are coloured similarly.

Figure 3.30 shows a split in the MCM chemistry for the Beijing mechanism. Here it is found that it is possible to separate the MCM network into an aromatic branch, a terpene branch, an alkane and straight-chain alkene branches. Such branches not only help us identify changes in chemistry due to biogenic or anthropogenic sources but also emphasise the path taken to carbon dioxide and water. As the MCM does not contain CO_2 , we see all the different groups converge on Carbon Monoxide at the centre of the figure (the white dot). Coupled with the last section, this suggests that following rotational and symmetric transformations, it is possible to compare different mechanisms.



Figure 3.30: Highlighting the groups of species, and their products within one of the MCM network graphs from Figure 3.27 These are Aromatics (gold), Terpenes (turquoise) and Alkane/Alkene carbon chains (red/blue)

3.6 Conclusion

Chapter 2 explained that the visual representation of data could make use of the pattern recognition side of the human brain. Similarly to the invention of cuneiform, we can use graphics to alleviate some of the cognitive strain from numerical data. It was noted that metaphor selection and storytelling are an essential part of conveying complex information to the reader and that the choice of encoding plays an integral part in this.

This chapter begins on building on those concepts of visualisation. We defined our system as a collection of relational interactions between species and chose the graph/sociograph design to represent these.

Next, we explored the different methods of graph design. These were semantic (meaning/design) and syntactic (structure). In the syntactic category, we found that the use of a force-directed graph provides a system most familiar to the reader. In addition to this, the ForceAtlas algorithm helped to produce a mathematical representation (practicality) of the relationships between the MCM network, while reducing the amount of clutter within the graph (visual aesthetics). In semantic design, it was noted that information about concentration, or functional groups might be represented within the network. Interactivity was found to be useful where additional information did not need to be provided but may be enquired by at a later point in the analysis. Edge shape and design was also explored. Here a confluent graph was seen to produce the easiest to understand the structure but was the most difficult to implement. The next most useful method was edge bundling, which was used in Figure 3.30, and future work.

Although graph layouts have a range of local minima, the overall network structure of the MCM is constrained by its construction protocol (due to the allowed chemical reactions) and thus can be used to produce comparable graphs. This method of visualisation, in combination with interactive querying techniques, can aid in the comparison and understanding of large/complex chemistry simulations. This can be particularly useful in the explanation of specific interactions within a mechanism, or the exploration of temporal changes within a box-model simulation.

The next chapter builds on the use of graphs in situations where visualisation may not be possible - for example, automatically generated graphs consisting of billions of species and reactions. To do this, we apply a series of graph metrics that allow the classification and ranking of graphs, and the nodes (species) within them.

Bibliography

- Archambault, D., Abello, J., Kennedy, J., Kobourov, S., Ma, K.-L., Miksch, S., Muelder, C., and Telea, A. C. (2014). *Temporal Multivariate Networks*, pages 151–174. Springer International Publishing, Cham. http://dx.doi.org/10.1007/978-3-319-06793-3_8.
- Aumont, B., Szopa, S., and Madronich, S. (2005). Modelling The Evolution Of Organic Carbon During Its Gas-Phase Tropospheric Oxidation: Development Of An Explicit Model Based On A Self Generating Approach. Atmospheric Chemistry and Physics, 5:2497–2517. https: //www.atmos-chem-phys.net/5/2497/2005/acp-5-2497-2005.pdf.
- Bach, B. (2020). Confluent Graphs. https://aviz.fr/~bbach/confluentgraphs/.
- Bach, B., Riche, N. H., Hurter, C., Marriott, K., and Dwyer, T. (2017). Towards Unambiguous Edge Bundling: Investigating Confluent Drawings For Network Visualization. *IEEE transactions on vi*sualization and computer graphics, 23(1):541–550. http://dx.doi.org/10.1109/TVCG.2016.2598958.
- Baronchelli, A., Ferrer-i Cancho, R., Pastor-Satorras, R., Chater, N., and Christiansen, M. H. (2013). Networks In Cognitive Science. Trends in cognitive sciences, 17(7):348–360. http://dx.doi.org/10. 1016/j.tics.2013.04.010.
- Bastian, M., Heymann, S., and Jacomy, M. (2009). Gephi: An open source software for exploring and manipulating networks. AAAI. http://www.aaai.org/ocs/index.php/ICWSM/09/paper/view/154.
- Ben Shneiderman (1985). The eight golden rules of interface design. http://www.cs.umd.edu/~ben/goldenrules.html.
- Bennett, C., Ryall, J., Spalteholz, L., and Gooch, A. (2007). The Aesthetics Of Graph Visualization. In Computational Aesthetics in Graphics, Visualization, and Imaging. The Eurographics Association.
- Bergwerf, H. (2019). Molview. http://molview.org/.
- Bostock, M. (2012). D3.js data-driven documents. http://d3js.org/.
- Chernobelskiy, R., Cunningham, K. I., Goodrich, M. T., Kobourov, S. G., and Trott, L. (2012). Force-Directed Lombardi-Style Graph Drawing, pages 320–331. Springer Berlin Heidelberg, Berlin, Heidelberg. http://dx.doi.org/10.1007/978-3-642-25878-7_31.
- Davidson, R. and Harel, D. (1996). Drawing graphs nicely using simulated annealing. ACM Trans. Graph., 15(4):301–331. http://doi.acm.org/10.1145/234535.234538.
- Deering, M. F. (1998). The Limits Of Human Vision. In 2nd International Immersive Projection Technology Workshop, volume 2. https://www.swift.ac.uk/about/files/vision.pdf.

- Di Battista, G., Eades, P., Tamassia, R., and Tollis, I. G. (1994). Algorithms for drawing graphs: An annotated bibliography. *Comput. Geom. Theory Appl.*, 4(5):235–282. http://dx.doi.org/10.1016/ 0925-7721(94)00014-X.
- Di Battista, G., Mariani, F., Patrignani, M., and Pizzonia, M. (2004). Bgplay: A System For Visualizing The Interdomain Routing Evolution, pages 295–306. Springer Berlin Heidelberg, Berlin, Heidelberg. http://dx.doi.org/10.1007/978-3-540-24595-7 27.
- Dianati, N. (2016). Unwinding the hairball graph: Pruning algorithms for weighted complex networks. *Phys. Rev. E*, 93:012304. https://link.aps.org/doi/10.1103/PhysRevE.93.012304.
- Dick Derwent, Andrea Fraser, John Abbott and Mike Jenkin (2010). Evaluating The Performance Of Air Quality Models. online. https://uk-air.defra.gov.uk/assets/documents/reports/cat05/1006241607 100608 MIP Final Version.pdf.
- Dixon, D. (2012). Analysis Tool Or Research Methodology: Is There An Epistemology For Patterns?, pages 191–209. Palgrave Macmillan UK, London. http://dx.doi.org/10.1057/9780230371934 11.
- Duyckaerts, C. and Godefroy, G. (2000). Voronoi tessellation to study the numerical density and the spatial distribution of neurones. *Journal of Chemical Neuroanatomy*, 20(1):83 – 92. http: //www.sciencedirect.com/science/article/pii/S0891061800000648.
- Dwyer, T., Koren, Y., and Marriott, K. (2006a). Drawing directed graphs using quadratic programming. *IEEE Transactions on Visualization and Computer Graphics*, 12(4):536–548.
- Dwyer, T., Koren, Y., and Marriott, K. (2006b). Ipsep-Cola: An incremental procedure for separation constraint layout of graphs. *IEEE Trans. Vis. Comput. Graph.*, 12(5):821–828.
- Dwyer, T., Marriott, K., and Stuckey, P. J. (2006c). Fast Node Overlap Removal, pages 153–164. Springer Berlin Heidelberg, Berlin, Heidelberg. http://dx.doi.org/10.1007/11618058 15.
- Eades, P. (1984). A heuristic for graph drawing. In proceedings.
- Ellis, D. (2018). Animation Of The Evolution Of Chemistry Graph Of Beijing. https://github.com/wolfiex/DanEllisThesis/blob/master/daynight_26mb.gif.
- Ellis, D. (2020). DSMACC-Testing. https://github.com/wolfiex/DSMACC-testing.
- Emmerson, K. M. and Evans, M. J. (2009). Comparison of tropospheric gas-phase chemistry schemes for use within global models. *Atmospheric Chemistry and Physics*, 9(5):1831–1845. https://www. atmos-chem-phys.net/9/1831/2009/.

- Fleming, Z. L., Lee, J. D., Liu, D., Acton, J., Huang, Z., Wang, X., Hewitt, N., Crilley, L., Kramer, L., Slater, E., Whalley, L., Ye, C., and Ingham, T. (2017). Dataset Collection Record: Aphh: Atmospheric Measurements And Model Results For The Atmospheric Pollution & Human Health In A Chinese Megacity. https://catalogue.ceda.ac.uk/uuid/648246d2bdc7460b8159a8f9daee7844.
- Foo, B. (2019). Memory Underground Convert Your Memories Into A Subway Map Home. http: //memoryunderground.com/.
- Friedrich, C. and Schreiber, F. (2004). Flexible layering in hierarchical drawings with nodes of arbitrary size. In *Proceedings of the 27th Australasian Conference on Computer Science - Volume 26*, ACSC '04, pages 369–376, Darlinghurst, Australia, Australia. Australian Computer Society, Inc. http: //dl.acm.org/citation.cfm?id=979922.979966.
- Fruchterman, T. M. J. and Reingold, E. M. (1991). Graph drawing by force-directed placement. Software: Practice and Experience, 21(11):1129–1164. https://onlinelibrary.wiley.com/doi/abs/10. 1002/spe.4380211102.
- Förster, H., Ganian, R., Klute, F., and Nöllenburg, M. (2019). On strict (outer-)confluent graphs.
- García-Pérez, G., Allard, A., Ángeles Serrano, M., and Boguñá, M. (2019). Mercator: Uncovering Faithful Hyperbolic Embeddings Of Complex Networks. *arxiv*. http://arxiv.org/abs/1904.10814.
- Gershon, N. and Page, W. (2001). What storytelling can do for information visualization. Commun. ACM, 44(8):31–37. http://doi.acm.org/10.1145/381641.381653.
- Goodrich, M. T. and Wagner, C. G. (1998). A Framework For Drawing Planar Graphs With Curves And Polylines, pages 153–166. Springer Berlin Heidelberg, Berlin, Heidelberg. http://dx.doi.org/ 10.1007/3-540-37623-2 12.
- Görg, C., Pohl, M., Qeli, E., Xu, K., Ebert, A., and Meyer, J. (2007). Visual Representations, pages 163–230. Springer Berlin Heidelberg, Berlin, Heidelberg. http://dx.doi.org/10.1007/ 978-3-540-71949-6_4.
- Harari, Y. (2015). Sapiens: A Brief History Of Humankind. Harper. https://books.google.co.uk/ books?id=FmyBAwAAQBAJ.
- Hazewinkel, M. (1997). Encyclopaedia Of Mathematics: Supplement. Number v. 1 in Encyclopaedia of Mathematics. Springer Netherlands. https://books.google.co.uk/books?id=3ndQH4mTzWQC.
- Holten, D. (2006). Hierarchical edge bundles: Visualization of adjacency relations in hierarchical data. *IEEE Transactions on Visualization and Computer Graphics*, 12(5):741–748. https://doi.org/10. 1109/TVCG.2006.147.

- Hu, Y. (2004). Efficient, High-Quality Force-Directed Graph Drawing. *web.* http://yifanhu.net/PUB/graph_draw.pdf.
- Jacomy, M., Venturini, T., Heymann, S., and Bastian, M. (2014). Forceatlas2, A Continuous Graph Layout Algorithm For Handy Network Visualization Designed For The Gephi Software. *PloS one*, 9(6):e98679. http://dx.doi.org/10.1371/journal.pone.0098679.
- Jankun-Kelly, T. J., Dwyer, T., Holten, D., Hurter, C., Nöllenburg, M., Weaver, C., and Xu, K. (2014). Scalability Considerations For Multivariate Graph Visualization, pages 207–235. Springer International Publishing, Cham. http://dx.doi.org/10.1007/978-3-319-06793-3 10.
- Jenkin, M., Watson, L., Utembe, S., and Shallcross, D. (2008). A common representative intermediates (cri) mechanism for voc degradation. part 1: Gas phase mechanism development. Atmospheric Environment, 42(31):7185 – 7195. http://www.sciencedirect.com/science/article/pii/ S1352231008006742.
- Jenkin, M. E., Saunders, S. M., and Pilling, M. J. (1997). The Tropospheric Degradation Of Volatile Organic Compounds: A Protocol For Mechanism Development. Atmospheric environment, 31(1):81–104. http://www.sciencedirect.com/science/article/pii/S1352231096001057.
- Johnson, S. (2010). Where Good Ideas Come From. Penguin Publishing Group. https://books.google. co.uk/books?id=3H2Xg5qxz-8C.
- JVC (2020). Videosphere Service And Repair Manuals (Model 3 240). https://thydzik.com/videosphere/.
- Kerren, A., Purchase, H. C., and Ward, M. O. (2014). Introduction To Multivariate Network Visualization, pages 1–9. Springer International Publishing, Cham. http://dx.doi.org/10.1007/ 978-3-319-06793-3_1.
- Kirkpatrick, S., Gelatt, C. D., and Vecchi, M. P. (1983). Optimization by simulated annealing. *Science*, 220(4598):671–680. http://science.sciencemag.org/content/220/4598/671.
- Klicpera, J., Bojchevski, A., and Günnemann, S. (2018). Predict Then Propagate: Graph Neural Networks Meet Personalized Pagerank. Arxiv. http://arxiv.org/abs/1810.05997.
- Kohlbacher, O., Schreiber, F., and Ward, M. O. (2014). Multivariate Networks In The Life Sciences, pages 61–73. Springer International Publishing, Cham. http://dx.doi.org/10.1007/ 978-3-319-06793-3 4.
- Kumar, G. and Garland, M. (2006). Visual exploration of complex time-varying graphs. IEEE Transactions on Visualization and Computer Graphics, 12(5):805–812.

- Lu, S. (2019). D3-Annotate. https://d3-annotation.susielu.com/.
- Lyons, K. A. (1992). Cluster busting in anchored graph drawing. In Proceedings of the 1992 Conference of the Centre for Advanced Studies on Collaborative Research - Volume 1, CASCON '92, pages 7–17. IBM Press. http://dl.acm.org/citation.cfm?id=962198.962200.
- Ma, K. and Muelder, C. W. (2013). Large-scale graph visualization and analytics. Computer, 46(7):39– 46.
- Maaten, L. v. d. and Hinton, G. (2008a). Visualizing Data Using T-Sne. Journal of machine learning research: JMLR, 9(Nov):2579–2605. http://www.jmlr.org/papers/v9/vandermaaten08a.html.
- Maaten, L. v. d. and Hinton, G. (2008b). Visualizing Data Using T-Sne. Journal of machine learning research: JMLR, 9(Nov):2579–2605. http://www.jmlr.org/papers/volume9/vandermaaten08a/ vandermaaten08a.pdf.
- Martin, S., Brown, W., Klavans, R., and Boyack, K. (2011). Openord: An open-source toolbox for large graph layout. *Proc SPIE*, 7868:786806.
- Martin Grandjean (2016). Connected World: Untangling The Air Traffic Network. http://www. martingrandjean.ch/connected-world-air-traffic-network/.
- Metropolis, N., Rosenbluth, A. W., Rosenbluth, M. N., Teller, A. H., and Teller, E. (1953). Equation of state calculations by fast computing machines. *The Journal of Chemical Physics*, 21(6):1087–1092. http://scitation.aip.org/content/aip/journal/jcp/21/6/10.1063/1.1699114.
- Michal, G. (1965). Metabolic Pathways. https://www.roche.com/sustainability/philanthropy/science_education/pathways/pathways-ordering.htm.
- Montañez, A. (2016). How Science Visualization Can Help Save The World. https://blogs. scientificamerican.com/sa-visual/how-science-visualization-can-help-save-the-world/.
- Mortenson, M. (1999). Mathematics For Computer Graphics Applications. Industrial Press. https://books.google.co.uk/books?id=YmQy799flPkC.
- Muelder, C., Gou, L., Ma, K.-L., and Zhou, M. X. (2014). Multivariate Social Network Visual Analytics, pages 37–59. Springer International Publishing, Cham. http://dx.doi.org/10.1007/ 978-3-319-06793-3 3.
- Needham, M. and Hodler, A. E. (2019). Practical Examples In Apache Spark & Neo4J. O'Reilly. https://neo4j.com/neoassets/graphbooks/Graph_Algorithms_Neo4j.pdf.
- Noack, A. (2004). An Energy Model For Visual Graph Clustering, pages 425–436. Springer Berlin Heidelberg, Berlin, Heidelberg. http://dx.doi.org/10.1007/978-3-540-24595-7_40.

- Norman, D. (2005). Emotional Design: Why We Love (Or Hate) Everyday Things. Basic Books. https://books.google.nl/books?id=h_wAbnGlOC4C.
- Over, E. A. B., Hooge, I. T. C., and Erkelens, C. J. (2006). A quantitative measure for the uniformity of fixation density: The voronoi method. *Behavior Research Methods*, 38(2):251–261. https://doi. org/10.3758/BF03192777.
- Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., and Duchesnay, E. (2011). Scikit-learn: Machine learning in Python. Journal of Machine Learning Research, 12:2825–2830.
- Pohl, M., Schmitt, M., and Diehl, S. (2009). Comparing The Readability Of Graph Layouts Using Eyetracking And Task-Oriented Analysis. In *Computational Aesthetics in Graphics, Visualization,* and Imaging. The Eurographics Association.
- Purchase, H. (1997). Which Aesthetic Has The Greatest Effect On Human Understanding?, pages 248– 261. Springer Berlin Heidelberg, Berlin, Heidelberg. http://dx.doi.org/10.1007/3-540-63938-1_67.
- Purchase, H. C. (2002). Metrics for graph drawing aesthetics. Journal of Visual Languages and Computing, 13(5):501 – 516. http://www.sciencedirect.com/science/article/pii/S1045926X02902326.
- Purchase, H. C., Colpoys, L., Carrington, D., and McGill, M. (2003). Uml Class Diagrams: An Empirical Study Of Comprehension, pages 149–178. Springer US, Boston, MA. http://dx.doi.org/ 10.1007/978-1-4615-0457-3 6.
- Rickard, A. (2020). MCM Website. http://mcm.york.ac.uk/.
- Roberts, J. C., Yang, J., Kohlbacher, O., Ward, M. O., and Zhou, M. X. (2014). Novel Visual Metaphors For Multivariate Networks, pages 127–150. Springer International Publishing, Cham. http://dx.doi.org/10.1007/978-3-319-06793-3 7.
- Sangers, A., van Heesch, M., Attema, T., Veugen, T., Wiggerman, M., Veldsink, J., Bloemen, O., and Worm, D. (2019). Secure Multiparty Pagerank Algorithm For Collaborative Fraud Detection. In *Financial Cryptography and Data Security*, pages 605–623. Springer International Publishing. http://dx.doi.org/10.1007/978-3-030-32101-7 35.
- Sangkloy, P., Burnell, N., Ham, C., and Hays, J. (2016). The sketchy database: Learning to retrieve badly drawn bunnies. ACM Trans. Graph., 35(4). https://doi.org/10.1145/2897824.2925954.
- Schreiber, F., Kerren, A., Börner, K., Hagen, H., and Zeckzer, D. (2014). Heterogeneous Networks On Multiple Levels, pages 175–206. Springer International Publishing, Cham. http://dx.doi.org/ 10.1007/978-3-319-06793-3_9.

- Shneiderman, B. (1996). The eyes have it: A task by data type taxonomy for information visualizations. In Proceedings of the 1996 IEEE Symposium on Visual Languages, VL '96, pages 336–, Washington, DC, USA. IEEE Computer Society. http://dl.acm.org/citation.cfm?id=832277.834354.
- Shneiderman, B. (1997). Designing The User Interface: Strategies For Effective Human-Computer Interaction. Addison-Wesley Longman Publishing Co., Inc., Boston, MA, USA, 3rd edition.
- Staples, J., Nickerson, D. A., and Below, J. E. (2013). Utilizing graph theory to select the largest set of unrelated individuals for genetic analysis. *Genetic Epidemiology*, 37(2):136–141. https:// onlinelibrary.wiley.com/doi/abs/10.1002/gepi.21684.
- Stefaner, M. (2020). Truth & Beauty Multiplicity. https://truth-and-beauty.net/projects/ multiplicity.
- Steven Franconeri (2018). Openvis Conference Proceedings. https://www.youtube.com/watch?v=Jq2Rc0WlYTE.
- Taylor, M. and Rodgers, P. (2005). Applying graphical design techniques to graph visualisation. In Ninth International Conference on Information Visualisation, 06-08 July 2005, London, England: Proceedings, pages 651–656. IEEE Computer Society. http://kar.kent.ac.uk/14297/.
- Thomas, P. (1952). Conformal Projections In Geodesy And Cartography. Special publication. Coast and Geodetic Survey. https://books.google.co.uk/books?id=7a60MQEACAAJ.
- VTL (2019). Visual Thinking Lab. http://visualthinking.psych.northwestern.edu/.
- Ware, C. (2013a). Chapter four color. In Ware, C., editor, Information Visualization (Third Edition), Interactive Technologies, pages 95 – 138. Morgan Kaufmann, Boston, third edition edition. http: //www.sciencedirect.com/science/article/pii/B9780123814647000041.
- Ware, C. (2013b). Chapter three lightness, brightness, contrast, and constancy. In Ware, C., editor, *Information Visualization (Third Edition)*, Interactive Technologies, pages 69 – 94. Morgan Kaufmann, Boston, third edition edition. http://www.sciencedirect.com/science/article/pii/ B978012381464700003X.
- Ware, C. (2013c). Chapter two the environment, optics, resolution, and the display. In Ware, C., editor, *Information Visualization (Third Edition)*, Interactive Technologies, pages 31 – 68. Morgan Kaufmann, Boston, third edition edition. http://www.sciencedirect.com/science/article/pii/ B9780123814647000028.
- Ware, C., Purchase, H., Colpoys, L., and McGill, M. (2002). Cognitive measurements of graph aesthetics. *Information Visualization*, 1(2):103–110. http://dx.doi.org/10.1057/palgrave.ivs.9500013.

Wybrow, M., Elmqvist, N., Fekete, J.-D., von Landesberger, T., van Wijk, J. J., and Zimmer, B. (2014). Interaction In The Visualization Of Multivariate Networks, pages 97–125. Springer International Publishing, Cham. http://dx.doi.org/10.1007/978-3-319-06793-3_6. Chapter 4

Chemical model diagnostics using graph theory and metrics.

"The complexities of cause and effect defy analysis."

- Douglas Adams, Dirk Gently's Holistic Detective Agency

4.1 Introduction

The node-link (ball-stick) style structure has long been used to represent real-world relationships between items (Subsection 3.1.2). Such a structure is complementary to our cognitive disposition towards pattern recognition, and it is for this reason that the node-link visualisation format has been used for anything ranging from transportation maps [Beck, 2017] to the differentiation of ancestorial lineages of the human race (Figure 4.1). However, the abundance and complexity of real-world data often present us with difficulties in manually representing it in a useful form. In Section 3.2, it was suggested this may be overcome with the use of computational analysis and automated visualisation tools. Such methods usually require a level of data manipulation to transform the data into a machine parseable form.



Figure 4.1: **The human family tree.** This is a visual depiction of the human lineage, starting with our common ancestorial roots. In Chapter 2 it was shown that trees / graphs¹ are useful in showing relationships between items. Source: [Wood, 2014]

In the field of mathematics a graph, $G(\nu, \epsilon, \omega)$, is defined as a function of items (vertices²), ν which are connected through a series of connections (or edges¹) representing any relationships between them, ϵ .

 $^{^1\}mathrm{A}$ tree is a special case of a graph

 $^{^2{\}rm The \ term \ node, \ item \ or \ vertex \ shall \ be used \ interchangeably \ for \ the \ remainder \ of \ this \ chapter. \ This \ also \ applies \ to \ links/relationships/edges \ and \ edge-weight/strength$

Since relationships in the real world are rarely equivalent, we then encode the importance of each link in the form of an edge weight, or strength, ω . Such formats allow both numerical and computational algorithms to understand and interpret the graph structure, providing us with information about the data or make use of automated layout programs for visualisation.

This chapter builds on the work shown in Chapter 3 - where the ability to represent complex data in the form of a graph was used to (visually) draw information regarding network structure and temporal changes. Here situations, where the visual representation of many, large or complex networks is impractical, will be explored. We start by introducing a series of mathematical approaches which are capable of quantifying the graph (and nodes within it) and apply them to the co-author network for papers using the Master Chemical Mechanism (Section 4.2). Following these global metrics are used to categorise the chemistry within different atmospheric chemistry mechanism subsets, and provide us with an insight to the chemistry structure (Section 4.4) and finally apply these to real-world simulations representing a range of environments (marine, rainforest and urban) in Section 4.6.

4.2 Graph Metrics

The increase in the ability to gather useful data has resulted in difficulty when trying to interpret it (Section 2.1). The production of large, multivariate networks of inexplicable complexity hinders our ability to draw out meaningful conclusions based on visualisation alone. This means that much like the generation of mechanism [Aumont et al., 2005], or creating semi-automated graph drawing layouts, we must rely on the field of mathematics coupled with computational aid (Chapter 3).

Numerical algorithms derived from the field of Graph Theory can be used to circumvent the need for individual graph analysis and provide us with information about the network. One such subset of numerical algorithms are regarded as "centrality metrics", and may be used to rank the role and importance (centrality) of a node [Ferguson, 2018]. In the following sub-section, the four most common centrality metrics are discussed and applied to the MCM citation network.

4.2.1 Centrality Metrics And Academic Publishing.

One common application for graph analysis and visualisation is the representation and prediction of citation counts within academic journals [Small, 1973; Page et al., 1999; Monastersky and Van Noorden, 2019; Molontay and Nagy, 2020]. Here network-visualisation techniques may be used to highlight the origins of a paper. For instance, Figure 4.2 shows the multi-disciplinary research which underpins six prominent discoveries in the last 150 years.

The next section looks at the four most common centrality metrics and explores their properties

with the use of an (approximate) citation graph showing the papers which cite the Master Chemical Mechanism (Subsection 4.2.2).

4.2.2 The Master Chemical Mechanism (MCM)

The MCM, [Rickard, 2020], is a near explicit representation of our foremost understanding of gasphase tropospheric chemistry. The mechanism describes the oxidation of 143 primary emitted VOCs and the respective rates at which this occurs. It has been tested on over 300 chamber experiments and used as a benchmarking mechanism to assist the development of reduced mechanism, providing a useful means for the evaluation of air quality models [Dick Derwent, Andrea Fraser, John Abbott and Mike Jenkin , 2010]. The current version (3.3.1) contains 5809 chemical species and 17224 reactions to describe them [Jenkin et al., 2015]. However there there are still a number of weaknesses that need to be considered. Firstly there very little Cl chemistry and no other halogens in the mechanism. Reactions with O_2 are implicit as are RO_2 - RO_2 reactions, which are shown through the reaction with an RO_2 pool.

4.2.3 Data Collection

To generate a dataset on papers related to the MCM. The academic search engine (Google Scholar [Google, 2019]) is queried for all articles containing the words { "Master", "Chemical", "Mechanism" and "MCM" }. For each match, the first 100 pages of results are selected. Each of these contains ten articles, from which the first 100 pages of related articles are chosen. In taking the top 1000 citations for each page, a network of 15744 papers and 30178 citations³ is created. This process made use of an edited version of the *etudier* Github repository, [Edsu and Ellis, 2019].

³Note: this had the potential of returning up to 1000,000 nodes


Figure 4.2: **150 years of letters to Nature.** A visualisation showing how previous research is used to inspire future studies. Important discoveries (DNA, Cloning(frogs), Bio-Currents, Ozone Hole, Molecular Sieves and Exoplanets) are split into research which contributed to their formation (below), and the consequent papers produced from each discovery. Use of colour is used to emphasise the multi-disciplinary nature of prolific scientific discovery. Source: [Barabási, 2019]

4.2.4 Visualising The Data.

The initial visualisation of the dataset is accomplished through the use of THREE.js [Cabello, 2019]. This makes use of WebGL bindings and allows for the efficient viewing, querying and interacting of the data in 3 dimensions. This helped identify the temporal changes within the network by mapping a papers publication year to the z direction, Figure 4.3, as discussed in Subsection 4.2.5.

4.2.5 Filtering The Data

In the method used to web scrape data, there are several features which need to be corrected/removed. The reasons for this are discussed below.

A note on unintentional filtering

The script used for web scraping extracts author names directly from the google scholar page, and not the articles themselves. This means some author names can be omitted and replaced by ellipses - producing an inaccurate graph. Therefore the results in this section are not explicit, but rather a demonstration of graph theory on a real-world dataset.

Pre-1996

There exist several papers predating the conception of the MCM (1996). A number of these are incorrect and contain publication dates <1900 which may be the result of missing information or a fault in googles web scraping algorithm. Any such papers are removed from the dataset.

Articles published before 1996 are deemed necessary in the creation of the MCM, but not its influence on the field of atmospheric science (see cone shape in Figure 4.3b) - it, therefore, makes sense to filter these from the dataset.

N-th degree research

Not all research articles in a field reference other articles with the same field. Figure 4.2 showed us that many of the great discoveries in science have a multi-disciplinary nature. It is for this reason that it is expected that articles from non-atmospheric areas of research may reference or build upon specific areas of research touched by the MCM. Such papers, and in consequence the papers which cite them, have little or no links to many of the core MCM papers. Such papers manifest themselves as a halo of satellite clusters which are connected by themselves but not with the main body of the graph, Figure 4.3a. In using a 3D perspective viewpoint (Figure 4.3c) it is possible to identify the paper which references the MCM and then the consequent papers which cite it by observing the satellite



(a) A 2D force directed representation of the network using the gephi software [Bastian et al., 2009]

(b) 3D orthographic camera (sideview). This shows a sideways view of the graph in (a) where time is across the x axis



(c) 3D perspective camera

Figure 4.3: Initial 3D graph representation of the scraped MCM citation graph. (a) shows the 'classic' graph representation of the network. (b) shows a size representation using an orthographic perspective. Here time is shown across the x axis, with yellow being the most recent. (c) uses a perspective camera, which emphasises the time component of the data. Still captures of 2D and 3D visualisations of the dataset.

Node size corresponds to the number of citations, and colour (and z-axis) corresponds to the publication year for each paper.

clusters, and the gradually lightening spiral of papers which emanate out of it. These groups of papers are now discussed.

Analysis of the network connections for each cluster can allow us to identify the indirect relationships between some of these diverse topics (Table 4.1) contained within the satellite nodes. Here it can be seen that the use of photochemical ozone creation potentials [Derwent et al., 1998; Jenkin and Hayman, 1999] are used for the Life cycle assessment of Italian high-quality milk production [Fantin et al., 2012]. Similarly, indirect paths such as the paper: "Temporal controls on dissolved organic matter and lignin biogeochemistry in a pristine tropical river" ([Spencer et al., 2010]) can be used to link to [Stubbins et al., 2008] and ultimately the MCM protocol paper [Saunders et al., 2003].

If we desired to remove such papers, the simplest method would be to recreate the graph into one where links are drawn between papers that are cited together (Subsection 4.2.6) and then removing any nodes without any external connections (isolates).

Fabrication of Bioinspired Actuated Nanostructures with Arbitrary Geometry and Stiffness	[Pokroy et al., 2009]
Temporal controls on dissolved organic matter and lignin bio- geochemistry in a pristine tropical river	[Spencer et al., 2010]
Neuroproteomics in Neurotrauma	[Ottens et al., 2006]
Fast start-up of a pilot-scale deammonification sequencing batch reactor from an activated sludge inoculum	[Jeanningros et al., 2010]
Red blood cell oxidative stress impairs oxygen delivery and induces red blood cell aging	[Mohanty et al., 2014]
Life cycle assessment of Italian high quality milk production.	[Fantin et al., 2012]
	1

Table 4.1: A selection of research papers not directly connected to the field of atmospheric modelling.

Unprobable occurances

Finally, the extracted network also contains many disconnected component subgraphs - graphs with no connection to atmospheric science. An example of this is seen in an article about neuroproteomics in neurotrauma [Ottens et al., 2006]. In analysing the paths which connect this, it is seen to cite the paper on "Large scale gene expression profiling of metabolic shift of mammalian cells in culture", [Korke et al., 2004]. This is an anomaly which within its structure contains the words "Master", "Chemical" and "Mechanism" (separately) and has 'MCM' as an abbreviation for one of the author names. Disconnected sub-components are omitted from the analysis to remove such papers.

4.2.6 The Co-Citation Network

The document coupling techniques of co-citation was introduced in the 1970s as an alternative approach for quantifying the results within the science citation index [Small, 1973]. Rather than representing a graph using backpropagation (through the use of referencing and citation counts), a co-citation network introduces a link between papers if, and only if, they have been cited together. Although this loses the directionality of a graph, it allows us to show forward propagating trends between papers within the same field.

Applying the above method allows us to reduce the citation graph of 451 papers and 5402 edges to an undirected co-citation graph of 2758 edges - halving the number of original links between papers.

4.2.7 The Co-Authorship Network

An alternative to exploring which papers which are cited together are to look at their authors. Here undirected links are drawn between authors on the same paper. This style of analysis was used to show that the number of papers per author, and the total number of authors per paper can vary between research fields, [Newman, 2004]. In combining this with a series of network centrality metrics, [Fujita et al., 2017] revealed that it is possible to discern promising researchers from both iter and Intra disciplinary groups.

In building a co-authorship network for the MCM, we can identify authors who publish together⁴ and highlight research groups who work with the MCM, Figure 4.4. This shows how authors with a similar geographic location/institution are more likely to publish together. The largest cluster here falls under the MCM developer team, which resides between the University of Leeds and York. Next two German institutions which are heavily involved in the atmospheric chemistry field (FZ-Julich and Max Planck for Chemistry, Mainz), followed by an assortment of Chinese authors, mainly centred around the Beijing or Hong Kong region.

⁴Disclaimer: as mentioned earlier, not all authors for every paper were recorded by the web scraping algorithm



Figure 4.4: **The co-author network.** In representing the authorship network as a force-directed graph, we see cliques or clusters of people who publish together. It is noted that this often occurs when they have a similar geographical location. Node sizes and colour represent author rankings using the PageRank algorithm (Subsection 4.3.5)

Labeled Graph of Figure 4.4

See Section C.1 for a heavily labelled version of the graph above graph showing the co-author network on work relating the Master Chemical Mechanism.

4.3 Metric Analysis

The co-author network (Figure 4.4) can be used to demonstrate the functions of each centrality metric. This subsection will access the efficiency of graph centrality metrics in their ability to identify influential nodes within a network.

4.3.1 Degree Centrality

The simplest, and most intuitive, metric is degree centrality [Freeman, 1978]. In counting the number of edges incident on a node (in and out), we calculate the degree of a node. In this instance, this corresponds to the number of papers co-authored by an individual. This gives us an idea of the importance of a node and has been used to calculate influence within social media or the probability of a profile committing online auction fraud [Gemma, 2019; Freeman, 1978].

Example analogy: If we take the UK rail network as an example, As individual stations, Warrington, Birmingham, Manchester and Doncaster all have a high degree (a large number of different rail networks passing through them. Similarly for the London network, Victoria and Kings Cross will have the highest degree value.). Examples are shown in Section C.2

The author network in Figure 4.5 shows many of the names with a high degree are contributors (or colleagues to the contributors) of the MCM at Leeds. The authors with the most collaborations, or links, are very likely to appear within the most cited or citing papers (Table 4.2 and Table 4.3 discussed below). This is likely because both development (well-cited) and the evaluation/usage (well citing) of a mechanism requires knowledge from a range of different fields, making it an interactively collaborative process.



	-
M Pilling	25
H Guo	24
L Whalley	23
L Xue	22
D Heard	19
X Wang	19
Z Ling	18
A Lewis	17

Figure 4.5: **Degree Centrality.** In applying the degree centrality to the co-authorship network, it is possible to pick the authors with the greatest number of papers, of which the top 10 have been listed.

Directed Degree

For graphs where link direction holds an inherent meaning regarding their representation (for example in the citation graph an outward link symbolises that paper citing the one that the link points to), it is possible to further divide the degree centrality metric into inwards and outward links. This can allow us to separate authors who are highly cited (in-degree) from those who use lots of papers (out-degree). In applying these metrics to the directed citation graph, it is possible to get an insight into the core MCM development papers (Table 4.2) and separate them from those who make use of the mechanism as part of a greater study (Table 4.3).

Protocol for the development of the Master Chemical Mech- anism, MCM v3 Part A tropospheric degradation of nonaro- matic volatile organic compounds	Saunders et al. [2003]
Protocol for the development of the Master Chemical Mecha- nism, MCM v3 Part B tropospheric degradation of aromatic volatile organic compounds	Jenkin et al. [2003]
Development of a detailed chemical mechanism MCMv3. 1 for the atmospheric oxidation of aromatic hydrocarbons	Bloss et al. [2005]

Table 4.2: In-Degree of the citation network: The top 3 most cited papers.

The MCM v3.3.1 degradation scheme for isoprene	Jenkin et al. $[2015]$
Atmospheric photochemical reactivity and ozone production at two sites in Hong Kong Application of a master chemical mechanismphotochemical box model	Ling et al. [2014]
HOx budgets during HOxComp A case study of HOx chem- istry under NOxlimited conditions	Elshorbany et al. [2012]

Table 4.3: Out-Degree of the citation network: The top 3 most citing papers.

4.3.2 Closeness Centrality

Often within a network, we are interested in how easy it is to to get information from one node to every other node. This is what the closeness centrality tells us. To calculate a nodes closeness, we begin by taking the reciprocal sum of all the Dijkstra paths⁵ to every other node [Poliaktiv, 2011; Sabidussi, 1966]. This gives a representation of how far information from a particular person (node) will need to travel to reach every other node. Such a metric has applications in intelligence gathering, telecommunications and word importance within key-phrase extraction [Krebs, 2002; Borgatti, 2005; Boudin, 2013].

Example analogy: If we take the UK rail network as an example, York station will have a high closeness value as it is well connected and central in location. This means it is easy to reach every other location when compared to other stations, Section C.2

⁵The shortest available path.

For the co-authorship network, Figure 4.6, nodes have been coloured by their closeness value. Here a heat-map-like effect may be observed, showing that information between the dense Leeds-York cluster makes it easier to disseminate information across all parts of the graph. The results of the closeness centrality suggest that if a problem (bug), or improvement (update) occurs, Michael Pilling would be the best served to pass that information to all other groups using the MCM.



Figure 4.6: Closeness centrality within the co-Author network. Here a colour/size gradient is seen, with the nodes that are more central (in location) and better connected having a higher closeness than those in the peripheries - which are harder to get to.

4.3.3 Betweenness

In social networks, it is often important not only to know who has the greatest reach (closeness centrality) but also where bottlenecks or 'broker' positions occur. Nodes with a high betweenness control, or limit, the amount of information that can be transferred across the network. If a node lies on a geodesic (the shortest path between two other nodes), we may consider it a 'pivotal' node, due to its role within the network [Needham and Hodler, 2019]. Should such a node then be removed, the overall flow of information incurs either a deviation, the information will either need to travel a longer (alternative) route or may not be able to reach its destination at all [Freeman et al., 1991; Freeman, 1977; Brandes, 2001; Borgatti, 2005]. Betweenness centrality is a count of the number of geodesics which pass through a node. If multiple 'shortest' paths are possible, these can be accounted for using division in the algorithm mathematics.

Example analogy: Expanding on the UK rail network analogy, Shrewsbury station serves the critical role of connecting many lines from England to Wales. In removing this station, routes from the Liverpool or Manchester to Cardiff will be greatly increased. Additionally, the Aberystwyth section of the line will then become isolated from the rest of the country.

Authors with a high betweenness in Figure 4.7 are seen to lie along the joints between clusters. Here we can imagine that removing Li, Griffin or Liu can disrupt the overall flow of collaboration, potentially isolating the work of the Max Planck for Chemistry from that of everyone else. Similarly, Jenkin and Pilling can be seen as holding much of the Leeds cluster together. In removing them from the network (if for example, the refused to collaborate) it is possible to see how many of groups within the Leeds environment may not have worked together, with the cluster potentially separating into several smaller groups. Finally, we see that Saunders (Australia) is highlightes as an important node - an action which can be attibuted her introducing the Chinese atmospheric community to the MCM. In removing her from the network, it can be seen that much of the collaboration which exists would have been significantly less likely.



Figure 4.7: Betweenness centrality within the co-Author network. Nodes which lie on a pivotal position (connecting/bottleneck) tend to have a high betweenness value due to their crutial role within the network. The colour represents the betweenness centrality

4.3.4 Spectral Methods And Matrix Analysis

Graphs can often be represented in the form of relationship (adjacency) matrixes (ref Chapter 1). This allows us to apply the theory of linear maps, such as eigenvectors and values, to stochiometric data in matrix form. Such methods have been around since the 1950s, [R. Seeley, 1949], but mainly became popular with the release of Larry Page's page-rank algorithm [Page et al., 1999] - the algorithm that began google. These methods, in addition to the HITS algorithm (Table 4.3.4), make use of a graphs

	No Normalisation	Row Normalisation
No Damping	Eigenvector [Bonacich, 1987, 2007]	Markov Chain Steady State
		[R. Seeley, 1949]
Damping	Katz [Goh et al., 2001]	Total Effect Centrality PageRank
		[Page et al., 1999]

native matrix representation to calculate node importance. Spectral algorithms can be broken down into four categories [Vigna, 2016]:

Here damping terms represent the probability of moving to the new random starting position, allowing for the user to 'randomly select a new webpage' or leave an isolated cluster. The normalisation of the matrix does not affect the node ranking, but merely adjusts the numerical output of the algorithm. It is for this reason that its overall practicality may be debated [Vigna, 2016]. Since page rank is the most common of these methods and allows for a tune-able degree of randomness within network propagation, this is discussed in more detail in the next subsection.

Hypertext Induced Topic Search (HITS)

A common eigenvector algorithm used for classifying webpages is the HITS algorithm. This helps categorise the role of a node as either a Hub or an Authority, [Kleinberg, 1999; Langville and Meyer, 2005; Kumar and Upfal, 2000]. Similar to the in and out-degree metrics, this algorithm separates nodes with many outgoing links (an authority) from those with many ingoing ones (an information hub). Overall this provides similar results to the in/out-degree, although since it looks more on how information propagates across the network as a whole, it often provides more accurate, and different, rankings to simple degree analysis.

4.3.5 Page Rank

Arguably the best-known centrality algorithm is PageRank. This is a spectral method for measuring the transitive influence of a node, by taking the effect of neighbours and by their neighbours into account [Needham and Hodler, 2019]. The page rank algorithm was initially developed to provide a better way of ranking web pages [Page et al., 1999]. Here an important page is not only one of many links, but links to other important sources. In the context of academic papers, that same paper also found that in predicting future citations, the page rank algorithm fared better than using the current citation count of a paper. To explain how this works, we will look at the mathematics behind the algorithm, and then eventually apply it to the co-authorship graph in Subsubsection 4.3.5.3

4.3.5.1 The Google Matrix

To solve for page rank, a 'google matrix' must first be constructed. Once done this is iterated until convergence is reached.

To build a google matrix, we must first generate a dyadic link map of the graph⁶ - its adjacency matrix $A_{i,j}$ (*i*, *j* are the source target indexes). This is then converted into a Markov matrix $M_{i,j}$ by dividing each column *j* by the sum of the total outgoing links of node *j*, Algorithm 1. Items with no outgoing links (sinks), are adjusted with either a personalised⁷ list of values or the constant 1/n, (where *n* is the number of nodes) to replace the zero-sum columns. This produces a normalised⁸ matrix of Markov chains representing the fractional production for node *j* from all other nodes.

Algorithm 1 Adjacency to Markov matrix.		
1: Obtain graph adjacency matrix, $A_{i,j}$.		
2: repeat		
3: for each $j \in \text{ columns } \mathbf{do}$		
4: $M(:,j) \leftarrow A(:,j) / \Sigma_{i=1,n} A(j,i)$		
5: end for		
6: until $\sum_{i=1,n} M(i,j) = 1$		

The google matrix $G_{i,j}$ can now be defined using Equation 4.1. Cyclic reactions and nodes that only point towards each other within a group can 'trap' the user, increasing their ranks. To account for this, a damping factor, typically $\beta = 0.85$ is used. This defines the probability that the user follows a link, and that for which they randomly select another page: $(1 - \beta)^{-9}$. The damping factor used varies significantly with the application, with values such as $\beta = 0.694$ having been found optimal for the use of biological data [Hobson et al., 2018].

$$G_{i,j} = \beta M + \frac{1-\beta}{n} \tag{4.1}$$

 $\begin{array}{rrrr} \beta & - & Probability \ the \ user \ follows \ a \ link \\ (1-\beta) & - & Probability \ the \ user \ does \ not \ follow \ a \ link \ (teleportation) \\ n & - & Number \ of \ items \ / \ species \end{array}$

M - Normalised markov matrix

 $^{^{6}\}mathrm{In}$ sociology a dyad is a group of two people - the smallest possible social group.

⁷The use of user chosen (beginning) values for each node are used.

⁸ $\Sigma_{i=1,n} M(i,j) = \text{unity}$

⁹Also known as teleportation.

4.3.5.2 Solving The Algebra

Once defined, the google matrix is solved by propagating a one's vector, r of length n, where n is the number of papers (or items) using Algorithm 2.

Algorithm 2 Solving the google matrix linear algebra

1: Define value vectors \bar{r}_t and \bar{r}_{t+1} : 2: $\bar{r}_t = [1_1, 1_2, ..., 1_n], \ \bar{r}_{t+1} = [0_1, 0_2, ..., 0_n]$ 3: 4: while $||\bar{r}_{t+1} - \bar{r}_t|| > \epsilon$ do 5: $\bar{r}_{t+1} \leftarrow M.\bar{r}_t$ 6: $\bar{r}_t = \bar{r}_{t+1}$ 7: end while

This is repeated until a pre-defined tolerance, ϵ is reached. For best results, this can be set to just under the numerical precision of the programming language/hardware.

For smaller systems, it is possible to use the LAPACK [LAPACK, 2019] library, as used by Oliphant [2006]. For a vast network, however, the computation of a $n \times n$ matrix can be very memory inefficient for small machines. It is then possible to apply the methods as described above using a sparse matrix on per-node bases as can be seen within the Python SciPy implementation of the Networkx source code [Jones et al., 01; Hagberg et al., 2008].

4.3.5.3 Prediction

As the PageRank algorithm loos at how quantities 'flow' within a network, it can be used to identify not only the bottlenecks (betweenness centrality) but also any nodes which are connected well within the network. As the flows between a node are somewhat governed by the number of links it contains, the PageRank algorithms tend to correlate, but not dependence, on the betweenness of a node. Figure 4.8 uses the PageRank algorithm to identify important authors within each 'cluster' or research group. Due to its propagating nature, authors connected to these important nodes are often also of greater importance. An application of this can again be the determination of how to best spread new results or information with the least number of people. *Note: if we only had one person we would probably use the node with the highest closeness centrality.*



M Jenkin	0.010435
L Whalley	0.006589
M Pilling	0.006488
S Saunders	0.005591
D Heard	0.005192
N Carslaw	0.004833
H Guo	0.004594
G Wolfe	0.004523
A Lewis	0.004508
R Griffin	0.004500

Figure 4.8: **Page Rank centrality within the co-Author network**. Node size and colour represent the ranking of each node from the page rank algorithm. Larger, lighter coloured nodes are more important.

4.3.6 Conclusions

In this section, we have explored the use of centrality metrics to provide us with information on an unweighted co-authorship network of the MCM. Having used these to demonstrate the different roles that may be extracted from a node, we can move on to applying them to a chemical mechanism. In the next section, a global (applying to the network as a whole) set of metrics will be used to determine the network type/structure of the MCM. Once this has been done, graph construction using simulation

results (a weighted graph) will be looked into in Subsection 4.5.1.

4.4 Classifying The Master Chemical Mechanism Network

Having shown that graph metrics can help the roles of individual nodes within the network, these are now applied to an atmospheric chemical system. Since computational efficiency and resources are often a limiting factor, many applications of the MCM only require a small subset of the entire mechanism. For this reason, it may be of interest to compare these against each other, in an attempt to classify the type of network the MCM chemistry falls under. In this section, we apply graph theory to the entire MCM network to determine its defining characteristics. This is achieved through the analysis of several hundred Monte Carlo selected subsets of the MCM. Each of these is a different combination of the primary emitted VOCs within the MCM v3.3.1.

4.4.1 Network Density

Network density is the easiest metric to understand. Visually this can induce complexity and obscure aspects in a graph; mathematically, it can greatly increase the computation time for metrics or algorithms. By definition, we can define network density as a measure of how well connected a node is to every other node. Mathematically it is the ratio of edges against the total number of possible edges for a complete graph¹⁰ of the same size. In chemical terms, we can use this to determine the sparsity of the graph (which has applications on model integrator selection) and give us insights on the chemical structure. In Figure 4.9, higher numbers of species (nodes) results in an overall decrease in the node-edge ratio - its density. This suggests a modular or hierarchical structure, where new species directly react only with a set number of species, and not the entire mechanism. An explanation for this is that the addition of larger species introduce new branches within the chemistry, which then need to be oxidised before they are small enough to react with the species from a different branch. Since these branches are somewhat isolated from the rest of the chemistry, they decrease the network density, even though their addition may increase the amount of chemistry that occurs within it.

¹⁰A complete graph is one where every node is connected to every other node.



Figure 4.9: How the MCM graph density scales with number of species. A figure showing that an increasing number of species within a mechanism subset results in an increased model sparsity (decreasing density).

4.4.2 Small World Phenomena

Within the biological or social sciences the small world phenomenon, colloquially known as 'six degrees of separation', is a common occurrence within network structure [Watts and Strogatz, 1998]. Such networks have a large number of localised clusters (cliques) all with a short path length between their elements [Humphries and Gurney, 2008]. This makes it easy to reach all parts of a network with only a couple of hops/reactions. In the initial interactive explorations of graph visualisation, it was found that in selecting the reactions of a node, and consequently the reactions of all the nodes which react with them, very quickly a large proportion of the network was highlighted. This suggests that the network may follow the small world phenomena, especially as it is a sparse network, Subsection 4.4.1. One of the possible methods for establishing the small world-ness of a graph by calculating the omega (ω) coefficient [Hagberg et al., 2008]:

$$\omega = L_r / L - C / C_l \tag{4.2}$$

Here C is the average clustering coefficient and L, the shortest path length of the graph. Comparing these with the average shortest path length, L_R , and clustering coefficient C_l (as calculated using an equivalent random and lattice graph) gives the above equation. The output is a result between positive and negative one $\{-1,1\}$, where a value of 0 suggests the graph exhibits perfect small world-ness.

In assessing the network structure of the MCM, a Monte Carlo (random) approach was taken to

extract several hundred subsets from the entire mechanism. For each of these, the omega coefficient was calculated and plotted in Figure 4.10. Here it is seen that subsets with a small number of species (for example those derived only from methane or ethane) exhibit a more lattice-style (grid) graph, with the majority of the networks showing a more random network structure Figure 4.11. All the results, however, show a prevalence of small-world features over any of the alternative network structures - they are closer to 0 than 1 or -1. This reflects the idea that large species react locally, forming branches (Chapter 3), before oxidising to smaller species with more reactions. This result is also seen within the Reaxys chemical database [Jacob and Lapkin, 2018].



Figure 4.10: A figure showing the small worldness for many Monte-Carlo selected MCM subsets. The network structure of these is then assessed using the omega coefficient, with [-1,0,1] corresponding to the perfect lattice, small-world and random network structure. Here Node size and colour represents the number of reactions in the mechanism subset and the number of primary VOCs (blue=small, green=large).

4.4.3 Power Law And Scale-Free Graphs

In real-world applications, it is common to have a hierarchical structure. These are often seen in the increase of citation counts in academic papers [de Solla Price, 1965], email threads [Ebel et al., 2002] and the world wide web [Needham and Hodler, 2019]. Unlike random or small-world graphs, scale-free graphs take a hub-and-spoke structure (Figure 4.11), which follows a power-law distribution - that is that scaling probability $p(x) \propto x^{-\alpha}$, where α is a constant and known as the scaling parameter.

Broido and Clauset [2019] suggests that scale-free networks are rare, and often misdiagnosed with incorrect tests, or the misinterpretation of power-law features in a network. Similarly, Clauset et al. [2009] suggests that even if the data distribution of a graph is well represented by the power-law distribution, in many cases a logarithmic or exponential distribution may have a better fit.



Figure 4.11: The different network structures. A visual depiction of the different graph structures. Source: Needham and Hodler [2019]

To assess the best distribution for describing the Monte Carlo subsets of the MCM, the Kolomogorov-Smirnov statistic [Press et al., 1992] was used to analyse the goodness of fit of the ω coefficient in Figure 4.10 to a number of distributions. This calculates the maximum distance D between the selected cumelative distribution function S(x) (In our case the Logarithmic, Exponential and Power Law) of the data and the fitted model P(x):

$$D = \max_{x \ge x_{min}} |S(x) - P(x)|$$
(4.3)

Using the MCM subsets from Figure 4.10, Figure 4.12 shows that out of the three tested distributions, the MCM is best represented as a power-law distribution (smaller KS distances are better). Although this is not entirely within the chosen 5% significance, it is highly indicative that some aspects of the network are indeed scale-free.



Figure 4.12: Comparing the MCM subsets against a power law, logarithmic and exponential distribution. The fit for different cumulative probability distributions of nodes in the MCM network is compared to determine the type of network hierarchy the chemistry follows. This is done by comparing the distance of the calculated distribution of data against a perfect one using the Kolmogorov-Smirnov test. The closer the two distributions, the lower the KS distance, and the better the fit.

4.4.4 Describing The MCM Network

To conclude, the MCM network exhibits both small world and scale-free (power-law) characteristics. This agrees with previous knowledge about the apparent network structure (Chapter 3). Here large primary emitted hydrocarbons produce branches of a hierarchical nature, as they are progressively broken down into smaller species. Since smaller species are then able to react with a much higher range of species, they then begin to form a tightly connected core, which exhibits many small-world features.

Having classified the MCM network type, the next section will look at how MCM based simulation results can be converted into the graph structure for a more in-depth analysis, Section 4.6.

4.5 Graph Construction Methodology

Thus far, we have only applied a qualitative analysis of the relationships between species in a mechanism. Although this can educate us about the chemistry within a specific system, often a quantitative value for the rate of reaction between different species is required when undergoing scientific evaluation or policy advice. A chemical mechanism is placed within an atmospheric model, initial concentrations are supplied, and the chemistry is propagated forwards¹¹ in time. Currently, there exist three primary

¹¹Or backwards if the adjoint is used.

model diagnostics which we may use to analyse the importance or role of a species from a simulation (model) output, concentration time-series, rates of loss and the Jacobian.

4.5.0.1 Concentration Time Series

The simplest of these methods look at the abundance of a species at a specific point in the atmosphere (its concentration ad a specific time). As time moves forwards, chemicals within the atmosphere undergo a range of reactions which result in the making and breaking of bonds - thus the changing from one species to another.

Using the species concentration as a metric, we can map how it changes over time, and how in changing the initial concentrations of a simulation can produce different results. This can be useful for looking at a range of possible scenarios and evaluating the potential outcome after a pre-determined amount of time. An example would be through the use of policy-based simulations to predict changes in air composition over cities.

Using a simple example from a methane only subset of the MCM (Figure 4.13), it is possible to observe the inverse relationship between NO_2 and NO using only their concentration profiles. Here nitrogen monoxide reacts with a RO_2 species to produce an RO and nitrogen dioxide. This then photolyses back to nitrogen oxide, releasing oxygen which may go on to form ozone (Subsection 1.3.1). The latter part of this reaction is dependent on photons and therefore can only occur during daytime (mostly).



Figure 4.13: A concentration (mixing ratios) time series from a simple methane-only simulation. This is the simplest method for identifying changes in species within a model simulation. This multi-plot shows the changes in concentration profiles for all initialised species (NOx:10ppb; CH₄:20ppb; O₃:30ppb) following an initial 3 day spin-up to steady state.

4.5.0.2 Rate Of Production And Loss

Analysing the concentration-time profiles allows the comparison of how a series of scenarios or runs change concerning their initial conditions and simulation length. Although these can tell us how, and how much, each species changes over time, it does not rank or quantifies the specific reactions to which this may be attributed. Rate of Production Analysis (ROPA)¹² provides a method for establishing the total contribution from each reaction by calculating the change of concentration (concerning time) for the produced species - the instantaneous reaction flux.

$$r_1 = A + B \xrightarrow{\kappa_1} \eta C$$
 Reaction 1 (4.4)

$$f(C) = \frac{\delta C}{\delta t} = \eta \kappa_1[A][B] \qquad \text{Instantanious Flux } (\Gamma)$$
(4.5)

Here A, B and C are example species; [A],[B] and [C] are species concentrations; η and ω are rate coefficients, and κ is the rate of the reaction.

Using a sample simulation representative of the conditions within Beijing (an urban environment), we explore the reactions contributing to the production and loss of CH_3CO_3 (Figure 4.14) [at noon]. The main reason for this specific example is that it can demonstrate how isolating a specific cause for the change within a species concentration may prove difficult in the context of atmospheric chemistry. Here we have many similarly weighted production and loss reaction, including that of peroxyacetyl nitrate (PAN) and nitrogen dioxide: $CH_3CO_3 + NO_2 \iff CH_3C(O)ONO_2$ (PAN). The reversible nature, coupled with its near-identical production and loss fluxes produce a tiny net change within our species of interest (CH_3CO_3). Although this may be seen by calculating the cumulative flux between individual species, it is evident that simply looking at the concentrations or highest-ranking reaction fluxes may not be the best method of determining influence. To account for this, we can look at how a change in one species can affect another using the Jacobian method.

 $^{^{12}}$ and loss



Figure 4.14: Rate of production and loss analysis plot for CH_3CO_3 exhibiting a net loss (daytime). An example ROPA plot from a simulation representing the chemistry within Beijing. This is used to identify the usefulness and weaknesses of using such a method. DUMMY represents the deposition term for any species.

4.5.0.3 The Jacobian

"The Jacobian [matrix] generalises the notion of gradient to describe the sensitivity to a vector" -Brasseur and Jacob [2017]. That this means is that in taking the partial derivatives of each reaction flux (e.g. from Equation 4.5), we can construct a representation of the influence each species has on itself - for example, the influence of species A on C and B on C (Equation 4.6-4.7).

$$\frac{\partial}{\partial A} \cdot \frac{\partial C_{r_1}}{\partial t} = \eta \omega B \kappa_1 \qquad \Gamma \text{ influence from A}$$
(4.6)

$$\frac{\partial}{\partial B} \cdot \frac{\partial C_{r_1}}{\partial t} = \eta \omega A \kappa_1 \qquad \Gamma \text{ influence from B}$$
(4.7)

These partial equations can then be aggregated for all reactions that contain the two species - taking the effect of species B on species C, for example, produces Equation 4.8. Using these aggregate sums, it is now possible to construct a pairwise relational matrix describing the influence each species has on every other species- Equation 4.9. This is known as the Jacobian matrix and is to solve the ordinary differential equations which describe the chemistry of the system (and propagate it forwards in time).

$$\mathbf{J}_{C,B} = \frac{\partial f(C)}{\partial B} = \frac{\partial}{\partial B} \cdot \left(\frac{\partial \Sigma_{r_1}}{\partial t} + \frac{\partial \Sigma_{r_2}}{\partial t} + \dots + \frac{\partial \Sigma_{r_n}}{\partial t}\right)$$
(4.8)

$$\mathbf{J}_{i,j} = \begin{bmatrix} \frac{\partial f_1}{\partial v_1} & \frac{\partial f_1}{\partial v_2} & \cdots & \frac{\partial f_1}{\partial v_n} \\ \frac{\partial f_2}{\partial v_1} & \frac{\partial f_2}{\partial v_2} & \cdots & \frac{\partial f_2}{\partial v_n} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial f_n}{\partial v_1} & \frac{\partial f_n}{\partial v_2} & \cdots & \frac{\partial f_n}{\partial v_n} \end{bmatrix}_{i,j=1}^{n,n}$$

$$(4.9)$$

4.5.1 Graph Construction Methodology For Simulated Data

Having covered the general definition of a Jacobian matrix and how it is constructed, we can now apply it to the context of mechanism analysis and comprehension. The first analogy that needs to be made is that for the flux is the change of a species concentration in time (the first differential with respect to time, d/dt). If we consider the change in a species concentration as a 'displacement', we can think of the flux as its 'velocity'. Similarly, the Jacobian provides us with a description of how the individual flux of a species changes concerning the concentration (or displacement) or another species (the second-order partial differential). This is analogous to the acceleration of the object or particle we first displaced. In using the Jacobian, we have constructed a relational matrix which outlines the effect a nominal change of a species has on all other species - a concept which is the foundation of the connectivity method (a mechanism reduction technique where all but essential species are removed) [Turányi and Tomlin, 2014].

Since the format of a Jacobian is already in the form of a relational matrix, it can easily be converted to a weighted adjacency matrix, and then directly into the graph format. Since it only considers the aggregated influence between species, much of the work that would otherwise be needed to convert a mechanism into a graph format has already been done. To make use of the Jacobian matrix, several extraction algorithms were written for an updated version of the Dynamically Simple Model of Atmospheric Chemical Complexity (DSMACC) [Emmerson and Evans, 2009; Ellis, 2020], as discussed in Chapter 1. Here we edit the kinetic pre-processor output, [Sandu and Sander, 2006] to release the values of the Jacobian Matrix and return them at each model timestep for analysis. The process for how this is done is described in Subsection 4.5.2.

A Note On Using The Flux Instead Of The Jacobian

Depending on the model setup or the users' capabilities, extraction of the Jacobian matrix for each timestep may not be possible. In many cases, the reaction rates and concentration may still be available, allowing for the calculation of reaction fluxes throughout the simulation. If this is the case, the total flux can be calculated using the method described in Equation 4.5. From this, an edge-weighted by a reaction flux can be created from every reactant to each product. This generates a multi-graph (A graph with multiple edges between nodes) which may be simplified by taking the net flux value for all edges between two nodes.

However, the potential for human/coding error, additional simplification and a non-explicit definition of the contribution of each species make the use of a Jacobian much more efficient in network generation from a chemical mechanism.

4.5.2 A Practical Example Using The MCM

Taking a single equation from the MCM, we may calculate the Jacobian relationships between species and convert them into a graph. A randomly chosen ethane reaction (Equation 4.10) from a simple mechanism was chosen. In general, the reaction consists of the following two steps: $C_2H_6 + OH \xrightarrow{\kappa_1} C_2H_5 \cdot H_2O$ and $C_2H_5 \cdot + O_2 \xrightarrow{\kappa_2} CH_5O_2$.

$$C_2H_6 + OH \xrightarrow{\kappa_3} C_2H_5O_2 \tag{4.10}$$

For simplicity, in this example, this will be the only equation for our mechanism. The resultant Flux Equation 4.11 and resultant Jacobian Equation 4.12 may be calculated.

$$\Gamma = [C_2 H_6][OH]\kappa_1 \tag{4.11}$$

$$\mathbf{J}_{i,j} = \begin{bmatrix} \frac{\partial f_{[C_2H_6]}}{\partial t \, \partial [C_2H_6]} & \frac{\partial f_{[C_2H_6]}}{\partial t [OH]} & \frac{\partial f_{[C_2H_6]}}{\partial t [C_2H_5O_2]} \\ \frac{\partial f_{[OH]}}{\partial t \, \partial [C_2H_6]} & \frac{\partial f_{[OH]}}{\partial t \, \partial [OH]} & \frac{\partial f_{[OH]}}{\partial t \, \partial [C_2H_5O_2]} \\ \frac{\partial f_{[C_2H_5O_2]}}{\partial t \, \partial [C_2H_6]} & \frac{\partial f_{[C_2H_5O_2]}}{\partial t \, \partial [OH]} & \frac{\partial f_{[C_2H_5O_2]}}{\partial t \, \partial [C_2H_5O_2]} \end{bmatrix}_{i,j=1}^{3,3}$$

$$(4.12)$$

Since not all species react with all other species, $(C_2H_6 \text{ does not react with } C_2H_5O_2)$ we can remove reactions that do not exist. To calculate the second differential, we begin by taking the flux of our equation:

$$\frac{dC_2H_5O_2}{dt} = \kappa_3[C_2H_6][OH]$$
(4.13)

Using this we can calculate the partial differential equations for OH and C_2H_6 :

$$\frac{\partial^2 C_2 H_5 O_2}{\partial t \ \partial OH} = \kappa_3 [C_2 H_6] \tag{4.14}$$

$$\frac{\partial^2 C_2 H_5 O_2}{\partial t \ \partial C_2 H_6} = \kappa_3 [OH] \tag{4.15}$$

This forms a 'sparse' Jacobian. Substituting numbers from subset mechanisms containing the methane and ethane precursors, we get Equation 4.17.

$$\mathbf{J}_{i,j} = \begin{bmatrix} \frac{\partial f_{[C_2H_6]}}{\partial t \ \partial [C_2H_6]} & -\kappa_3[C_3H_6] & \kappa_3[C_3H_6] \\ -\kappa_3[OH] & \frac{\partial f_{[OH]}}{\partial t \ \partial [OH]} & \kappa_3[OH] \\ & & \frac{\partial f_{[C_2H_5O_2]}}{\partial t \ \partial [C_2H_5O_2]} \end{bmatrix}_{i,j=1}^{3,3}$$
(4.16)

$$\mathbf{J}_{i,j} = \begin{bmatrix} \frac{\partial f_{[C_2H_6]}}{\partial t \ \partial [C_2H_6]} & -2 \times 10^{-7} & 2 \times 10^{-7} \\ -0.1 & \frac{\partial f_{[OH]}}{\partial t \ \partial [OH]} & 0.1 \\ & & \frac{\partial f_{[C_2H_5O_2]}}{\partial t \ \partial [C_2H_5O_2]} \end{bmatrix}_{i,j=1}^{3,3}$$
(4.17)

This creates the Jacobian - a matrix representing the reactions of a mechanism. Here we can calculate the production of a species, by summing of its column (except the diagonal), or its loss (from reacting to produce other species) by summing the row. This relational matrix can be used to generate a weighted graph of the chemistry (Figure 4.15).



Figure 4.15: A graphical representation of Equation 4.17 derrived from the Equation 4.10

Since any loss edges contain a negative value (orange numbers), it is possible to reverse the direction of the links to produce a positive edge of the same value (Figure 4.16).



Figure 4.16: Reversing the directions on negatively weighted edges from Figure 4.15

After reversing the links, we see that concentration for the reaction between C_2H_6 and OH follow the paths:

$$OH \xrightarrow{C_2H_6} C_2H_5O_2 \tag{4.18}$$

$$C_2H_6 \xrightarrow[2 \times_{10}^{-7}]{OH} C_2H_5O_2$$

$$(4.19)$$

As links in the graph are of the same units, we can simplify the equations by propagating the values of each edge. This results in a graph with only one link between each product reactant pair (Figure 4.17). It is worth noting that although this method of simplification produces a more intuitive graph, eigenvector metrics such as PageRank automatically transfer the 'flow' of information through the system to produce the same result.



Figure 4.17: Simplifying Figure 4.16

4.6 Case Study

below.

In this section, the centrality metrics discussed in Section 4.3 are applied to a range of scenarios. These range from polluted urban environments such as London [Bandy et al., 2012] and Beijing Fleming et al. [2017], to marine and terrestrial forest- Cape Verde [Read, 2010] and Borneo [Hewitt and Edwards, 2009]. We determine the main drivers for the chemistry and compare the species which are important across each simulation.

4.6.1 Establishing Initial Conditions From Observational Data

Within experimental data assimilation, it is not uncommon to face problems which result in unreliable or missing data. These can range from anything as little as measuring below the instrument sensitivity to powercuts and equipment damage/theft from the local wildlife. This can result in problems when analysing the results and combining them to create a simulation of the chemistry for that environment. To overcome this, traditionally a combination of data filtration, smoothing and interpolation is required. Although it is possible to fit a diurnal profile, through iterative methods of comparison, and cubic splines, it is more straightforward to implement the method (especially if more data will be added at a later date) is through the use of a Multi-Layer Perceptron Regressor model (MLPRegressor) as provided by the Python package Scikit-Learn, [Pedregosa et al., 2011]. This is described

4.6.1.1 The Origin Of Artificial Neural Networks

The concept of a neural network originated within the field of neuroscience. In biological neurons, signals are sent through the use of electrical impulses using their synapses. When a sufficient number of signals are received within a short timeframe, a neurone will respond, often firing a range of its signals. Using this as a foundation, McCulloch and Pitts [1943] presented a computational model of the biological neuron - the artificial neuron. This has a series of binary inputs and produces a single binary output. This idea was later improved with the invention of the perceptron - a linear classifier which classifies categories by separating them with a straight line. Invented by Rosenblatt [1958], this was popularised as a device representative of a modern-day shallow neural network - [John Hay, 1960], Figure 4.18. Unlike the artificial neuron, however, the perceptron can take non-binary (numerical) inputs of an associated weight which allows for the computation of simple linear binary classification. Much like Logistic regression, the perceptron produces a positive or negative classification based on a certain threshold¹³.



Figure 4.18: **The Mark 1 perceptron** Both software and hardware are different manifestations of a flow chart. The perceptron hardware accomplished what is now done using software. Source: Cornell [2020]

4.6.1.2 The Multi-Layer Perceptron

Limitations of the perceptron include the classification of complex patterns such as the XOR problem (where a category appears between two other categories, e.g. 1|0|1 - this cannot be classified by a

 $^{^{13}}$ It is worth noting that while a Logistic Regression classifier can output a class probability, the use of a hard threshold means that this is not done within the perceptron algorithm [Géron, 2017]

single linear split). In taking inspiration from nature (Figure 4.19) it is possible to overcome this with the use of multiple layers. This creates a deep (> 2 two hidden (non-input) layers of perceptrons¹⁴) artificial neural network (ANN)

The multi-layer perceptron (MLP) model now represents a simple feed-forward network, much like a decision tree. However, unlike a decision tree, the MLP ANN can describe the probability a branch is taken using non-linear activation (threshold) functions. These are discussed in detail as part of Subsection 6.3.5. The weighting thresholds for each neuron are then calculated by backwards propagation of results through the network until a suitably good result is produced.

Example analogy: Backpropagation can be likened to the iterative calibration of scientific instrumentation. In the field of atmospheric chemistry, laser-induced fluorescence is used to measure species concentrations and reaction rates within the troposphere, [Dillon et al., 2006; Bloss et al., 2004]. Here the frequency of a laser can be tuned to a resonant frequency of a known target (e.g. OH, NO_2 and SO_2) to produce a response curve. Similarly, a neural network can be 'trained' (calibrated). This is done through the use of a 'training dataset' - a set of input-output pairings which represent a random selection of 2/3rds of the total dataset. Next, the neurons within each layer (similar to the potentiometer dials on an instrument) are adjusted in sequence through the layers to match the known result (a standard of known concentration) to the input values provided. This process is repeated until for many iterations, or until a sufficiently 'good' prediction is attained for the entire training dataset (early termination). The power of ANNs comes from the ability to adjust neuron thresholds whilst moving both forwards and backwards through the network (Note: predictions of an MLP are still only passed forwards). Finally, model performance is evaluated against the remaining 1/3rd of the total dataset.

¹⁴These are sometimes referred to as Linear Threshold Units.



Figure 4.19: The Human Cortex - A biological neural network. A vertical cross section of the human cortex between an adult (top) and 1.5 month old infant (bottom) showing a layer like structure with a change in depth (left to right). Source: Cajal [2020]

4.6.1.3 Applying The Mlpregressor To Observational Data

In the application of any type of machine aided algorithms, it is important to evaluate the results provided. In this section data collected from Cape Verde ([Read, 2010]) containing 12 years of observations are shown. A MLPRegressor of 10 hidden layers, and a hyperbolic tan (tanh) activation function is used Section B.4. Additionally, the limited-memory Broyden–Fletcher–Goldfarb–Shanno (l-BFGS) solver (a quasi-newton method which minimises the inverse of the Hessian matrix¹⁵ to steer through space and obtain a solution) and an adaptive learning rate¹⁶ is used.

The input of the regressor is in the form of a month and an hour, to represent each measurement. This allows it to find not only daily trends but also seasonal trends within the data. Once trained, the regressor is then used to predict a diurnal profile for each month based on the observational data provided. For simplicity \log_{10} values of the concentrations obtained have been used. The predicted MLPRegressor line is compared to a transparent scatterplot for all the results. In addition to this, a boxplot showing the Inter Quartile Range (The range between the 25th and 75th percentile), median and mean (green line) plotted alongside to evaluate the predictor output. In this study, we only take the values for the month of June (or closest available depending on the dataset).

In providing the MLPRegressor with both month and hour inputs, the data is not only fitted hourly

 $^{^{15}}$ The hessian is a square matrix of second-order partial derivatives of a scalar-valued function/field describing the local curvature of a function (of many variables).

 $^{^{16}}$ Each time the model improvement fails to decrease the learning loss, the learning rate is reduced by 1/5. This means smaller jumps are made towards the curve peak.

(a diurnal average) but also across the seasonal/monthly cycles. This accounts for the variation between years and datasets. Since \log_{10} values of the concentrations are used, species such as ozone (Figure 4.20) which for the Cape Verde dataset (clean air) do not change more than one order of magnitude, the effects of neighbouring months, which shift the diurnal away from the mean (the green line on the boxplot), can be seen. However, since this is overall a small change, and the diurnals lie within the interquartile range, they still provide an adequate approximation. NO (Figure 4.21) on the other hand, has a concentration change of several orders of magnitude. Here a distinct daytime peak is seen and is centred around a seasonally consistent mean value of the data. Here the multi-magnitude change in concentration also provides a striking silhouette of the data to which we may compare the fitted line. Finally the plots of NO₂ and iso-Pentane (Figure 4.22-4.23) vary both in diurnal magnitude and seasonally. Within these plots, changes in the data in the January and December months produce deceptively misleading results. Here although the diurnals are not symmetric, they fit well within the median, mean and interquartile range values, as well as the general data silhouette behind them. This suggests that it is a property of the data that we are fitting, and not that the regressor is producing incorrect results. It is however noted that for a more accurate seasonal prediction, periodic boundary conditions should be employed in the training dataset, where an additional two months are added before January and after December. As only a single value estimate from the summer region will be taken, this does not affect the result accuracy.



Figure 4.20: Cape Verde MLP predicted and observational data of Ozone (mixing ratio). Each segment represents data from a different month. Within each month segment exists 24 hour segments to create a diurnal. Observational concentrations are plotted in the form of a translucent scatterplot and summarised using the boxplot on the right of the month segment. MLP predicted results are shown using the solid lines. Concentration in mixing ratio.



Figure 4.21: Cape Verde MLP predicted and observational data of NO (mixing ratio). Each segment represents data from a different month. Within each month segment exists 24 hour segments to create a diurnal. Observational concentrations are plotted in the form of a translucent scatterplot and summarised using the boxplot on the right of the month segment. MLP predicted results are shown using the solid lines. Concentration in mixing ratio.


Figure 4.22: Cape Verde MLP predicted and observational data of NO_2 (mixing ratio). Each segment represents data from a different month. Within each month segment exists 24 hour segments to create a diurnal. Observational concentrations are plotted in the form of a translucent scatterplot and summarised using the boxplot on the right of the month segment. MLP predicted results are shown using the solid lines. Concentration in mixing ratio.



Figure 4.23: Cape Verde MLP predicted and observational data of iso-Pentane. Each segment represents data from a different month. Within each month segment exists 24 hour segments to create a diurnal. Observational concentrations are plotted in the form of a translucent scatterplot and summarised using the boxplot on the right of the month segment. MLP predicted results are shown using the solid lines. Concentration in mixing ratio.

4.6.1.4 Model Initialisation Procedure

The aim is to generate a set of initiation conditions which are representative of the species found for different environments around the world. In this section, we are not interested in the exact concentration modelling for specific times or scenarios. Instead, we seek to generate representative of the processed chemistry under a range of conditions.

Species concentrations are extracted from an MLP regressor trained on observational data for each scenario. Each concentration is that of noon local time from the generated diurnal from summer observations at each location. This produces a monthly error of $\pm 2months$ from June. As both nitrogen oxide and dioxide are supplied, the total NO_x for each simulation are *not* constrained. The initial conditions are shown in Table 4.4.

In general observational measurements are not able to detect all the species presented within the MCM. This means that to be able to compare model scenarios, the chemistry must first be spun up. In propagating the chemistry forwards in time, primarily emitted and measured species are broken up forming the intermediate species which exist within a mechanism. To reach a steady-state, the model is initiated at noon, and the observational concentrations are rest every 24 hours. For each diurnal, the fractional difference between the concentrations at each day are compared. If the difference between these is less than 0.001, the model is left to run unconstrained for five days (right of the dashed line in Figure 4.25). Model results are then taken after three days of unconstrained runs. The reason for this is that the total RO_2 concentration takes longer to stabilise in the polluted environments (London and Beijing). This falls into a periodic cycle beginning noon on the third day and can provide a representation of the processed chemistry within each environment.

NOTE: It should be noted that some of the concentration plots may appear to lose their diurnal dependability. This may be attributed to the changing order of magnitude of the concentrations, and that the species are still responding as expected.

4.6.1.5 Extracting The Required Results

Model diagnostics such as concentration and the net flux passing through a species may be extracted directly from the DSMACC box model. These provide the baseline comparison and can be directly compared to the graph metrics. Species concentration tells us the abundance of different species, and the net-flux tells us how fast this is changing in time.

As some species may have a fast inwards and outwards flux (low net-flux), the absolute flux is also included. Finally, the sensitivity of each species for other species is also extracted (the Jacobian matrix). This serves to not only generate the graph used to represent the chemistry, (Subsection 4.5.1) but also to identify the overall influence a species has on others in the network. This can be calculated by taking the net sum of the influence a species has on every other from the Jacobian. This is analogous to calculating the out-degree of a node in the Jacobian network.

Species	Beijing(APHH)	Borneo(OP3)	London(ClearFlo)	CapeVerde
LAT	39.9	0.96	51.0	16.5
LON	116.3	114.5	0.00	23.4
СО	3.829e-06	3.321e-07	7.780e-09	0.0*
O_3	6.883e-08	8.939e-09	3.819e-08	2.629e-11*
NO	1.660e-09	2.668e-14*	2.350e-09	2.358e-12
NO_2	1.226e-08	$1.081e-13^*$	7.445e-09	8.447e-12
HCHO	4.472e-09		1.119e-08	
C_2H_6	3.163e-09	7.315e-10	2.133e-09	4.539e-10
C_2H_4	1.004e-09	1.152e-10	4.893e-10	2.481e-11
C_3H_8	3.019e-09	1.924e-10	1.128e-09	1.728e-11
C_3H_6	1.335e-10	1.333e-11	1.784e-10	9.343e-12
IC_4H_{10}	6.412e-10	8.742e-11	5.142e-10	2.486e-12
NC_4H_{10}	1.593e-09	5.698e-11	1.058e-09	4.481e-12
C_2H_2	1.058e-09	1.825e-10	3.018e-10	1.848e-11
TBUT2ENE	4.198e-11		1.815e-11	
CBUT2ENE	4.454e-11		1.305e-11	
IC_5H_{12}	1.047e-09	2.883e-11	7.424e-10	3.470e-12
NC_5H_{12}	4.650e-10	2.090e-11	2.792e-10	2.513e-12
TPENT2ENE	3.939e-11			
CPENT2ENE	3.982e-11			
NC_6H_{14}	2.057e-10	6.437e-12	6.357e-11	
C_5H_8	7.134e-10	1.957 e-09	1.640e-10	
NC_7H_{16}	7.905e-11		5.222e-11	
BENZENE	4.045e-10		1.137e-10	7.682e-12
NC_8H_{18}	3.091e-11		1.442e-11	
TOLUENE	6.767e-10		3.205e-10	3.121e-12
EBENZ	3.115e-10		6.017e-11	
OXYL	1.677e-10		5.049e-11	
CH_3CHO	4.783e-10		4.095e-09	
C_2H_5OH	4.655e-09		3.125e-09	
CH_3COCH_3	3.328e-09		2.924e-09	
NC_9H_{20}	1.336e-11		7.922e-11	
$NC_{10}H_{22}$	1.062e-12		1.602e-10	
α -PINENE ¹⁷	7.341e-11	15e-11	1.105e-10	
LIMONENE	5.836e-11	1.351e-10	3.566e-11	
$\mathrm{PXYL}^+\mathrm{MXYL}^{18}$	4.943e-10			
IPBENZ	4.567 e-10			
PBENZ	3.996e-10			
HONO	6.479e-10		4.109e-10	
MACR		6.948e-11	1.862e-11	

Species	$\operatorname{Beijing}(\operatorname{APHH})$	Borneo(OP3)	London(ClearFlo)	CapeVerde
PENT1ENE			2.383e-11	
MVK			2.091e-11	
NPROPOL			2.883e-10	
NBUTOL			4.535e-10	
STYRENE			2.241e-11	
MEK			5.494e-11	
C_3H_7CHO			$9.534e{-}12$	
C_4H_9CHO			1.865e-11	
$C_5H_{11}CHO$			1.201e-11	
CYHEXONE			9.790e-12	
BENZAL			1.510e-11	
PAN			1.791e-10	

Table 4.4: (2-page split) The initial conditions created from the MLPRegressor prediction of observational data. Although not specified the mixing ratios for methane is set by the model at 1770ppb, the temperature is 298K, and water vapour is at 2%. * Starred values are of the wrong units and should be multiplied by 1000. As there was no time to rerun these, their results have been omitted from this chapter.

 $^{^{18}}$ This is (incorrectly) written as '?-pinene' in the merged CEDA dataset for the Borneo OP3 campaign. This is due to character conversion errors.

 $^{^{18}\}mathrm{The}$ mixing ratios for these is split evenly between both species



Figure 4.24: The mixing ratio profile for London. This shows a the change in mixing ratio over time for HO_x, NO_x , HCHO, Ozone and RO_2 species for a simulation run generated by the mlpregressor. Left of the dashed line shows the last 6 days of spinup, where the values are reset at noon each day until the species fractional difference is less than 0.001.



Figure 4.25: The mixing ratio profile for Beijing. This shows the change in mixing ratio over time for HO_x , NO_x , HCHO, Ozone and RO_2 species for a simulation run generated by the MLPRegressor. Left of the dashed line shows the last six days of spinup, where the initial values are reset at noon each day until the species fractional difference is less than 0.001.

4.6.1.6 Unifying The Results

Each metric provides a different range in which it ranks the importance of a node. All results are scaled to the range $\{0,1\}$, where one is the highest. Entries, where the results span several orders of magnitude (e.g. concentration, flux, influence), are flattened using the \log_{10} scale before being normalised.

4.6.2 Comparing Results

This subsection juxtaposes the use of traditional model diagnostic methods against a selection of graph metrics. As there are several thousand species within each simulation run, the keyword extraction algorithm Term Frequency - Inverse Document Frequency (TF-IDF), is used to identify the top most prominent species for each metric (traditional and graph). From this, the ten highest-ranking species from each category are collated into a single diagram for comparison.

4.6.2.1 What Is TF-IDF

TF-IDF is a numerical statistic used in text natural language processing and text mining. It is designed to identify the importance of a word concerning its context.

It provides a value for the frequency a word appears within a document, offset by the number of times it appears in other documents within the corpus - It is for this reason that 83% of text recommender systems in digital libraries use TF-IDF, [Beel et al., 2016].

In [Ellis, 2019] this was applied this to the chapters of Frankenstein and found the keywords extracted almost exactly replicated those from the synoptic description of the novel. Although TF-IDF is a text mining procedure, the algorithm itself is mathematical, meaning that it may be applied to our diagnostic dataset. The working of the algorithm is discussed below.

Term frequency

The TF from the algorithm name stands for term frequency. This is an analysis of the number of times a word exists within a dataset. There are several ways in which this can be done; these are:

- Raw Count The number of times a word exists within the document.
- Boolean/Logistic True if the word exists, false otherwise.
- Adjusted for Document Length word frequency/total number of words

As the scaled values for each item are taken, we can liken our results to the 'Adjusted for Document length' equation and use the scaled ranking value for each group.

Inverse Document Frequency

Inverse document frequency tells us how much information a word provides concerning a specific context. While a word may be used extensively throughout a set of documents (a corpus) - (i.e. term frequency), often that we are interested in the words which frequently appear only within an individual document. This is one of the reasons TF-IDF is useful in the extraction of keywords from a document.

The inverse frequency of a word is usually calculated as the log of the number of documents (N) divided by the number of documents the word appears in (D_f) , Equation 4.20.

$$IDF = \log(\frac{N}{D_f}) \tag{4.20}$$

If required, changes can be made to produce results which show a better representation of words which are important in all documents (probabilistic, Equation 4.21) or individually (smooth, Equation 4.22). However in looking at Figure 4.26, it can be seen that the basic IDF formula mentioned has a limit of zero the higher the document frequency (D_f) , which makes it easy to normalise against (divide by 2) as this is the value tended to if the document frequency tends to 0.

$$IDF_{prob} = \log(\frac{N - D_f}{D_f}) \tag{4.21}$$

$$IDF_{smooth} = \log(\frac{N}{1+D_f}) + 1 \tag{4.22}$$

To complete the TF-IDF equation, the frequency terms and inverse document frequency terms are multiplied together.

Applying TF-IDF to chemical metrics

To identify metrics selection criteria, we seek only species (words) which are important only for that metric. The TF-IDF algorithm can be adapted for use with the graph metric output. Here 'Term Frequency' corresponds to the number of times a value appears within the body of a document and can be seen as the scaled $\{0,1\}$ metric output. This is then divided by the log of the 'Inverse Document Frequency'. D_f is the sum of values across all the metrics. This makes the TF-IDF equation:



$$TF.IDF = metric_value \ . \ \log(\frac{N_o \ documents}{\Sigma_{\forall} \ metric \ values})$$
(4.23)

Figure 4.26: The different IDF outputs. A plot showing Inverse Document Frequency profiles against Document Frequency. This shows that the probabilistic IDF highlights words that are more important across all items, while the smooth IDF shows files which are more important individually. The general IDF (which is used) produces a result starting at two and tending to zero. This provides the best response and can easily be scaled between the range of [0,1] by dividing the output by 2. Source: [Mquantin, 2020]

4.6.2.2 Metric Comparison

This section aims to compare the efficiency of graph metrics against a list of traditional methods in determining species which play an important role within the network. To do this a bivariate colourmap (Figure 4.27) is used. Each figure consists of a red-hued image/heatmap representing the scaled values {0,1}:{white, red} for each of the individual metrics (graph columns). As each simulation contains thousands of species, only the top 10 species from each column/category are selected. These are then sorted by the average sum of their closeness, betweenness and page-rank values (blue column). Superimposed on this reds-only heatmap is a blue heatmap representing the average sum of the three metrics for comparison. Such a method allows for the comparison of individual values against an approximation of species importance, by the sum of graph metrics - allowing an easy categorisation of the data:

- Purple This value is high in both the individual category and the metric sum.
- Red This value is high for the individual category but not the metric sum.
- Blue This value is high for the metric sum but not the individual category.
- White This value is low for all categories.



Figure 4.27: The bivariate colourplot key.

4.6.2.3 Individual Categories

Individual categories are split between traditional metrics and graph centrality metrics. To represent the importance of a species, the following values may be extracted through the use of a simple box model:

- Concentration this describes the abundance of a species within the atmosphere.
- Net flux this describes the rate of net (absolute) change of concentration over time for a species.
- Absolute flux some species may have a large flux going through them (production and loss), resulting in a small net flux. This sums the production and loss fluxes.
- influence influence is the total magnitude of an effect that changing a species concentration by 1% would have on other species within the network. Since the graph is generated using the Jacobian matrix, an alternative method for calculating this can be by calculating the total out-degree of a node.

The importance of a species is then compared through the use of three of the most common centrality metrics. These are:

- **Closeness** This describes how easily information from one node can be disseminated to all other nodes (see Subsection 4.3.2).
- **betweenness** This describes the number of shortest paths (fastest fluxes/greatest influences) that are routed between nodes adjacent to our chosen node. Species with a high betweenness hold a brokering position and can act as a bottleneck between different groups of chemistry (see Subsection 4.3.3).
- **PageRank** PageRank looks at the flow in a system. It ranks nodes not only on the number of species it reacts with but also the importance of the species it has reacted with (see Subsection 4.3.5).

Finally, the 'Metric Sum' is the sum of all the metric values scaled between 1 and zero (the mean).

4.6.3 Scenario Analysis

In selecting the top 10 ranking species for each category, it is possible to examine if the importance of a species with centrality metrics varies from the results suggested by traditional metrics. In this subsection, we explore the TF-IDF rankings of each metric and use this to decide if species importance is local to a specific metric. We look at what species are highlighted by each scenario (Figures 4.30 - 4.31) and compare them against the primary emitted species shown in Table 4.4. Finally, we compare the total metric sum against the traditional metrics of concentration and flux and compare the correlation.

London

The London dataset (Figure 4.28) contains a mix of anthropogenic and biogenic aromatics and longchain alkanes. We have a section of alkanes which have a low overall metric sum and a small value for closeness and page rank. Combined with their high net flux, absolute flux and influence values, this suggests that they have a moderate directional flux, most likely influencing the production of many other species at a consistent rate. In addition to these, we have species with a moderate closeness but a high betweenness. These are often species such as formaldehyde (HCHO), glyoxal and acetaldehyde which can serve as tracers for fast photolytic reactions. This is because on the graph structure (Figure 6.4) they sit between the dense centre of the network (high closeness) and the branches formed from each primary emitted species (a low closeness value). Their high connection density and importance in the network is also picked up by the page rank algorithm. Other species with high betweenness and a low centrality are the monoterpenes limonene and α pinene, as well as hexane (NC₆H₁₄) and butane products. These are (or are close to) primary emitted species and therefore have a low closeness. Since much of the chemistry originates with such species, the outward 'flow' of information also results in a lower page rank value.



Figure 4.28: An example force graph showing the complex chemistry of London. (Weightings not from initial conditions described.) Source: [Lewis, 2018]

Beijing

Similar to London, the fast photochemical tracers are identified, although some have a slightly lower flux between them (Betweenness) and page rank values for Beijing (Figure 4.29). This suggests that the network structure or weightings may have shifted slightly from London, creating more links, or importance in a specific branch of chemistry. Additionally, their overall metric sum is lower. Glyoxal, Methyl Vinyl Ketone (MVK) and their associated criegee configurations all feature heavily in the middle of Figure 4.31. These are important as they represent the fast chemistry formed by both the anthropogenic and biogenic chemistry that is within the simulation. These tend to have a high closeness and page rank centrality, a pattern that is also seen with the long-chain alkane products from Octane $(n-C_8H_{18})$, Hexane $(n-C_6H_{14})$ and Isoprene.



Figure 4.29: An example force graph showing the complex chemistry of Beijing. (Weightings not from initial conditions described.)



Figure 4.30: A bivariate heatmap comparison of London.



Figure 4.31: A bivariate heatmap comparison of Beijing.

4.6.4 Providing An Overall Overview Using The TF-IDF And The Metric Sum.

In the previous section, it was shown that centrality metrics could be used to complement the use of traditional metrics in the analysis of the chemical network. As each metric represents a different aspect of importance, should a single ranking value for a node be required, it is possible to take the average sum of all three metric values. Looking at Figure 4.31 it is possible to see similar trends in colour gradient between the purples of the traditional metrics of flux and concentration with the total metric sum (the blue column). This suggests that it is possible to compare each scenario with the use of the metric sum.

In selecting the ten highest-ranking species from the mean centrality metric table for each simulation, Table 4.5 can be created. Unlike the previous method, we are now looking at species which are essential across all metrics in a simulation. Beijing consists mainly of Quinones and Dialdehydes, which are both derivatives of Benzene. London again has Benzene related compounds, mixed with the fast photochemical indicators, which were also ranked highly in Figure 4.30. Looking at the highestranking sum (NaN-mean), it is seen that Isoprene, hept/hexane and glyoxal products highlighted as the most consistently important across all four simulations.

London	Beijing	
HCHO	PTLQONE	
CH3CHO	PBZQONE	
C5CO14OOH	HOHOC4DIAL	
PBZQOOH	MNNCATCOOH	
MALANHY	C6H5CO3H	
CH3CO3	EPXDLPAN	
C57OH	C5DIALO	
C624CHO	NBZFUOOH	
GLYOX	TLBIPEROOH	
HCOCOHCO3	NCRESOOH	

Table 4.5: A table of the top 10 ranked species for each simulation. Only species that exist within at least 3 out of the four simulations are used. The Nan-Mean takes the mean of all available data, ignoring runs where a species is not present. Species presented within the table follow the MCM naming convention.

A note on finding the precursors

Graphs are also useful in the back navigation of a network. It is possible to discover the most probable primary emitted species (nodes with no in-degree) by comparing the shortest path lengths for all primary emitted species (not including inorganic species). Here the primary emitted species with the smallest number of connections are often the most likely source.

4.7 Causality Analysis using the Jacobian and Pagerank

Due to the complex inter-dependency on species within an atmospheric network, it is often hard to calculate how much of an effect perturbing one species will have on another. Classically we can find the instantaneous change at a specific point in time by close examination of the Jacobian matrix - also known as the connectivity method. Here Turányi and Tomlin [2014] states that the value $J_{A,B}$ of a log-normalised Jacobian represents the effect changing a species A will have on species B - allowing us to determine a set of important species with an aim (e.g. mechanism reduction).

If we take the sum of all items in the B column $(\sum_{i\neq B} J[i, B])$ we can calculate the total production of species B. Expanding this method we can calculate the percentage importance of B's precursors. Here the species with the highest values present the most significance to the formation of B. If we were to look at the rows instead of columns, a similar analysis may be executed for a species loss. It is for this reason that in preparing the Jacobian for use within a graph, we take the net difference between production and loss contributions to determine edge direction and weight.

4.7.1 Source Analysis using PageRank

The PageRank value of a species can be calculated through the solving of eigenvalues and eigenvectors of a google matrix, or propagating a vector of unity through the network in small increments (Equation 2). These solutions are both very similar to the integrator we use to propagate the chemistry within our box model - where instead of concentration we move 'information' between nodes in our network. This means that for a network of nodes {A, B, C, D, E} (Figure 4.32a) we can use page rank to determine where the flow of information ends up (ultimately E). However, if we are interested to find out how much each of these nodes contributes to the total amount of information gained by E, we need to look at the whole sequence in reverse. For a directed-graph, we may reverse the direction of the links (so that source \rightarrow target becomes source \leftarrow target) and then assign an arbitrary amount of information to our queryable node (E) and then propagate it forwards using page rank on our reversed network using the same edge weighting as before, Figure 4.32b.

4.7.2 Mathematical Equivalent of Edge Reversal

The physical 'flipping' of an edge direction is the most visually intuitive method reversing the direction, however, in Chapter 3 we discussed that the 2D relational data of a network may also be represented through the use of an adjacency matrix. Since this is a matrix of links where row elements \rightarrow column elements, the reversal of these edges is mathematically identical to taking the transpose of the matrix. In terms of our chemistry, our network is derived using the net values of our second order, partial derivative, relational matrix - the Jacobian. This makes the link reversal procedure equivalent to transposing the Jacobian, and thus analogous to the creation of the adjoint. The adjoint matrix is often used within the modelling world, to run backwards in time and make historic predictions based on current data [Henze et al., 2007]. This has its application in determining the geolocational source of pollution, or in our case the source of concentration change in a species of interest.



(b) Reversed-link (adjoint) influence graph

Figure 4.32: Link reversal of the Jacobian Sensitivity matrix graph results in a graph of the adjoint. Showing changing the direction of links in a graph is equivalent to applying the transpose to an adjacency matrix (right). In the case of a Jacobian based graph, this is analogous to using the adjoint to propagate the model back in time - something that can be used to identify the influence upon a species with a model.

4.7.3 A Calculated Case Study

To illustrate the points made within this section we select a species (MCM name: NC101CO) from a sample simulation of the MCM (Borneo initial conditions are given in Table 4.4). A brief mechanism analysis allows us to determine the precursors of NC₁₀₁CO by isolating its local sub-graph and reversing it. This is done by looking at outwards arrows in Figure 4.33.



Figure 4.33: The reversed subgraph between α -pinene, and NC101CO. This is a subgraph showing the production of NC₁₀₁CO. Arrows point towards a species precursor.

4.7.3.1 Personalised PageRank

Using the reversed network we apply a 'personalised' version of PageRank. This means that we initiate NC101CO with a ranking vector of 1000000 and -1 for all other species. A damping factor of 0.01 is also used within the algorithm to produce the results in Table 4.6. As is often seen within the algorithm, top-scoring values often have a notable split from mid and lower scoring values. We find that NC101CO has the strongest influence on itself (it is where we start our information), followed

that of α -pinene - its primary emitted precursor. Other influences such as NAPINBOOH, NAPINBO, NAPINBO₂ are seen, where NAPINBO has twice the influence from the others. This is most likely as this has the highest net-flux from the model (Table 4.7).

Species	PageRank Ranking
NC101CO	9.920000e-01
APINENE	9.210000e-06
NAPINBO	4.540000e-03
NAPINBO2	2.770000e-03
NAPINBOOH	2.690000e-03
C511OOH	-9.990000e-07
C527NO3	-9.990000e-07

Table 4.6: A reversed graph Page Rank test with $NC_{101}CO$

Species	$\dot{v} \ ({ m net} \ { m flux})$
NAPINBOOH	-0.458271
NAPINBO	-1840.391917
NAPINBO ₂	-0.037366

Table 4.7: The net flux of the three species: NAPINBOOH, NAPINBO, NAPINBO₂, taken from a simulation of Borneo at 2020-06-24 12:09:56

4.7.3.2 Iterative Analysis of the Jacobian

The total production of a species was also calculated by taking the sum of all items in the corresponding column (not including the diagonal), Table 4.8. As we are ignoring the diagonal¹⁹ the first generation influence matches that of the reversed page rank algorithm. To get the total influence however we now have to repeat the process for each of the species producing NC101CO, and so forth. This process is identical to the iterative sum of in-degree links going backwards through our jacobian derived network - a process visually illustrated within Figure 4.34.

$J_{precursor->NC101CO}$	Value
$\begin{array}{l} \text{NAPINBO} \longrightarrow \text{NC}_{101}\text{CO} \\ \text{NAPINBO}_2 \longrightarrow \text{NC}_{101}\text{CO} \\ \text{NAPINBOOH} \longrightarrow \text{NC}_{101}\text{CO} \end{array}$	$\begin{array}{c} 47123.629418\\ 0.000014\\ 0.000001\end{array}$

Table 4.8: The influence on $NC_{101}CO$ from other species, taken from a simulation of Borneo at 2020-06-24 12:09:56

 $^{^{19}}$ The diagonal of a Jacobian tells us the net change of a species, which can also be calculated as the difference of the column and row sums.

4.7.4 Verdict

As the PageRank algorithm is applied to the whole network and contains teleportation, it provides small values for species without a direct link to the species in question. This requires some sort of changepoint analysis to filter.

A better method would be the calculation of the shortest simple path between a species in question and all other species, and then subtract the value obtained within each step to get its contribution. For the example $A \xrightarrow{4} B \xrightarrow{6} C$ the shortest path from A to C would be ten and B to c would be 6. The influence of A on C is calculated by the influence of A on B divided by the total influence on B. This is similar to the process required if we were to use the Jacobian Matrix directly.

Overall, as we use the Jacobian matrix to create the graph, writing a script to process this may be the most computationally efficient (and comparable to other results). The graph representation and page rank methods may at first glance be quicker to implement, but do not necessarily produce numerical output that is immediately interpretable. They do however illustrate what the more traditional (mathematical) methods are doing in the background. Therefore both methods are equally difficult to implement, as well as being useful, and therefore it is up to the reader to select the one most appropriate to their aims.

4.8 Conclusions

Chapter 2 and Chapter 3 explained the importance of visualisation and graph structure in communicating the complexity of an atmospheric chemical system. This chapter explores using a graph framework in cases where a network is too large or complex to represent visually.

We start by looking at several centrality metrics, which categorise nodes of importance within the network. Despite showing great promise in other fields, these do not seem to supply significant new insights into the chemical mechanisms of a model simulation when compared to traditional methods. Next, we look at the classification of the MCM as a network. Here it is shown that this contains both small world and power-law features. This means that the network has both a local and global structure. The MCM network consists of a series of small 'communities' of species which react with each other, connected by a hierarchical structure - a structure which can be utilised in the use of mechanism reduction (as seen in Figure 2.16).



Figure 4.34: Showing total the influence from each species on HCHO for a sample MCM subset of Butane. Species importance (node size) is determined using the reverse PageRank algorithm starting at HCHO (middle). It is calculated by taking a snapshot of a chemical simulation and rendered using the transpose of the Jacobian relational matrix. Link width is representative of the cumulative sum of the weights contributing to a concentration change of HCHO.

Finally, it is possible to apply graph theory for the use of source analysis. Although this is analogous to manual matrix methods on the Jacobian, the use of a graph structure makes the explanation and understanding of the procedure more intuitive. Again there is no gain over the traditional Jacobian analysis methods.

To conclude, the graph-based analysis offers much of the same results as current methods. Since we do not obtain any new significant insight in applying these the new learning curve in embedding them into current practices of model analysis does not appear worthwhile. Instead, we further explore the modular structure of the MCM in Chapter 5, where graph clustering techniques are applied to group species with a fast-flux between them.

Bibliography

- Aumont, B., Szopa, S., and Madronich, S. (2005). Modelling The Evolution Of Organic Carbon During Its Gas-Phase Tropospheric Oxidation: Development Of An Explicit Model Based On A Self Generating Approach. Atmospheric Chemistry and Physics, 5:2497–2517. https: //www.atmos-chem-phys.net/5/2497/2005/acp-5-2497-2005.pdf.
- Bandy, B., Belcher, S. E., Bloss, W., Blyth, A., Faloon, K., Finessi, E., Gallagher, M. W., Grimmond, C. S. B., Herndon, S., Laufs, S., Lee, J. D., Leigh, R. J., Liu, D., Monks, P. S., Nemitz, E., Reeves, C. E., Oram, D., Sokhi, R., Young, D., Visser, S., Whitehead, J., and Zotter, P. (2012). Dataset Collection Record: Clearflo (Clean Air For London) Project: Meteorology, Composition And Particulate Loading Measurements Of London'S Urban Atmosphere. https://catalogue.ceda. ac.uk/uuid/cee49a1f044b79d5413b7a0282467508.
- Barabási, A.-L. (2019). Nature-150-Cover.Pdf. Nature, 575(7781). https://www.nature.com/ immersive/d42859-019-00121-0/public/pdf/nature-150-cover.pdf.
- Bastian, M., Heymann, S., and Jacomy, M. (2009). Gephi: An open source software for exploring and manipulating networks. AAAI. http://www.aaai.org/ocs/index.php/ICWSM/09/paper/view/154.
- Beck, H. (2017). London Icon: A History Of Harry Beck'S Iconic Tube Map. https://londontopia. net/site-news/featured/london-icon-tube-map/.
- Beel, J., Gipp, B., Langer, S., and Breitinger, C. (2016). Research-paper recommender systems: A literature survey. International Journal on Digital Libraries, 17(4):305–338. https://doi.org/10. 1007/s00799-015-0156-0.
- Bloss, C., Wagner, V., Jenkin, M. E., Volkamer, R., Bloss, W. J., Lee, J. D., Heard, D. E., Wirtz, K., Martin-Reviejo, M., Rea, G., Wenger, J. C., and Pilling, M. J. (2005). Development of a detailed chemical mechanism (mcmv3.1) for the atmospheric oxidation of aromatic hydrocarbons. *Atmospheric Chemistry and Physics*, 5(3):641–664. https://www.atmos-chem-phys.net/5/641/2005/.
- Bloss, W. J., Lee, J. D., Bloss, C., Heard, D. E., Pilling, M. J., Wirtz, K., Martin-Reviejo, M., and Siese, M. (2004). Validation of the calibration of a laser-induced fluorescence instrument for the measurement of oh radicals in the atmosphere. *Atmospheric Chemistry and Physics*, 4(2):571–583. https://www.atmos-chem-phys.net/4/571/2004/.
- Bonacich, P. (1987). Power and centrality: A family of measures. American Journal of Sociology, 92(5):1170–1182. http://www.jstor.org/stable/2780000.

- Bonacich, P. (2007). Some unique properties of eigenvector centrality. Social Networks, 29(4):555 564. http://www.sciencedirect.com/science/article/pii/S0378873307000342.
- Borgatti, S. P. (2005). Centrality And Network Flow. Social networks, 27(1):55–71. http://www.sciencedirect.com/science/article/pii/S0378873304000693.
- Boudin, F. (2013). A Comparison Of Centrality Measures For Graph-Based Keyphrase Extraction. online. https://hal.archives-ouvertes.fr/hal-00850187/document.
- Brandes, U. (2001). A faster algorithm for betweenness centrality. *Journal of Mathematical Sociology*, 25:163–177.
- Brasseur, G. and Jacob, D. (2017). *Modeling Of Atmospheric Chemistry*. Cambridge University Press. https://books.google.co.uk/books?id=k9 PDgAAQBAJ.
- Broido, A. D. and Clauset, A. (2019). Scale-Free Networks Are Rare. Nature communications, 10(1):1017. http://dx.doi.org/10.1038/s41467-019-08746-5.
- Cabello, R. (2019). Three.Js Javascript 3D Library. https://threejs.org/.
- Cajal, S. R. (2020). Cortex drawings. *web.* https://upload.wikimedia.org/wikipedia/commons/5/5b/ Cajal_cortex_drawings.png.
- Clauset, A., Shalizi, C. R., and Newman, M. E. J. (2009). Power-Law Distributions In Empirical Data. SIAM Review, 51(4):661–703. https://doi.org/10.1137/070710111.
- Cornell, L. (2020). Mark 1 Perceptron. https://en.wikipedia.org/w/index.php?title=Perceptron& oldid=935763442.
- de Solla Price, D. J. (1965). Networks of scientific papers. *Science*, 149(3683):510–515. https://science.sciencemag.org/content/149/3683/510.
- Derwent, R. G., Jenkin, M. E., Saunders, S. M., and Pilling, M. J. (1998). Photochemical Ozone Creation Potentials For Organic Compounds In Northwest Europe Calculated With A Master Chemical Mechanism. Atmospheric environment, 32(14):2429–2441. http://www.sciencedirect.com/science/ article/pii/S1352231098000533.
- Dick Derwent, Andrea Fraser, John Abbott and Mike Jenkin (2010). Evaluating The Performance Of Air Quality Models. *online*. https://uk-air.defra.gov.uk/assets/documents/reports/cat05/1006241607 100608 MIP Final Version.pdf.
- Dillon, T. J., Tucceri, M. E., and Crowley, J. N. (2006). Laser induced fluorescence studies of iodine oxide chemistry part ii. the reactions of io with ch3o2, cf3o2 and o3. *Phys. Chem. Chem. Phys.*, 8:5185–5198. http://dx.doi.org/10.1039/B611116E.

Ebel, H., Mielsch, L.-I., and Bornholdt, S. (2002). Scale-free topology of e-mail networks. *Phys. Rev.* E, 66:035103. https://link.aps.org/doi/10.1103/PhysRevE.66.035103.

Edsu and Ellis, D. (2019). Etudier. https://github.com/wolfiex/etudier.

Ellis, D. (2019).Using Tf-Idf To Form Descriptive Chapter Summaries Via Keyword Extraction. https://towardsdatascience.com/ using-tf-idf-to-form-descriptive-chapter-summaries-via-keyword-extraction-4e6fd857d190.

Ellis, D. (2020). DSMACC-Testing. https://github.com/wolfiex/DSMACC-testing.

- Elshorbany, Y. F., Kleffmann, J., Hofzumahaus, A., Kurtenbach, R., Wiesen, P., Brauers, T., Bohn, B., Dorn, H.-P., Fuchs, H., Holland, F., Rohrer, F., Tillmann, R., Wegener, R., Wahner, A., Kanaya, Y., Yoshino, A., Nishida, S., Kajii, Y., Martinez, M., Kubistin, D., Harder, H., Lelieveld, J., Elste, T., Plass-Dülmer, C., Stange, G., Berresheim, H., and Schurath, U. (2012). Ho X Budgets During Hoxcomp: A Case Study Of Ho X Chemistry Under No X -Limited Conditions. *Journal of geophysical research*, 117(D3). https://www.academia.edu/29719780/HO_x_budgets_during_HOxComp_A_case_study_of_HO_x_chemistry_under_NO_x_-limited_conditions.
- Emmerson, K. M. and Evans, M. J. (2009). Comparison of tropospheric gas-phase chemistry schemes for use within global models. *Atmospheric Chemistry and Physics*, 9(5):1831–1845. https://www. atmos-chem-phys.net/9/1831/2009/.
- Fantin, V., Buttol, P., Pergreffi, R., and Masoni, P. (2012). Life Cycle Assessment Of Italian High Quality Milk Production. A Comparison With An Epd Study. *Journal of cleaner production*, 28:150–159. http://www.sciencedirect.com/science/article/pii/S095965261100388X.
- Ferguson, N. (2018). The Square And The Tower: Networks And Power, From The Freemasons To Facebook. Penguin Group, The.
- Fleming, Z. L., Lee, J. D., Liu, D., Acton, J., Huang, Z., Wang, X., Hewitt, N., Crilley, L., Kramer, L., Slater, E., Whalley, L., Ye, C., and Ingham, T. (2017). Dataset Collection Record: Aphh: Atmospheric Measurements And Model Results For The Atmospheric Pollution & Human Health In A Chinese Megacity. https://catalogue.ceda.ac.uk/uuid/648246d2bdc7460b8159a8f9daee7844.
- Freeman, L. (1977). A set of measures of centrality based on betweenness. online, 40:35–41.
- Freeman, L., Borgatti, S., and White, D. (1991). Centrality in valued graphs: A measure of betweenness based on network flow. *Social Networks*, 13:141–154.
- Freeman, L. C. (1978). Centrality in social networks conceptual clarification. *Social networks*, 1(3):215–239.

- Fujita, M., Inoue, H., and Terano, T. (2017). Searching promising researchers through network centrality measures of co-author networks of technical papers. In 2017 IEEE 41st Annual Computer Software and Applications Conference (COMPSAC), volume 2, pages 615–618.
- Gemma, J. (2019). The Most Influential Men And Women On Twitter 2017. https://www.brandwatch. com/blog/react-influential-men-and-women-2017/.
- Géron, A. (2017). Hands-On Machine Learning With Scikit-Learn And Tensorflow: Concepts, Tools, And Techniques To Build Intelligent Systems. O'Reilly Media. https://books.google.co.uk/books? id=khpYDgAAQBAJ.
- Goh, K. I., Kahng, B., and Kim, D. (2001). Universal Behavior Of Load Distribution In Scale-Free Networks. *Physical review letters*, 87(27 Pt 1):278701. http://dx.doi.org/10.1103/PhysRevLett.87. 278701.
- Google (2019). Google Scholar. https://scholar.google.com/schhp?hl=en.
- Hagberg, A. A., Schult, D. A., and Swart, P. J. (2008). Exploring network structure, dynamics, and function using networkx. In Varoquaux, G., Vaught, T., and Millman, J., editors, *Proceedings of* the 7th Python in Science Conference, pages 11 – 15, Pasadena, CA USA.
- Henze, D. K., Hakami, A., and Seinfeld, J. H. (2007). Development of the adjoint of geos-chem. Atmospheric Chemistry and Physics, 7(9):2413–2433. https://www.atmos-chem-phys.net/7/2413/ 2007/.
- Hewitt, N. and Edwards, P. (2009). Dataset Record: Op3-1 Campaign: Leeds Merged Chemistry Data At Bukit Atur. https://catalogue.ceda.ac.uk/uuid/81892deb2dd5e7f0d26b9c587af45f3d.
- Hobson, E. A., Mønster, D., and DeDeo, S. (2018). Strategic Heuristics Underlie Animal Dominance Hierarchies And Provide Evidence Of Group-Level Social Knowledge. *online*. http://arxiv.org/abs/ 1810.07215.
- Humphries, M. D. and Gurney, K. (2008). Network 'Small-World-Ness': A Quantitative Method For Determining Canonical Network Equivalence. *PloS one*, 3(4):e0002051. http://dx.doi.org/10.1371/ journal.pone.0002051.
- Jacob, P.-M. and Lapkin, A. (2018). Statistics of the network of organic chemistry. *React. Chem. Eng.*, 3:102–118. http://dx.doi.org/10.1039/C7RE00129K.
- Jeanningros, Y., Vlaeminck, S. E., Kaldate, A., Verstraete, W., and Graveleau, L. (2010). Fast Start-Up Of A Pilot-Scale Deammonification Sequencing Batch Reactor From An Activated Sludge Inoculum. Water science and technology: a journal of the International Association on Water Pollution Research, 61(6):1393-1400. http://dx.doi.org/10.2166/wst.2010.019.

- Jenkin, M. E. and Hayman, G. D. (1999). Photochemical Ozone Creation Potentials For Oxygenated Volatile Organic Compounds: Sensitivity To Variations In Kinetic And Mechanistic Parameters. Atmospheric environment, 33(8):1275–1293. http://www.sciencedirect.com/science/article/ pii/S1352231098002611.
- Jenkin, M. E., Saunders, S. M., Wagner, V., and Pilling, M. J. (2003). Protocol for the development of the master chemical mechanism, mcm v3 (part b): Tropospheric degradation of aromatic volatile organic compounds. Atmospheric Chemistry and Physics, 3(1):181–193. https: //www.atmos-chem-phys.net/3/181/2003/.
- Jenkin, M. E., Young, J. C., and Rickard, A. R. (2015). The mcm v3.3.1 degradation scheme for isoprene. Atmospheric Chemistry and Physics, 15(20):11433–11459. https://www.atmos-chem-phys. net/15/11433/2015/.
- John Hay, Ben Lynch, D. S. (1960). Mark 1 Perceptron Operators' Manual. Cornell Aeronautical Laboratory. https://apps.dtic.mil/dtic/tr/fulltext/u2/236965.pdf.
- Jones, E., Oliphant, T., Peterson, P., et al. (2001–). Scipy: Open source scientific tools for Python. http://www.scipy.org/.
- Kleinberg, J. M. (1999). Authoritative Sources In A Hyperlinked Environment. Journal of the ACM, 46(5):604–632. http://doi.acm.org/10.1145/324133.324140.
- Korke, R., Gatti, M. d. L., Lau, A. L. Y., Lim, J. W. E., Seow, T. K., Chung, M. C. M., and Hu, W.-S. (2004). Large Scale Gene Expression Profiling Of Metabolic Shift Of Mammalian Cells In Culture. *Journal of biotechnology*, 107(1):1–17. http://dx.doi.org/10.1016/j.jbiotec.2003.09.007.
- Krebs, V. E. (2002). Mapping Networks Of Terrorist Cells. *Connections*, 24(3):43–52. http://ecsocman.hse.ru/data/517/132/1231/mappingterroristnetworks.pdf.
- Kumar, R. and Upfal, E. (2000). The Web As A Graph. *online*. http://cs.brown.edu/research/webagent/pods-2000.pdf.
- Langville, A. and Meyer, C. (2005). A Survey Of Eigenvector Methods For Web Information Retrieval. SIAM Review, 47(1):135–161. https://doi.org/10.1137/S0036144503424786.
- LAPACK (2019). Lapack Linear Algebra Package. http://www.netlib.org/lapack/. http://www.netlib.org/lapack/.
- Lewis, A. C. (2018). The Changing Face Of Urban Air Pollution. *Science*, 359(6377):744–745. http://dx.doi.org/10.1126/science.aar4925.

- Ling, Z. H., Guo, H., Lam, S. H. M., Saunders, S. M., and Wang, T. (2014). Atmospheric photochemical reactivity and ozone production at two sites in hong kong: Application of a master chemical mechanism-photochemical box model. *Journal of Geophysical Research: Atmospheres*, 119(17):10567–10582. https://agupubs.onlinelibrary.wiley.com/doi/abs/10.1002/2014JD021794.
- McCulloch, W. S. and Pitts, W. (1943). A logical calculus of the ideas immanent in nervous activity. The bulletin of mathematical biophysics, 5(4):115–133. https://doi.org/10.1007/BF02478259.
- Mohanty, J. G., Nagababu, E., and Rifkind, J. M. (2014). Red Blood Cell Oxidative Stress Impairs Oxygen Delivery And Induces Red Blood Cell Aging. Frontiers in physiology, 5:84. http://dx.doi.org/10.3389/fphys.2014.00084.
- Molontay, R. and Nagy, M. (2020). Twenty Years Of Network Science: A Bibliographic And Co-Authorship Network Analysis. arXiv. http://arxiv.org/abs/2001.09006.
- Monastersky, R. and Van Noorden, R. (2019). 150 Years Of Nature: A Data Graphic Charts Our Evolution. Nature, 575(7781):22–23. http://dx.doi.org/10.1038/d41586-019-03305-w.
- Mquantin (2020). Idf response functions. *wikipedia commons*. https://upload.wikimedia.org/ wikipedia/commons/0/05/Plot IDF functions.png.
- Needham, M. and Hodler, A. E. (2019). Practical Examples In Apache Spark & Neo4J. O'Reilly. https://neo4j.com/neoassets/graphbooks/Graph_Algorithms_Neo4j.pdf.
- Newman, M. E. J. (2004). Coauthorship networks and patterns of scientific collaboration. Proceedings of the National Academy of Sciences, 101(suppl 1):5200–5205. https://www.pnas.org/content/101/ suppl_1/5200.
- Oliphant, T. (2006). Guide to numpy. https://docs.scipy.org/doc/ static/numpybook.pdf.
- Ottens, A. K., Kobeissy, F. H., Golden, E. C., Zhang, Z., Haskins, W. E., Chen, S.-S., Hayes, R. L., Wang, K. K. W., and Denslow, N. D. (2006). Neuroproteomics In Neurotrauma. *Mass spectrometry reviews*, 25(3):380–408. http://dx.doi.org/10.1002/mas.20073.
- Page, L., Brin, S., Motwani, R., and Winograd, T. (1999). The pagerank citation ranking: Bringing order to the web. *Stanford InfoLab*, 1(1999-66). Previous number = SIDL-WP-1999-0120 http: //ilpubs.stanford.edu:8090/422/.
- Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., and Duchesnay, E. (2011). Scikit-Learn: Machine Learning In Python . Journal of Machine Learning Research, 12:2825–2830.

- Pokroy, B., Epstein, A. K., Persson-Gulda, M. C. M., and Aizenberg, J. (2009). Fabrication Of Bioinspired Actuated Nanostructures With Arbitrary Geometry And Stiffness. Advanced materials, 21(4):463–469. http://doi.wiley.com/10.1002/adma.200801432.
- Poliaktiv (2011). Social Network Analysis: Theory And Applications. *online*. https://www.politaktiv. org/documents/10157/29141/SocNet TheoryApp.pdf.
- Press, W. H., Teukolsky, S. A., Vetterling, W. T., and Flannery, B. P. (1992). Numerical Recipes In C (2Nd Ed.): The Art Of Scientific Computing. Cambridge University Press, USA.
- R. Seeley, J. (1949). The net of reciprocal influence: A problem in treating sociometric data. Canadian Journal of Psychology/Revue canadienne de psychologie, 3:234–240.
- Read, K. A. (2010). Dataset Record: Cape Verde Atmospheric Observatory: Meteorological Davis Weather Station Measurements. https://catalogue.ceda.ac.uk/uuid/ a457d9715f3c4bc295ef975932e491d9.
- Rickard, A. (2020). MCM Website. http://mcm.york.ac.uk/.
- Rosenblatt, F. (1958). The perceptron: A probabilistic model for information storage and organization in the brain. *Psychological Review*, pages 65–386.
- Sabidussi, G. (1966). The centrality index of a graph. Psychometrika, 31(4):581–603. https://doi. org/10.1007/BF02289527.
- Sandu, A. and Sander, R. (2006). Technical note: Simulating chemical systems in fortran90 and matlab with the kinetic preprocessor kpp-2.1. Atmospheric Chemistry and Physics, 6(1):187–195. https://www.atmos-chem-phys.net/6/187/2006/.
- Saunders, S. M., Jenkin, M. E., Derwent, R. G., and Pilling, M. J. (2003). Protocol For The Development Of The Master Chemical Mechanism, MCM V3 (Part A): Tropospheric Degradation Of Non-Aromatic Volatile Organic Compounds. *Atmospheric Chemistry and Physics*, 3(1):161–180. https://hal.archives-ouvertes.fr/hal-00295229.
- Small, H. (1973). Co-citation in the scientific literature: A new measure of the relationship between two documents. Journal of the American Society for Information Science, 24(4):265–269. https: //asistdl.onlinelibrary.wiley.com/doi/abs/10.1002/asi.4630240406.
- Spencer, R. G. M., Hernes, P. J., Ruf, R., Baker, A., Dyda, R. Y., Stubbins, A., and Six, J. (2010). Temporal Controls On Dissolved Organic Matter And Lignin Biogeochemistry In A Pristine Tropical River, Democratic Republic Of Congo. *Journal of geophysical research*, 115(G3):2069. http://doi.wiley.com/10.1029/2009JG001180.

- Stubbins, A., Hubbard, V., Uher, G., Law, C. S., Upstill-Goddard, R. C., Aiken, G. R., and Mopper, K. (2008). Relating Carbon Monoxide Photoproduction To Dissolved Organic Matter Functionality. *Environmental science & technology*, 42(9):3271–3276. http://dx.doi.org/10.1021/es703014q.
- Turányi, T. and Tomlin, A. S. (2014). Reduction Of Reaction Mechanisms. Springer Berlin Heidelberg, Berlin, Heidelberg. https://doi.org/10.1007/978-3-662-44562-4 7.
- Vigna, S. (2016). Spectral Ranking. Network Science, 4(4):433–445. https://www.cambridge.org/core/ journals/network-science/article/spectral-ranking/99ACDCD0CC1B774AB0041FB16AB43D1B.
- Watts, D. J. and Strogatz, S. H. (1998). Collective Dynamics Of 'Small-World' Networks. Nature, 393(6684):440–442. http://dx.doi.org/10.1038/30918.
- Wood, B. (2014). The Origin Of Humans Is Surprisingly Complicated. *Scientific American*. https://www.scientificamerican.com/article/the-origin-of-humans-is-surprisingly-complicated/.

Chapter 5

Using Graph Clustering And Natural Language Processing To Aid Mechanism Reduction.

"Entities should not be multiplied beyond necessity."

- William of Ockham, Summa Logicae

5.1 Introduction

In the previous chapters, we have discussed visualisation and its role in bridging the gap between data and understanding. We have applied centrality metrics to a chemical network to tell us what species are of importance and experimented in getting machine learning models to learn the chemical structure of the species in a mechanism. This final research chapter provides a (brief) overview of current mechanism reduction techniques while providing two novel alternatives to aid the process.

Science often deals with the problem of understanding complexity. Such a task may be accomplished through organisation and partitioning (e.g. chunking a problem into smaller problems) and processing these at the same time using many workers (parallelism). In cases where such methods fail, we are forced to 'disregard' complexity. To do this physical processes may be simplified¹, or described using mathematics. Theorems and ideas may be applied to emulate 'real-world' outcomes based on the Platonian concept of an abstract 'Ideal' world [Welton et al., 2002; Ostrovsky, 2005].

The process of lumping has long been used to replace a complex, changing process (e.g. Quantum Mechanics or Boundary Layer Fluid Dynamics) with a more straightforward constant process, [Maha-jan, 2008]. In such cases, an approximation may be far more useful than a lengthy exact solution, or none at all provided the primary criteria/outcome is identified and optimised for (evaluated against a benchmark or standard).

Similar problems of complexity are seen within the chemistry of the atmosphere. An example is seen within the Master Chemical Mechanism (MCM v3.3.1), [Rickard, 2020], this contains 1228 RO₂ reactions. If written explicitly, all RO₂-RO₂ (gross and self) interactions would result in a total of 1,507,984 reactions. Instead, the MCM overcomes this problem by creating a RO₂ pool, with which all RO₂ species react. This results in a mechanism which preserves the quality of science (the primary goal of the MCM is to preserve O₃ prediction) with only 0.000814 of the total possible RO₂ - RO₂ reactions.

However, even with such simplifications, atmospheric chemical mechanisms have been increasing in size over the last ten years ([Dick Derwent, Andrea Fraser, John Abbott and Mike Jenkin , 2010],Figure 2.2). With the ability to automate their construction, mechanisms with species numbers of the millions become possible. Although the existence of more-explicit mechanisms may improve the quality of science produced, they can cause problems for efficient computation, diagnosis and analysis. This chapter shall look at two methods in which we may simplify a mechanism by grouping species with similar reaction patterns together. These are through the use of species lifetime (Section 5.7)

¹It is common to approximate a year as 365 days, an atom as a sphere and replace the Van der Walls equation with the ideal gas law (for normal pressures).
and graph-based clustering (Section 5.4).

5.2 Mechanism Reduction

Although this chapter discusses work which can contribute to a chemists toolkit for mechanism reduction, it does not concern itself with this task specifically - and will not contain any work directly related to this nor analyse the results of a mechanism where the species suggested may be lumped together. Although the framework for such a task exists, this shall be left as a task for future work.

Currently, there exist two main reasons for trying to find species of similar chemical properties. These are searching for pysciochemical analogues within the field of cheminformatics (see Chapter 6) and mechanism reduction. This chapter concerns itself with using the graph structure to group species with reactions on fast timescales, or similar connectivity patterns - a problem commonly presented in the production of a reduced (simplified) chemical mechanism. For this reason, the following section provides a short literature review on several different reduction techniques.

5.2.1 Species Categories

As with any problem, the first step to simplifying a complex task involves the partitioning data into categories. For a mechanism, we begin by observing the foci of an experiment and defining some critical (necessary) species for the task. Following these needed species (species which are required by the essential species for the chemistry to work) are identified and added. Finally, species which provide a negligible input to the aims of an experiment are labelled as redundant and often removed. The outline of each category is given below.

- **Important** reactions or species directly related to the topic/outcome we are interested in (for the MCM this is ozone)
- **Needed** reactions/species required by the important species/pathways, such that they may perform their desired function
- **Redundant** those we may remove with little or no consequence to the outcome of the model. These are determined through the use of sensitivity analysis.

5.2.2 Reaction Removal

Since atmospheric chemical mechanism forms a numerically stiff system (Chapter 1), a reduction in the number of reactions within a mechanism leads to a reduction in the computational burden experienced

by a model each iteration forwards in time. Classically the identification of important reactions may be accomplished through the use or rate of production and loss analysis (Subsubsection 4.5.0.2). This allows us to filter reactions contributing less than 5% (in the MCM reaction pathways < 5% are disregarded) to the formation of any species we are interested in. Other methods using principal component analysis of the sensitivity of species (PCAS) also exist and are discussed in [Vajda et al., 1985; Wyche et al., 2015].

5.2.3 Species Removal

Similar to reaction removal, the removal of species is useful because the removing or combining of species inherently reduces or simplifies the reactions within a mechanism. This method also has added benefit of reducing the size of the Jacobian matrix used to propagate the chemical system forwards. For large systems which do not use a sparse framework, storing a n^2 matrix in memory can prove difficult.

Many methods of species reduction are possible. The simplest of these is through the use of trial and error [Turanyi, 1990] (Method 1). Here the consuming reactions for a species are removed, and if the resulting deviation in results between the full and reduced mechanism is small within a certain threshold, their results are retained. The main downside to this is that it only works on a per-species level, which may be very resource-consuming for large mechanisms.

With the use of sensitivity analysis, it is possible to remove species whose reaction are much slower than the rate-determining steps of a mechanism, [Oran and Boris, 1991]. However, even after removing all slow-reacting species, those on a fast timescale remain. Here the use of Quasi-Steady-State Approximation (QSSA), [Whitehouse et al., 2004a], can be used to identify species associated with fast timescale reactions. QSSA works on the assumption that such species have little to no change in concentration over time - i.e. the net flux (\dot{v}_i) is zero. Such an assumption causes an error Δc_i of :

$$\Delta c_i = \frac{\dot{v_i}}{J_{ii}} \tag{5.1}$$

where J_{ii} is the diagonal of the Jacobian matrix. Here if the error for a species is small, the species may be removed from the mechanism.

Finally, investigation of the system Jacobian can be used to identify redundant species, which is a 'capable' and 'efficient' method for removing most redundant reactions and species from the MCM, [Whitehouse et al., 2004a]. Use of a log-normalised Jacobian to determine which species can be removed is found in the connectivity method [Turányi and Tomlin, 2014; Turányi, 1990]. Here the influence a 1% change in a species concentration has on the concentration of 'important' species can

be determined by

$$B_i = \sum_j ((y_i/f_i)(\partial f_i/\partial y_i))^2$$
(5.2)

where $(y_i/f_i)(\partial f_i/\partial y_i)$ is element *i* of the normalised Jacobian (see Chapter 4 for information about the construction of the Jacobian matrix). Through an iterative process species with a low contribution to our important species can be found and removed.

5.2.4 Lumping

Rather than removing species or reactions from a mechanism, we may combine them to form a new composite species. This is referred to as species lumping. To do this, we must first consider how we define species that are to be joined together, and then how their grouped reactions will contribute to every other species, it reacts with. Some of the more general types of lumping styles are outlined below.

5.2.4.1 Chemical Lumping

Mechanisms follow protocols in their generation. This produces reaction styles that many likestructured species follow in their degradation. In determining such classes, we may be able to generalise like-species reactions and group them as one. An example of this is the Common Representative Intermediates (CRI) Mechanism (described in Subsection 5.3.1). Here the 'ozone production potential' of the species within the MCM is used to simplify and reduce it. This is a function of the C-C and C-Hbond ratio of a species (its CRI index). Species with the same CRI index are lumped (grouped) into a proxy species. Alternatively, time scale analysis for species lumping has been successfully applied by [Whitehouse et al., 2004b]. Here it is seen that many groups of species have rate coefficients that are identical or sufficiently similar due to the generic rules/protocols of the MCM. This results in a similar overall lifetime for species in the same group, allowing them to be lumped together with little overall consequence to the experimentation criterion.

5.3 Data Setup

Unlike manual reduction, this chapter does not concern itself with the intricacies of the chemistry behind a mechanism. Instead, we search for an automated method of simplifying the mathematical structure behind a mechanism while preserving the quality of science it represents. Although this may not directly replicate real-world scenarios, it can provide an accurate test of the robustness of a mechanism and the equations within it. Work is carried out on the assumption that the equations within the MCM benchmark mechanisms are representative of experimental results, and in simplifying these, their usefulness in modelling the real data is preserved. This section describes the experimental setup for the experiment.

5.3.1 The Mechanism

The mechanism used is the Common Representative Intermediates (CRI) Mechanism v2.2 [Jenkin, 2019]. This is an already reduced version of the MCM v 3.3.1, where species are grouped based on their ozone formation potential - i.e. the C–C and C–H ratio of bonds. Reductions have been made on a compound-by-compound basis and compared to the MCM using a series of 5-day box-model simulations, [Jenkin et al., 2008].

Why further simplify the CRI network?

CRI v2.2 [Jenkin et al., 2008] is a mechanism of 422 species and 1261 reactions - that is 7% of the full MCM (5809 species and 17224 reactions). Although this is significantly smaller than the full MCM, it may still prove problematic if used within a global model - for comparison, the GEOS-Chem² standard chemistry is approximately half the size of this, [Community, 2020].

5.3.2 The Box-Model

The box model is an adapted version of the Dynamically Simple Model of Atmospheric Chemical Complexity (DSMACC) [Emmerson and Evans, 2009; Ellis, 2020]. Recent updates allow for multiple parallel runs, easy extraction of rates, fluxes and the Jacobian matrix as well as a simple Neurses (a command like semi-graphic interface) interface for loading and parsing new files.

The DSMACC model works by using the Kinetic PreProcessor (KPP), [Sandu and Sander, 2006], to generate Fortran code, which can then be used to integrate the provided mechanism. As there were some issues presented a pre-pre parser code is used before running KPP. Occasionally a post parser may be required on some of the files to produce the desired output.

5.3.3 Model Inputs

The aim of this experiment is not to replicate a specific case study or scenario. Instead, we extract all non-lumped species which appear in both CRI and the MCM and provide an assortment of initial

 $^{^2{\}rm A}$ global 3D model of atmospheric chemistry driven by meteorology from NASA's Goddard Earth Observing System (GEOS), [GEOS-Chem, 2020].

condition concentrations to cover the entirety of the input space.

To select the initial conditions there exist several sampling styles [McKay et al., 2000]. The most common style is the random or 'Monte Carlo' approach. However, this does not guarantee a homogeneous distribution of points. A lattice or grid approach is also possible, but that can result in a large number of sample points to produce a complete distribution of the input space. To overcome this, a Latin hypercube can be used. This is a generalisation of the Latin square - a square matrix containing n items, arranged in such a way that they only appear once in each row and column (akin to a sudoku puzzle) [Dodge, 2008]. The experimental setup uses a Latin hypercube to define the initial condition range for 148 primary emitted species, and 300 simulations follow the formula below:

concentration
$$\begin{cases} min = 100ppbv \ max = 1pptv, & \text{if } NO, NO_2, O_3 \\ min = 10ppbv \ max = 0.1pptv, & \text{otherwise} \end{cases}$$
(5.3)

5.4 Graph Based Reduction

It has been shown that a graph-based representation of the atmospheric chemical network proves useful in both the visual and mathematical analysis of simulation results (Chapters ??-??. It, therefore, follows that the network representation of mechanism may also have its uses in the simplification, and thus reduction, of chemical complexity. This section will outline the basic methods of modularity (cluster) detection with the graph framework, the different methods in which this may be done and eventually apply it to a case example representative of the chemistry within the London environment.

5.4.1 Graph Parallels

Graph structure can be used to analyse changes of reactions or relationships between species - providing an alternative representation and method to access such data. Additionally, clustering techniques may be used to locate groups of highly connected, fast reacting/strongly related species. This has applications in both understanding the data, but more importantly, chemical lumping. In creating a graph from a model simulation, we encode not only information about the chemical structure, but also the influence between species in the mechanism. By grouping species which have a strong dependence upon each other, we can simplify the provided network or mechanism.

5.4.2 Types Of Graph Clustering

Unlike vector clustering algorithms (such as DBSCAN, UMAP and K-means - see Subsection 6.5.1), graph clustering metrics do not rely on the spatial orientation of the data to determine groups or 'clusters'. Instead, these may partition the network into segments, group nodes by structural equivalence or explore the 'flow' dynamics of the network.

Algorithms such as Label Propagation [Raghavan et al., 2007] and spin-glass [Newman and Girvan, 2004] work by randomly assigning nodes with a property or label. This property is then transferred to its neighbours. Other algorithms such as the nested block model can decompose a graph into clusters of similar properties, [Fortunato, 2010]. These are often grouped in the form of topological equivalence which can be either:

- structual equivalence vertices are similar if they have like neighbours, [Zhou, 2003].
- regular eqivalence retrieves nodes with similar connection patterns (e.g. parent child node hierarchicl structures), [Everett and Borgatti, 1994].

This works similarly to an AutoEncoder (discussed in Subsection 6.3.5), where topological similarities are used to simplify (or encode) the network structure, in a way which it may be decoded again.

Finally, there exist a set of 'flow' based models which use the network dynamics to determine the modularity of a network. These are discussed below.

5.4.3 Walk/Flow-Based Clustering

Temporal networks result in a change in the relationships between items (magnitude/type). Such changes in the network dynamics are encoded within the edges of a graph. The primary function of a random walk or 'flow' algorithms is to capture the changes between the real-world systems represented by the network.

In Subsubsection 6.5.1.1, the silhouette coefficient was used to compares the vector position of clusters with regards to the distance of data points between them. Translating this to the graph framework, topological (graph) clustering defines a cluster, or module, as a region with a higher inter-cluster degree or density³ compared to their intra-cluster density⁴. This results in a system, that is sorted by group, has more links between elements of the same group than with those in other groups - such patterns can be seen within the sorted adjacency matrix in (Chapter 2).

³The number of links or edges between items in the same group.

⁴The number of edges to other clusters

Since flow-based methods are more interested in the network dynamics, than structure, the number of links or density is replaced with the time a random 'walker' spends 'trapped' between a set of nodes. A real-world analogy would be to view the flow of water in a slowly filling river, Figure 5.1. Here a walker (or water molecule) traverses the entirety of the river/graph network, occasionally getting trapped between a set of nodes. Here although the water is still moving, it ends up spending more time going back and forth between a set of nodes, than exploring the rest of the network. It is these regions of stalled progress that form network clusters.



Figure 5.1: The proposed plans for the change of the UK National Watersports Centre Whitewater Course (Holme Pierrepont). Walk based clustering is analogous to the movement of a river. Clusters (or modules) are identified as areas where the 'flow' becomes trapped, much like water in the pools immediately following a hydraulic jump. Source: [Cornes, 2008]

5.4.4 Louvain Clustering

The Louvain clustering algorithm is one of the most popular of the clustering algorithms due to its algorithmic and qualitative robustness, [Blondel et al., 2008; Lu et al., 2015]. On the simplest level, this works by maximising the modularity for each configuration. Modularity is a value between positive and negative unity which measures the density of edge between inter and intra communities and compares it to an equivalent random network. The Louvain is a hierarchical clustering algorithm; this means that after each iteration, all nodes which belong to the same cluster are consolidated to form a new 'grouped' item. Inter-cluster links are converted into self-links, and intra-cluster links are updated accordingly.

5.4.5 Infomap Clustering

The Infomap clustering algorithm is a flow-based method which operates on system dynamics rather than structure [Rosvall et al., 2009]. It works by trying to minimise the Huffman code [Huffman, 1952] - a type of optimal prefix code used for lossless compression in computer science and information theory. Here the frequency of items is grouped to produce a binary decision tree (Figure 5.2 left). Here the letters a,d,r,g and b appear much more frequently than! and c. Navigation this binary tree can then be used to produce a Huffman code for each letter (Figure 5.2 right) where letters with a higher frequency have a shorter code length. This method is used within the infomap algorithm to compresses information about the probability of a random walker transitioning between pairs of nodes in a network. Here the prefix of a Huffman code works much like a postcode (City+Areacode, StreetCode) to identify regions where the walker gets trapped (clusters of high density with a low probability that the walker leaves).

As the number of partitions grows exponentially with the number of nodes⁵ it is not possible to apply a brute force approach to find the best number of partitions. Instead, a variation of the Louvain method applies 'submodule exploration' which verifies that large modules should not be broken down into smaller ones and 'single-node movements' which allow individual nodes to move between modules. These optimisations avoid the situation of too large modules (where each node has a category to itself) or too small where the number of prefix codes is great.



Figure 5.2: A Huffman tree, and the Huffman code generated from it. A Huffman tree is created using the frequency of occurrence of an item. The more often it appears in the source, the shorter the path to it. Source: [Sad CRUD Developer, 2016]

 $^{^{5}}$ See 'Bell Numbers' - these count the possibles partitions in a set and have root going back to medieval Japan.

5.5 Selection Criteria For Graph Clustering

The main two criteria in selecting an algorithm for grouping atmospheric reactions are:

- 1. The algorithm can deal with a directed network chemistry is directional.
- 2. The algorithm can handle temporal data The chemistry within a system changes depending on the time of day. This is mainly due to the change in the amount of radiation available from photolytic reactions of oxidants and photolysis reactions.

As the InfoMap algorithm implements a directed approach coupled with a multi-level clustering approach able to capture node-layer interaction in temporal networks, [Aslak et al., 2018], it makes the right candidate for the task of mechanism reduction.

5.6 Evaluation Of Infomap On A Real Simulation.

Using the initial conditions for London from (Table 4.4), a spun up a model simulation with the CRI v2.2 mechanism was run. Since this does not contain C_5H_{11} CHO, MVK, MACR or Limonene as primary species, these are omitted from the initialisation. Following a spinup to radical steady-state, a graph is generated for noon after one day of an unconstrained run. The InfoMap algorithm is then applied to the generated graph.

The coarsest level of clustering is shown in Figure 5.3. Here nodes are coloured by their cluster, and approximate polygon hulls (a shape of multiple corners which enclose a set of points) surround the nodes closest to the median cluster centre. Much like the findings in (Chapter 3), it is seen that different sections of the graph network represent different types of chemistry - for example, hull 4 contains aromatic species, hull 2 contains the products of linear alkanes and hull 3 contains the terpenes.



Figure 5.3: A graph of CRI v2.2 showing the hulls of the first level of hierarchical clustering. Nodes are coloured by the splits in branches, and the hulls enclose the nodes which lie within 95% of the (median) centre of the cluster.

As the InfoMap algorithm provides a finer level of clustering, it is essential to evaluate its performance. Using a graph-hull approach, as in Figure 5.3, becomes cluttered and unusable. Instead, a bubble plot may be used. Although this sacrifices the ability to view links, it allows for the complete overview of the hierarchical structure. In Figure 5.4 shows the nested structure of each clustered group. In an electronic mail correspondence with Mike Jenkin (Section D.1), the origin of the naming convention of reduced species was explained. Individual nodes are coloured by their prefixes. This allows the further categorisation of the species structure within each category.



Figure 5.4: **Species structure within each cluster.** A nested bubble chart is used to show the full hierarchical structure of the mechanism. This allows us to evaluate the species structure/type that has been extracted in each level of the hierarchical split. Node sizes are representative of the \log_{10} number of walkers that have become trapped by the flow algorithm at a location.

5.6.1 Species Type And Clustering

The bubble chart provides an intuitive way to represent groups for interactive or small systems but is less useful for larger numbers of species and print (Figure 5.4). Instead, a tree approach is better suited to revealing the hierarchical structure of the network, as shown in Figure 5.5. Here branches within Figure 5.5 are numerically labelled for each level. This allows us to navigate the hierarchy using a sequence of numbers (e.g. to get to C_4H_6 we take the branch 1 from the centre, followed by branch 5 - resulting in the notation 1.5.C4H6).

This split notation allows a general overview of the mechanism structure, as well as the reasoning/process of the clustering algorithm. The first level split in Figure 5.3 shows branches 1,2 and 5 to



Figure 5.5: A radial treemap showing the hierarchical clustering of the CRI mechanism. The simulation results used are representative of the chemistry within London at Noon local time and generated using DSMACC and the InfoMap algorithm.



have origins in the linear (n-) alkane species. This can be seen through both the emitted species (bold) and the RN prefix of the species. Here the linear alkanes can react with OH to extract hydrogen and then from a RO₂, or produce a carbonyl *CARBxx*, which can then go on to produce the $RNxxO_2$ peroxy radical.

Except for benzene in 2.14, branches 3 and 4 contain the aromatic species in the network. Branches $4.\{2,5,9,11\}$ all consist of $RAxxO_2$ species, which are the product of the addition of OH to toluene/benzene ringed species. $4.\{1,7,8\}$ and 1.5 contain peroxy radicals formed from the degradation of conjugated dienes $RUxxO_2$. For the CRI v2.2 mechanism these are only isoprene and 1,3-butadiene. Such peroxy radicals often go on to form unsaturated carbonyls, as denoted by UCARBxx.

Branch 3 contains the monoterpenes. This can be seen in 3.{2,5} (α -pinene) and 3.6 (β -pinene). Here peroxy radicals formed from the reaction with the endocyclinc⁶ and exdocyclinc⁷ double bonds of α - and β - pinene are denoted with the prefix *RTN* and *RTX*.

The $RIxxO_2$ prefix was used initially for the peroxy radicals iso ('i-') alkanes and their carbonyl products - branches 3.{1,4}, however, they tend to mainly be used for smaller branched precursors which produce acetone (CH₃COCH₃) as a significant product in their oxidation chain (branch 3.1). Acetone is a relatively unreactive carbonyl, the fact that it is water-soluble means that they may be washed out of the atmosphere by precipitation, [Andersson-Sköld et al., 1992]. This may have been seen to interrupt the ozone formation process under regional-scale photochemical smog conditions in north-western Europe.

Finally, since the CRI index is representative of the oxidation potential, it is common to see species containing the CRI value within a cluster. Cluster typically contain a combination of carbonyl (R(=O)R', CARBxx), hydroperoxy (R-OOH, RxxOOH), peroxy $(ROO \cdot, RxxO_2)$ and nitrate $(R-ONO_2, NO_3)$ groups. For the lumped species, it can be common for a RO₂ species to react with NO or NO₃ to produce a carbonyl with a CRI index of two values lower. This can be attributed to the loss of hydrogen in the oxidation process. Similarly, a reaction with NO or HO₂ can produce a hydroperoxy or nitrate species, which in turn react with OH to produce the equivalent carbonyl.

5.6.2 Number Of Clusters

Sometimes it may be required to have a preset (target) number of clusters. The InfoMap algorithm contains a *preffered number of modules* parameter which can either terminate the algorithm early, should the number be reached (or continue splitting if it has not). Since we are interested in merging

⁶Inside the pinene ring.

⁷Outside the pinene ring.

smaller numbers of nodes, this can be seen as a useful parameter to have. However, in selecting a number too large, (e.g. 200 clusters, which should result in groups of 2-3 nodes), it is seen that much of the hierarchical information from the network is lost, Figure 5.6. It is for this reason that forcing the number of nodes without reason will not be attempted.



Figure 5.6: A radial tree of the InfoMap algorithm with a forced number of groups. Here a loss of hierarchical structure can be seen when compared to Figure 5.5. By setting a high number of required clusters, many species are grouped by themselves, which does not provide a useful output for mechanism lumping.

5.7 Reduction Through Lifetime

In Subsubsection 5.2.4.1 is was mentioned that a species lifetime could be used to decide on which species may be lumped together. Here Whitehouse et al. [2004b] found that large groups of species within the MCM had similar or identical lifetimes and that in many cases this could be attributed to similar/identical rate coefficients for the same type of reaction. This was then used as a methodology for automatic mechanism reduction.

This section describes a method in which this may be performed without prior knowledge of the mechanism. Natural language processing tools are applied first to determine species of a similar lifetime across a range of timesteps (Subsubsection 5.7.1.1), and then their standardised temporal profiles are compared (Subsubsection 5.7.1.2). However, we first begin by defining the lifetime of a species.

5.7.0.1 Calculating The Lifetime

The chemical lifetime is of a species is defined by calculating the lifetime of a species against the loss of individual reactions:

$$\tau_A = \frac{1}{\sum_{i=1,n} 1/\tau_{A_i}}$$
(5.4)

where τ_A is the overall lifetime of species A and τ_{A_i} are the lifetimes of a die to a loss from reaction 1 to *n* [Jacobson, 2005]. It can be noted that this equation is calculated as part of the diagonal (J_{ii}) of the Jacobian matrix [Turanyi and Tomlin, 2015; Whitehouse et al., 2004b]:

$$\tau_A = -\frac{1}{J_{ii}} \tag{5.5}$$

Where *i* corresponds to the index of species A in the matrix, since this is a loss, the value of J_{ii} will be negative unless a species does not contain a consuming reaction (then it will be zero).

5.7.1 Comparing Magnitude And Direction

The most significant changes of a reaction rate within a simulation are due to photolytic reactions. Here the solar zenith angle determines the amount of radiation which can reach (and excite) a molecule. Since the chemical lifetime of a species takes into account its loss due to other reaction, such temporal changes in reaction rates need to be taken into account. The simplest method is to construct a vector for each species, showing how their lifetimes change over time. In doing so, we can apply natural language processing techniques, such as euclidean (magnitude) and cosine (angle) distance, co compare them.

5.7.1.1 Euclidian Distance

This is the simplest method of vector comparison and works by calculating the distance between all points in two vectors. For the vectors

$$v1 = [a, b, c, \dots n]$$

 $v2 = [i, j, k, \dots z]$ (5.6)

This can be done using Pythagoras' theorem in Equation 5.7:

$$e_{dist} = \sqrt{(a-i)^2 + (b-j)^2 + (c-k)^2 + \dots + (n-z)^2}$$
(5.7)

This transformation converts the straight line distance between each vector into metric space, allowing us to represent the difference in their magnitudes as a single scalar. Unfortunately, as this requires the difference between all permutations of rows, it cannot be done as a single operation.

5.7.1.2 Cosine Distance

Similar to euclidean distance, if we wish to calculate the angle between two vectors, we may use the cosine difference. In starting with the definition of the dot product

$$v1 \cdot v2 = \|v1\| \|v2\| \cos \theta$$

this may be arranged

$$\cos\theta = \frac{v1 \cdot v2}{\|v1\| \|v2\|}$$
(5.8)

The problem is that for a meaningful representation for the cosine inequality, the Cauchy-Schwarz (triangle) inequality needs to be satisfied. This states that for all sequences of real numbers a_i and b_i :

$$(\Sigma a_i^2)(\Sigma b_i^2) \ge (\Sigma a_i b_i)^2 \tag{5.9}$$

Each vector needs to be normalised before the calculation of the angle. This eliminates information about the magnitude of the vectors but also allows for a better comparison of the distribution (or shape). This normalisation factor makes it particularly useful in the analysis of text documents, where a word may appear multiple times in different length segments.

5.7.2 Temporal Lifetime Vector Comparison

To compare a species diurnal profile with its absolute lifetime, we can plot the cosine and euclidian distance against each other on a x - y scatterplot, Figure 5.7b. In this subsection, we compute the Euclidean and cosine distances for all remaining reaction pairs (88410 pairs) for a single simulation from the setup described in Subsection 5.3.3. We start by looking at the species density profiles, Figure 5.8.



(a) Original



Figure 5.7: Cosine distance against Euclidian Distance. Normalised version of the two distance metrics are plotted in an x - y scatterplot. Each dot represents a different species pair plotted at the location of their value for each metrics. Species pairs with similar values and profiles have small values for each metric and are located towards the upper left hand corner. (b) shows the same results as (a) but has a no-overlapping (collision detection) algorithm applied show all points, and therefore aid interactive selection.

NOTE: the kernel density plot x axis shows 1-(the value shown in the scatter plot). This is because output values from each distance closer to 1 are similar. In the scatter plot inverting this however proved more straightforward to plot and explain

Similar to Whitehouse et al. [2004b], we find there are several groups of species with similar lifetimes. In general, we have two main peaks where the temporal profile and concentration differences are similar. Here the first peak (Figure 5.8 from the right) shows a significant agreement between both similarities. This suggests that most of the species within this section react similarly, and will very likely have the same inorganic reactions at a similar rate. The second peak, however, shows species which have a similar diurnal response, with different magnitude differences. These species are likely affected by photolysis reactions directly or indirectly but have a differing set of reactions controlling them. In a concentration line-plot we would expect them to have peaks in the same location, but to change at a different rate/magnitude to each other.



Figure 5.8: Gaussian Kernel Density Estimate plot showing the distributions present for the $\{0,1\}$ scaled euclidean and cosine distances. This graph shows the density profiles for each metric in Figure 5.7a - these show 1-(normalised metric distance), and therefore the peak at 1 corresponds to that in the top left of Figure 5.7a. Peaks here correspond to the regions of high densitry within the scatterplot Figure 5.7b.

A comparison of both similarities on the x - y plot is shown in Figure 5.7a. As many species have similar lifetimes, these are often situated within the same temporal space, which can make it hard to visually or interactively separate them. It is possible to convert the scatter plot into a force-simulation, Figure 5.7b. Here nodes repulse each other and are attracted to their original location. This expands the graph and prevents overlapping nodes. In doing so, it is possible to interactively query the pairs of nodes which are represented by each point if required. This information can be better shown in a Kernel density plot comparing the distribution of cosine and euclidean distances.

The agreement of both metrics suggests a similarity between the lifetime values and their change in time for simulation. This is in agreement of with the x - y plot of the species. In selecting species that are part of the same initial cluster and have a high agreement between both similarities, it is possible to gauge the suitability for two species to be lumped together.

5.7.3 A Quick Comparison

Having described how the similarity distances work, Figure 5.7b showed the locations of the best and worst matched pairs. This subsection looks a the differences between these using a log10 ensemble of the mixing ratios for the 300 simulations used in the results section. Figure 5.9(a,b) show that the best matching pairs contain an easy to match flat decay curve, with the worst Figure 5.9(c,d) often



containing a combination of a species which decays with one which undergoes a photolytic reaction.

Figure 5.9: Comparing the best (a-b) and worst (c-d) species combinations using the combined similarity metrics. Here species which only undergo a simple decay seem to be the easiest to group together. Species pairs between an photolytic and non photolytic species produce different profiles at differing magnitudes and are therefore difficult to match.

5.8 Results

To get a representation of the mechanism, we run 300 randomly initiated scenarios (Subsection 5.3.3). The experimental setup is one such that it is possible to add more data points at a later date. From each simulation, the no diagonal elements of the Jacobian are used to construct a graph representative of the aggregated hourly means of the simulation output. Each of these graphs is then run through the infomap algorithm, and a grouping/clustering produced. Each infomap is run 100 times, where the result with the best fit (shortest code length) is taken - this is an optional parameter on the algorithm.

5.8.1 The Co-Grouping Network

To aggregate the groupings produced by each algorithm an $n \times n$ matrix is created for each of the (n) species in the mechanism. This is treated as a relational graph matrix. If species A is in the same group as species B, then a link (or value +1) is added to the [A,B] (A \longrightarrow B) and [B,A] (B \longrightarrow A) column. Using this matrix format, it is possible to generate a graph showing the relationship between species that were clustered in the same group.

This relational matrix can then easily be converted into the network format: Figure 5.10a. Starting with this it is then possible to filter edges below a certain weight, Figure 5.10b-d. Finally, isolates (nodes with no links) are removed, leaving only those clusters where each species has a strong relationship between every other.

In the context of this section, we select only relationships that appear in over 45% of all the clustered simulation runs. The reasoning is that there may exist a pairing which only appears during day or night time.



Figure 5.10: Filtering the infomap clustering relationship matrix/graph How the clustering relationship network changes as weak links (links between species which do not appear in many of the infomap groupings) are removed.

5.8.2 Comparing Daytime And Nighttime Groups

In determining a group of species which are commonly clustered together in most simulation results, we are next interested in seeing if these groups change with day or night.

To do this, we use an alluvial diagram [Rosvall and Bergstrom, 2010]. This is a cross between a parallel-line plot and a Sankey diagram and is particularly suitable for showing the changes in clusters within a temporal network.

In taking the common clusters formed at midnight (Figure 5.11 left) and midday (Figure 5.11 right) we are able to compare these to the overall selection (all hours - middle). Here, as is expected, any parings which persist in over 45% of all the timesteps, exist in all three categories. We see a selection of species which are grouped at 12:00 and 0:00 hours. This suggests that they may not be grouped with some of the intermediate hours and that if the threshold of selection is lowered below 45, they may appear in the overall result—finally a selection of species which are only grouped in daytime or night time only results.

5.8.3 Determining Cluster Suitability

Having selected clusters that appear for most graphs in the network, it is now important to assess the suitability of each node for being lumped together. Using a normalised similarity matrix, we extract the values for each of the lumped groups, Table 5.1. Here the best values are provided by the PECOH and DIEK species, Figure 5.12a. These both have linear decaying concentrations within the same order of magnitude. This is probably due to PECOH being the only precursor to DIEK, where DIEK accounts for 0.436% of its total products. This makes them a suitable candidate for lumping. $HOCH_2CO_3H$ and $HOCH_2CO_3$ make the worst possible lumping combinations. This is because the radical HOCH₂CO₃ can react with many of the inorganics, while HOCH₂CO₃H can only dissociate into formaldehyde or react with OH to reproduce HOCH2CO3. Although these species both have differing profiles, of several orders of magnitude difference, their cyclic nature $HOCH_2CO_3H \xrightarrow[HO_2]{HO_2} HOCH_2CO_3$ most likely proved to trap the 'flow' of the network, producing the cluster. Additionally there are also several clusters consisting of (N)RIxxOOH and (N)RIxxO₂ species. These are generally species formed from iso-alkanes (Section D.1), and both produce acetaldehyde (CH₃CHO) as a product. Here the peroxy radical $(R-O_2)$ produces a diurnal profile. Regardless of this, the cosine similarity is still relatively small. This may be attributed to the 'flat' periods of slow decay that is experienced at nighttime (due to the reduction of available HO₂ and NO) which follow the loss trend of the peroxide (R-OOH) species. Since OH addition and H-abstraction are both fast reactions these often form species with similar CRI numbers, the clustering algorithm often identifies peroxides and their peroxy

NR122000F NR122002				NRT 22000H NRT 22002
RTN23O2 RTN23OOH				RTN2302 RTN2300H
RTN25NO3 TNCARB15				RTN25NO3 TNCARB15
RTN2500H RTN2502				RTN2500H RTN2502
NRI1500H NRI1502		N	IRI15OOH NRI15O2	NRI15OOH NRI15O2
NRI1200H NRI1202			IRI1200H NRI1202	NRI12OOH NRI12O2
RTN24O2 RTN24OOH				RTN24O2 RTN24OOH
RN802 CH3COCH3				RN802 CH3COCH3
RU1102 RU1100H				RU1102 RU1100H
RTN2600H RTN2602 RTN26PA	N		R	IN2600H RTN2602 RTN26PAN
RUA2000H RUA2002				RUA2000H RUA2002
PHAN HOCH2CO3H HOCH2CO3	3	PI	HAN HOCH2CO3H HOCH2CO3 F	HAN HOCH2CO3H HOCH2CO3
				IEPOX RU14OOH
IEPOX UCARB12 RU1400H	<00:00) I	12:00>	IEPOX RU1400H
IEPOX UCARB12 RU14OOH	<00:00	9	12:00>	IEPOX RU1400H
IEPOX UCARB12 RU14OOH PPN C2H5CO3H C2H5CO3 CH3CO3H CH3CO3 PAN	<00:00		12:00>	IEPOX RU1400H
IEPOX UCARB12 RU14OOH PPN C2H5CO3H C2H5CO3 CH3CO3H CH3CO3 PAN BU12OOH BU12O2 DHCABB9	<00:00		12:00>	IEPOX RU1400H
IEPOX UCARB12 RU14OOH PPN C2H5CO3H C2H5CO3 CH3CO3H CH3CO3 PAN RU12OOH RU12O2 DHCARB9	<00:00		12:00>	IEPOX RU1400H
IEPOX UCARB12 RU14OOH PPN C2H5CO3H C2H5CO3 CH3CO3H CH3CO3 PAN RU12OOH RU12O2 DHCARB9 RU14NO3 RU12NO3 RU10NO3	<00:00		12:00>	IEPOX RU1400H
 IEPOX UCARB12 RU14OOH PPN C2H5CO3H C2H5CO3 CH3CO3H CH3CO3 PAN RU12OOH RU12O2 DHCARB9 RU14NO3 RU12NO3 RU10NO3 RA16BO2 BENZAL RA16BOOH 	<00:00		12:00>	IEPOX RU1400H
IEPOX UCARB12 RU14OOH PPN C2H5CO3H C2H5CO3 CH3CO3H CH3CO3 PAN RU12OOH RU12O2 DHCARB9 RU14NO3 RU12NO3 RU10NO3 R16BO2 BENZAL RA16BOOH NRUA20OOH NRUA20O2	<00:00		12:00> urrence imesteps)	IEPOX RU1400H
 IEPOX UCARB12 RU14OOH PPN C2H5CO3H C2H5CO3 CH3CO3H CH3CO3 PAN RU12OOH RU12O2 DHCARB9 RU14NO3 RU12NO3 RU10NO3 RA16BO2 BENZAL RA16BOOH NRUA20OOH NRUA20O2 MPAN MACO3 	<00:00 >459 (for a		12:00> urrence imesteps)	IEPOX RU1400H
 IEPOX UCARB12 RU14OOH PPN C2H5CO3H C2H5CO3 CH3CO3H CH3CO3 PAN RU12OOH RU12O2 DHCARB9 RU14NO3 RU12NO3 RU10NO3 RU14NO3 RU12NO3 RU10NO3 RA16BO2 BENZAL RA16BOOH NRUA20OOH NRUA20O2 MPAN MACO3 NRN6O2 NRN6OOH 	<00:00 >459 (for a		12:00> urrence imesteps)	IEPOX RU1400H
 IEPOX UCARB12 RU14OOH PPN C2H5CO3H C2H5CO3 CH3CO3H CH3CO3 PAN RU12OOH RU12O2 DHCARB9 RU14NO3 RU12NO3 RU10NO3 RA16BO2 BENZAL RA16BOOH NRUA20OOH NRUA20O2 MPAN MACO3 NRN6O2 NRN6OOH PECOH DIEK 	<00:00 >459 (for a		12:00> urrence imesteps)	IEPOX RU1400H
 IEPOX UCARB12 RU14OOH PPN C2H5CO3H C2H5CO3 CH3CO3H CH3CO3 PAN RU12OOH RU12O2 DHCARB9 RU14NO3 RU12NO3 RU10NO3 RU14NO3 RU12NO3 RU10NO3 RA16BO2 BENZAL RA16BOOH NRUA20OOH NRUA20O2 MPAN MACO3 MPAN MACO3 NRN6O2 NRN6OOH PECOH DIEK NRN9O2 NRN9OOH 	<00:00 >45% (for a		12:00> urrence imesteps)	IEPOX RU14OOH

Figure 5.11: An alluvial diagram showing the changes in clusters between noon and midnight. On the left are all groups that appear in >45% of the midnight simulation results. On the right are groups which appear >45% of the midday results. In the middle exist the clusters extracted which appear in >45% of all runs. Here it is seen that there exist a series of species which may exist in the daytime or nighttime chemistry, but do not persist between both. Sizes represent the number of species and colours (greys) has no purpose other than to differentiate different items. radical equivalent as a group, RO₂ $\xleftarrow{\rm HO_2}{\rm OH}$ ROOH.

Sepecies Pair	Euclidean	Cosine
NRI15OOH NRI15O2	0.4624	0.2885
NRI12OOH NRI12O2	0.4617	0.2986
PHAN HOCH2CO3	0.5103	0.9998
HOCH2CO3H HOCH2CO3	0.8350	0.9892
RI14OOH RI14O2	0.4922	0.2275
NRN9O2 NRN9OOH	0.4620	0.2818
PECOH DIEK	0.0172	0.0011

Table 5.1: A table of the **normalised** similarity values for the lumped species. Numbers closest to 1 show the worst possible paring in the mechanism, and numbers approaching 0 show the best.



(c) RI14OOH RI14O2

(d) NRI₁₂OOH NRI₁₂O₂

Figure 5.12: Comparing the best (a-b) and worst (c-d) species pairs from Table 5.1. Species which make a good candidate for reduction have a similar diurnal profile and production/loss patterns as well as ranges of magnitude in which the concentration lies. This is seen in subplots (a) and (b). Bad pairings either cover very different magnitude ranges (d) or have dice different temporal profiles (c and d). Time is in the format DD-MM HH

5.9 Conclusions

Chapter 3 discussed graphs as a useful method for representing the chemistry within a mechanism. Building on that Chapter 4 showed that graph centrality metrics could be used to mathematically locate nodes (species) of importance from the chemical network from a chemical simulation. This chapter explores the chemical structure of the MCM network and uses graph clustering methods to locate groups of similar chemistry (Figure 5.5).

This process was trialled, using 300 randomly initiated simulations, on the CRI v2.2 mechanism [Jenkin et al., 2008]. In addition to graph clustering, we use cosine and Euclidean distances to compare the concentration magnitude and profile for species which may be lumped together. These are natural language processing techniques which allow the comparisons of two (temporal) arrays comparing their geometric distance and the angle between them.

For example, only six pairs of species were identified to be potentially lumped together. This has shown that in using the methodology presented, it is possible to located potential candidates - although both the parameters and the sensitivity of grouping these within a chemical simulation need to be further explored. Future work should involve the use of real-world chemical scenarios, and clear optimisation criteria to which to benchmark the results against - if using CRI and the MCM this will likely be ozone-forming potential. Since CRI v2.0 has an additional five reduced states, it would be useful to attempt to reduce that and compare the results against the pre-existing mechanisms.

This chapter has shown a novel way for querying and representing a mechanism. This methodology needs to be further developed and applied to real-world chemistry before any conclusive comments on its applicability to atmospheric chemistry are made.

Bibliography

- Andersson-Sköld, Y., Grennfelt, P., and Pleijel, K. (1992). Photochemical Ozone Creation Potentials: A Study Of Different Concepts. Journal of the Air & Waste Management Association, 42(9):1152– 1158. https://doi.org/10.1080/10473289.1992.10467060.
- Aslak, U., Rosvall, M., and Lehmann, S. (2018). Constrained information flows in temporal networks reveal intermittent communities. *Phys. Rev. E*, 97:062312. https://link.aps.org/doi/10.1103/ PhysRevE.97.062312.
- Blondel, V. D., Guillaume, J.-L., Lambiotte, R., and Lefebvre, E. (2008). Fast unfolding of communities in large networks. *Journal of Statistical Mechanics: Theory and Experiment*, 2008(10):P10008. https://doi.org/10.1088%2F1742-5468%2F2008%2F10%2Fp10008.
- Community, T. I. G.-C. (2020). Geoschem/geos-chem: Geos-chem 12.7.1. Zenodo. https://doi.org/ 10.5281/zenodo.3676008.
- Cornes, P. (2008). Proposed Plans For Holme Pierrepont Whitewater Course. https://hppconcern. wordpress.com/2008/08/04/proposed-plans-for-holme-pierrepont-whitewater-course/.
- Dick Derwent, Andrea Fraser, John Abbott and Mike Jenkin (2010). Evaluating The Performance Of Air Quality Models. online. https://uk-air.defra.gov.uk/assets/documents/reports/ cat05/1006241607 100608 MIP Final Version.pdf.
- Dodge, Y. (2008). Latin Square Designs, pages 297–297. Springer New York, New York, NY. https: //doi.org/10.1007/978-0-387-32833-1_223.
- Ellis, D. (2020). DSMACC-Testing. https://github.com/wolfiex/DSMACC-testing.
- Emmerson, K. M. and Evans, M. J. (2009). Comparison of tropospheric gas-phase chemistry schemes for use within global models. *Atmospheric Chemistry and Physics*, 9(5):1831–1845. https://www. atmos-chem-phys.net/9/1831/2009/.
- Everett, M. G. and Borgatti, S. P. (1994). Regular equivalence: General theory. The Journal of Mathematical Sociology, 19(1):29–52. https://doi.org/10.1080/0022250X.1994.9990134.
- Fortunato, S. (2010). Community Detection In Graphs. Physics reports, 486(3):75–174. http://www.sciencedirect.com/science/article/pii/S0370157309002841.
- GEOS-Chem (2020). Geos-Chem Publications. *online*. http://acmg.seas.harvard.edu/geos/geos_pub. html.

- Huffman, D. A. (1952). A method for the construction of minimum-redundancy codes. Proceedings of the IRE, 40(9):1098–1101.
- Jacobson, M. (2005). Fundamentals Of Atmospheric Modelling. Cambridge University Press. https://www.cambridge.org/core/books/fundamentals-of-atmospheric-modeling/ A6B866737D682B17EE46F8449F76FB2C.
- Jenkin, M. (2019). Http://Cri.York.Ac.Uk . Online.
- Jenkin, M., Watson, L., Utembe, S., and Shallcross, D. (2008). A common representative intermediates (cri) mechanism for voc degradation. part 1: Gas phase mechanism development. Atmospheric Environment, 42(31):7185 – 7195. http://www.sciencedirect.com/science/article/pii/ S1352231008006742.
- Lu, H., Halappanavar, M., and Kalyanaraman, A. (2015). Parallel Heuristics For Scalable Community Detection. *Parallel computing*, 47:19–37. http://www.sciencedirect.com/science/article/pii/ S0167819115000472.
- Mahajan, S. (2008). The Art Of Approximation In Science And Engineering. MIT OpenCourseWare. http://web.mit.edu/6.055/book/book-draft.pdf.
- McKay, M. D., Beckman, R. J., and Conover, W. J. (2000). A comparison of three methods for selecting values of input variables in the analysis of output from a computer code. *Technometrics*, 42(1):55–61. https://amstat.tandfonline.com/doi/abs/10.1080/00401706.2000.10485979.
- Newman, M. E. J. and Girvan, M. (2004). Finding and evaluating community structure in networks. *Phys. Rev. E*, 69:026113. https://link.aps.org/doi/10.1103/PhysRevE.69.026113.
- Oran, E. and Boris, J. (1991). Numerical approaches to combustion modeling. progress in astronautics and aeronautics. vol. 135. U.S. Department of Energy Office of Scientific and Technical Information.
- Ostrovsky, V. N. (2005). Towards a philosophy of approximations in the 'exact' sciences. *Hyle*, 11(2):101–126.
- Raghavan, U. N., Albert, R., and Kumara, S. (2007). Near linear time algorithm to detect community structures in large-scale networks. *Phys. Rev. E*, 76:036106. https://link.aps.org/doi/10.1103/ PhysRevE.76.036106.
- Rickard, A. (2020). MCM Website. http://mcm.york.ac.uk/.
- Rosvall, M., Axelsson, D., and Bergstrom, C. T. (2009). The map equation. The European Physical Journal Special Topics, 178(1):13–23. https://doi.org/10.1140/epjst/e2010-01179-1.

- Rosvall, M. and Bergstrom, C. T. (2010). Mapping Change In Large Networks. *PloS one*, 5(1):e8694. http://dx.doi.org/10.1371/journal.pone.0008694.
- Sad CRUD Developer (2016). The Huffman Tree. *StackOverflow*. https://i.stack.imgur.com/9T1Am. png.
- Sandu, A. and Sander, R. (2006). Technical note: Simulating chemical systems in fortran90 and matlab with the kinetic preprocessor kpp-2.1. Atmospheric Chemistry and Physics, 6(1):187–195. https://www.atmos-chem-phys.net/6/187/2006/.
- Turanyi, T. (1990). Reduction Large Reaction Mechantsms. New journal of chemistry = Nouveau journal de chimie, 14:795–gO3. http://garfield.chem.elte.hu/Turanyi/pdf/14_Turanyi_NJC_1990. PDF.
- Turányi, T. (1990). Sensitivity Analysis Of Complex Kinetic Systems. Tools And Applications. Journal of mathematical chemistry, 5(3):203–248. https://doi.org/10.1007/BF01166355.
- Turanyi, T. and Tomlin, A. (2015). Analysis Of Kinetic Reaction Mechanisms. Springer. http: //eprints.whiterose.ac.uk/84294/.
- Turányi, T. and Tomlin, A. S. (2014). Reduction Of Reaction Mechanisms. Springer Berlin Heidelberg, Berlin, Heidelberg. https://doi.org/10.1007/978-3-662-44562-4 7.
- Vajda, S., Valko, P., and Turainyi, T. (1985). Principal component analysis of kinetic models. International Journal of Chemical Kinetics, 17(1):55–81. https://onlinelibrary.wiley.com/doi/abs/10. 1002/kin.550170107.
- Welton, W., Benso, S., and Bowery, A. (2002). *Plato'S Forms: Varieties Of Interpretation*. G
 Reference, Information and Interdisciplinary Subjects Series. Lexington Books. https://books.google.co.uk/books?id=vbtbQk A0YoC.
- Whitehouse, L. E., Tomlin, A. S., and Pilling, M. J. (2004a). Systematic reduction of complex tropospheric chemical mechanisms, part i: Sensitivity and time-scale analyses. *Atmospheric Chemistry* and Physics, 4(7):2025–2056. https://www.atmos-chem-phys.net/4/2025/2004/.
- Whitehouse, L. E., Tomlin, A. S., and Pilling, M. J. (2004b). Systematic Reduction Of Complex Tropospheric Chemical Mechanisms, Part Ii: Lumping Using A Time-Scale Based Approach. Atmospheric Chemistry and Physics, 4:2057–2081. https://www.atmos-chem-phys.net/4/2057/2004/ acp-4-2057-2004.pdf.
- Wyche, K. P., Monks, P. S., Smallbone, K. L., Hamilton, J. F., Alfarra, M. R., Rickard, A. R., McFiggans, G. B., Jenkin, M. E., Bloss, W. J., Ryan, A. C., Hewitt, C. N., and MacKenzie, A. R.

(2015). Mapping gas-phase organic reactivity and concomitant secondary organic aerosol formation: Chemometric dimension reduction techniques for the deconvolution of complex atmospheric data sets. *Atmospheric Chemistry and Physics*, 15(14):8077–8100. https://www.atmos-chem-phys.net/ 15/8077/2015/.

Zhou, H. (2003). Distance, dissimilarity index, and network community structure. *Phys. Rev. E*, 67:061901. https://link.aps.org/doi/10.1103/PhysRevE.67.061901. Chapter 6

Computational Learning of Species Structure using Visualisation and Vector Clustering

"So, in the interests of survival, they trained themselves to be agreeing machines instead of thinking machines. All their minds had to do was to discover what other people were thinking, and then they thought that, too."

- Kurt Vonnegut, Breakfast of Champions

6.1 Introduction

6.1.1 Historical Significance

The established process of trial and error has always underpinned our survival [Noble, 1957]. Babies are born to rely on a set of sensory reflexes and a framework for physical reasoning [Baillargeon and Carey, 2012], and with these, we develop methods to navigate the influence of change within a physical, and auditory space [Lynch, 2011]. This method of decision making is reflected in our adult lives with ideas and actions being limited in choice by our intuition and experience [Descartes and Lafleur, 1960]. In science, we apply a methodological framework consisting of a continuous assessment of scepticism, educated guessing (hypothesising) and rigorous practical testing. Specialists accrue years of practical and theoretical knowledge within a narrow field and can identify areas of potential gain and futility. Nevertheless, even with all prior experience, the discovery of new and untested techniques involve the tortuous traipsing through a sea of uncertainty.

Such methods sometimes prove fruitful, through accidental discoveries of items such as polyetheylene, penicillin, x-rays, nylon, teflon, velcro etc. [Roberts, 1989]; finding novel applications for existing methods such as optical tweezers for chemistry or the abstract field of maths utilised by Einstein, but more often than not end in the constant evolution of a pre-existing project with no apparent result.

6.1.2 Theory And Simulation In Science

Until recently much of the experimentation possible was limited by resources, levels of knowledge and available technology. With the increase of computation power, we have been able to not only increase our understanding but also run theoretical simulations to guide exploratory efforts with an impact on real-world applications [Oliveira et al., 2006; T. Leube et al., 2018; Morozov, 2016; Yu-ChenLo, 2018]. However, as our ability to record and produce data increases, the need for the scientific method diminishes [Anderson, 2008]. Here the application of 'big data' tools and algorithms can provide insights and correlations much more compelling than the predictive capabilities of constantly changing models - "Since all models are wrong the scientist cannot obtain a "correct" one by excessive elaboration" - Box [1976]. As our level of attainable technology increases, so does the complexity of the data collected. New datasets tend to be very large, complex and highly multivariate. Although this dramatically improves the quality of science, the difficulty lies in trying to represent it in such a way that we may efficiently access the reliability of the results. Since simple bar and line graphs are no longer applicable, one solution falls within a class of unsupervised machine learning techniques called dimensionality reduction (DR).

6.1.3 Chapter Aims

In Chapter 2, we looked at visual representation as a way of better understanding of complex systems. Chapter 3 showed that the chemical properties could be inferred (visually) from the node-link graph structure of a mechanism. Similarly, Chapter 4 and Chapter 5 located the presence of important species and clusters of similar properties by applying mathematical algorithms to the graph network. As opposed to attempting to visualise complex data, this chapter looks at learning the structure of a chemical species and simplifying it into two dimensions. Here it is possible to extract key features of like-groups through the use of vector clustering, which unlike the graph clustering in Chapter 5 works by determining the density between points on a plane.

The chapter begins with the introduction of the chemical system, and the various methods for representing species structure within it (Section 6.2). Next, we define the dimensionality reduction methods, which are to be used to simplify the inputs above (Section 6.3). This is followed by a brief overview of the visualisation methodology (Section 6.4). Finally, all three sections are combined to produce a set of result and conclusions about the use of DR to identify species structure. This chapter aims to access the efficiency of a machine learnt (dimensionality reduced) models in simplifying the chemical structure, and decide upon the best input for future deep learning tasks (e.g. for mechanism construction and emulation).

6.2 Species Of The MCM And Ways To Represent Them.

The master chemical mechanism (as defined in all previous chapters), represents our foremost knowledge of gas-phase chemistry within the troposphere. Chapter 3 shows that information about a species structure is encoded within its reactions, much of which can be attributed to the well-defined construction protocols.

This section explores the different methods of representing a species structure, intending to provide a machine built algorithm with the highest amount of information about each species and its functionality. A range of input types will be evaluated against several dimensionality reduction algorithms to isolate which chemical properties are most extractable.

6.2.1 Input Generation

The MCM database provides species information in the form of a species 'SMILES' (Subsubsection 6.2.3.2) and the IUPAC InChi string [Heller et al., 2013]. Within this chapter, we use only the SMILES string, which is either manually processed using regular expressions or with the aid of pythons RDKIT package [Landrum et al., 2019]. There are seven different methods for representing the chemistry; these are outlined below.

6.2.2 Manual Categorisation

Reactions within the MCM are determined by a set of rules (Figure 2.9). These mimic the process a chemist may use to build a species degradation mechanism and often rely on understanding the bond availability and functionalisation of a species. Since the present functional groups are the benchmark of whether a DR algorithm has successfully separated species structure, it makes sense to run a unit test using the known functional groups of a species as the input.

To generate the functional groups the regular expressions in Table 6.1 are used¹ on the SMILES strings (described in Subsubsection 6.2.3.2) for each species. In extracting the functional groups, we can plot the likeliness a species with a certain group is likely to have another using a chord diagram - Figure 6.1. Since most species contain a multitude of functional groups, the separation of these into 'tidy' clustered groups seems unlikely.

PAN	C\\(=0\\)00N\\(=0\\)=0\$ ^\\[0-{0,1}\\]\\[N\\+{0,1}\\]\\(=0\\)00C 0=N\\(=0\\)00C\\(=0\\) C\\(=0\\)00\\[N\\+{0,1}\\]\\(=0\\)\\[0-{0,1}\\]
Carb. Acid	[^0](C\\(=0\\)0\$ ^0C\\(=0\\))
Ester	[\^0](C\(=0\)0\b 0C\(=0\))C
Ether	(\([\^O=]+\))*C(\([\^O=]+\))*O(\([\^O=]+\))*C(\([\^O=]+\))*
Per. Acid	c\\(=0\\)00\$ ^00\\(=0\\)C
Nitrate	O(NO2\b NOO\b N\(=O\)=O \[N\+\](?:\[O-\\] \(=O\)){2})
Aldehyde	C=0\$ ^0=C
Ketone	C\(=O\)C
Alcohol	CO/\b (?=^\\b)(?!^\\[)CO. (?=^\\b)(?!^\\[)OC. \\(O\\) C\\)O(\\b [^O]
Criegee	\[0-\]\[0\+\]
Alkoxy rad	\[[\/]{0,1}CH{0,1}\] \b[\^0]\[0\.{0,1}\]
Peroxyacyl rad	\\ w\(=0\)0\[0\.{0,1}\]

Table 6.1: A set of regular expressions that are used to determine the number of occurrences of a functional group within a SMILES string. These were written to scan the SMILES string a match specific patterns corresponding to each functional group. A similar process is used within [rdkit, 2019] to construct MACCS keys (discussed later).

 $^{^1\}mathrm{To}$ see the structure of each functional group type, go to Section D.2.


Figure 6.1: **The multifunctionality of the MCM.** A chord diagram showing the functionalisation of all species within the MCM v3.3.1. Arc sizes represent what percentage of all functional groups in the MCM mechanism a group contains. Translucent areas of no outwards links represent species with multiples of a certain functional group, of which Alcohols and Ketones have the most. Source: [Ellis, 2019]

6.2.3 Tokenization

As computer algorithms are unable to understand words or their meaning, we have first to categorise the data into groups. Tokenisation is the conversion of a string into characters and representing them with a numerical equivalent. In doing so, a string of characters can be converted into a numerical vector, allowing for its representation in a latent vector space. Within our input selection, we have two

 $^{^2\}mathrm{Check}$ the correct table has been used.

sets of inputs we can convert. These are the species names, and their SMILES string representation.

6.2.3.1 Species Names

In Chapter 5 it was shown that the dedicated species names for species in the CRI mechanism were often representative of their structural properties. This also applies for the MCM, where an intuitive naming convention following the FACSIMILE format is used. This is often derived as part of the construction protocol, where a species names reflect its own, or its precursor's structure (which it will have at least in-part inherited). Although this is not the most robust method of defining the structure, it allows for a straightforward test of the algorithms, for which the user can quickly compare the human-readable output.

6.2.3.2 SMILES Strings

SMILES ('Simplified Molecular-Input Line-Entry System') provide a human-readable representation of the molecular structure, [Weininger, 1988]. They offer a linear human-readable description of the chemical composition within a molecule - making it easy to visually check the construction of a species without any additional work. Besides, their role in generating the molecular fingerprints in Subsection 6.2.5, SMILES strings provide a useful tool for quickly comparing species structure.

Construction Methodology of SMILES strings

The construction of a SMILES string happens in three parts:

- The SMILES string is built by creating the longest possible chain to form a molecule backbone. Figure 6.2b
- 2. This may within itself contain aromatic rings denoted by the lowercase carbons and a number corresponding to the location of each break cycle. Figure 6.2c
- 3. Finally all the functional groups and branches attached to the main backbone are added. These are nested within the parenthesis to show that they are not part of the skeletal backbone. Figure 6.2d



Figure 6.2: Construction process of a SMILES string. The example compound is Melatonin. Although this does not exist within the atmosphere, it provides a clear example of the SMILES string methodology. Figure 6.2a is made using SMILES drawer: [Probst and Reymond, 2018]

6.2.4 Graph Inspired

Chapter 3 - 5 have shown the role of graphs in revealing network properties and structure. Graphs in themselves can simplify relational data into two/three dimensions for visualisation and algorithmic clustering. Continuing this trend, we can represent a species structure in the form of a graph (Subsubsection 6.2.4.1), as well as converting the structure of a mechanism for dimensionality reduction (Subsubsection 6.2.4.2)

6.2.4.1 The Species Graph (Fingerprint)

The structure of a species has long represented using a graph-like layout, Chapter 3. It, therefore, follows that other methods for representing the graph structure would also apply. One such way is the use of an adjacency (or relational) matrix to describe the relationships between atoms and bonds in a species. Such a methodology is already used in the construction of bond and z-matrixes [Aumont et al., 2005; Parsons et al., 2005].

The construction of a structure matrix/graph begins with a chemical species. Here the relationships between atoms (Figure 6.3b) is converted into an adjacency matrix (Figure 6.3c). However, since species have different numbers of each atom, a template allowing us to compare different graphs is required. To do this a maximum occurrence table (Figure 6.3a) is created. Here, for example, β caryophyllene (BCARY) C₁₅H₂₄, contains the most carbon atoms of any species within the MCM. This universal matrix is now able to contain any possible combination of atoms in a species. As machine learning algorithms only use vectors as an input, it is possible to decompose the 37^2 element adjacency matrix into rows, which can then be joined together, Using this method we create a one-dimensional array (vector) of 259 elements (518 bytes) to represent our species.



Figure 6.3: Constructing a graph from species structure. (a) shows the maximum number of times an atom occurs for any single species in the MCM. (b) depicts the graph-like chemical structure of INB1NBCO3(a product from isoprene). This is a highly processed species stemming from Isoprene, and this makes for a good example of the bond matrix. Finally, a matrix representing the bonds in INB_1NBCO_3 is created from the maximum possible occurrence matrix from (a). For simplicity, empty row/column pairs have been removed to produce (c). This matrix will always be symmetrical as the bonds do not have a direction.

6.2.4.2 Node Embeddings (Node2Vec)

Chapter 3 and Chapter 4 showed that the underlying structure of a chemistry mechanism graph contains information about the species and reactions within it. Here as a species is oxidised the O-C ratio increases. Long-chain VOCs are likely to fragment into two radicals, producing smaller more oxidised species. Eventually, this process leads to the production of carbon dioxide and water. Figure 6.4 shows a subset of the MCM representing the chemistry in Beijing. Node colour and size show the increase of oxidation as species head towards CO at the centre) - lighter colour and larger node.

This type of structural information can be extracted through the use of a natural language processing package capable of transforming a graph into a vector - Node2Vec [Grover and Leskovec, 2019]. Since this may also be used for dimensionality reduction, it is described within the next section (Subsection 6.3.6).



Figure 6.4: A graph of an MCM subset representing the chemistry within Beijing. Here colours show the increase of O-C ratio as species are oxidised (lighter). All emitted species ultimately tend towards carbon monoxide, which is at the centre of the graph. Node clusters symbolise groups of species which react more between themselves and less with others (This graph only represents the mechanism structure).

6.2.5 Molecular Fingerprints

In the field of chemical informatics, molecular fingerprints (or structural keys) are used to encode and query structural properties of species. Their binary representation makes them suitable for dimensionality reduction and the exploration of chemical space (a type of property space constructed using pre-determined features and boundary conditions).

Here species properties are often split into structural and psyico-chemical groups - which has used such as the discovery of natural analogues (which circumvent problems such as intolerances in medicine³

 $^{^{3}\}mathrm{Certain}$ molecules can cause all ergies in people, but whist their natural analogues do not.

[Spahn et al., 2017]). Although there exist many different types of molecular fingerprints, the two main ones that will be explored are molecular quantum numbers (MQN) and the molecular access system (MACCS).

6.2.5.1 Molecular Quantum Numbers (MQN)

In chemistry the shape, phase and electron occupancy of an atom may be described through the use of four quantum numbers: the *n* principle quantum number, *I* angular momentum quantum number, M_i magnetic quantum number and M_s spin quantum number. The rationalisation of elements based on their structure, and by consequence reactivity, has led to the most iconic tool of the modernday chemist - the periodic table, where increasing atomic numbers follow the principal quantum number [Wang and Schwarz, 2009]. In representing a molecule as a set of 42 quantum numbers, MQN fingerprints produce a multi-dimensional mapping of an atom, bond, polarity and topology count [Nguyen et al., 2009].

6.2.5.2 Molecular Access System (MACCS)

MACCS keys are a 164⁴ bit structural keys formulated through answering a series of structure-related questions. Developed by MDL Information Systems [MDL, 1984], their main purpose lies in being a SMILES Arbitrary Target Specification (SMARTS) system for substructure searching. However, their distinct structure key format makes them highly suitable for similarity detection. In many cases, the optimised version of MACCS keys is cited ([Durant et al., 2002]), although most use cases exploit a variation of the undocumented 166bit keys. We use the implementation presented by [Landrum et al., 2019; rdkit, 2019] for all molecular fingerprints in our work.

6.3 Dimensionality Reduction Methods

In the last section, we described several methods in which the chemical structure of a species could be encoded for direct comparison. However, since each input consists of a multitude of elements, it is still not a simple task to determine the differences and similarity between all species in mechanisms. Dimensionality reduction is the process of reducing the number of random variables and only presented a set of principal values, by mapping a high-dimensional space into a low-dimensional one [Roweis and Saul, 2000]. This allows us to flatten a multivariate input into the two dimensions required for a simple scatter plot.

 $^{^{4}}$ They are 166-bit keys, although there is no real agreement to what the 44th keys' purpose is, and therefore it is often omitted. Within RDKIT this is denoted by a ? [rdkit, 2019].

In this section, we begin by explaining the data preparation required for dimensionality reduction (Subsection 6.3.1) before describing the different possible methods of reducing the dimensions of a dataset through Principle Component Analysis, Auto Encoders and t-Distributed Stochastic Neighbor Embedding.

6.3.1 Preparation Of The Data

Real-world data is rarely preformatted in such a way that it can be used directly within a computational model. Often values need to be cleaned and corrected to be fit for purpose. In the interest of completeness, the two main methods of data adjustment for machine learning are outlined below. These are (i) normalisation and (ii) standardisation.

(I) Normalisation

In the data is without (dimensionless) or of a single unit, it is possible to rescale the data between a range - most commonly 0,1. In doing so, it is possible to interpret the importance of value in contrast to the largest recorded value. This gives us a percentage scale spanning the range of the data. Such a range is useful in the definition of colourmaps and describing the importance of value relative to the dataset. To rescale a dataset, we shift the minimum value to zero, then divide by the new maximum of the dataset (Note this is equivalent to the range of the unshifted dataset.)

$$n(x_i) = \frac{x_i - \min_x}{\max_x - \min_x} \tag{6.1}$$

229

(Ii) Standardisation

If the components we wish to compare are of different units or are expressed with a different scale, normalising them can not produce meaningful data. Instead, it is possible to standardise the data by looking at each points deviation from the mean. Here the variation of the mean for a dataset is divided by the standard deviation to produce a value between $\{-1,1\}$, Equation 6.2. In statistics this is known as the 'z-score'⁵

$$z(x_i) = \frac{x_i - \mu_x}{S} \tag{6.2}$$

⁵Possibly because of the American spelling of standardiZation?

6.3.2 Principle Component Analysis (PCA)

One of the most well-known dimensionality reduction methods is the determination of the principal components through the use of Principal Component Analysis (PCA). PCA increases the readability of a dataset by creating a set of new uncorrelated variables which maximise the variance [Jolliffe and Cadima, 2016].

PCA works on the assumption that components within a dataset are linear combinations of each other. By simplifying these linear combinations, it is possible to identify the elements which explain the most variability in a dataset - these are the principal components.

A more straightforward interpretation of this would be to adjust the direction of each axis of the data, such that its projection has the most prominent variability. In doing so, it is possible to determine which components contribute the most to changes in the dataset [F.R.S., 1901; Hotelling, 1933]. An example of this is seen in Figure 6.5, where the second component of the original data can be removed with little effect on the overall result of the data. Such methods have applications in compression and signal filtering [Hernandez and Mendez, 2018; Hamadache and Lee, 2017].



PCA is useful for eliminating dimensions. Below, we've plotted the data along a pair of lines: one composed of the x-values and another of the y-values.





If we're going to only see the data along one dimension, though, it might be better to make that dimension the principal component with most variation. We don't lose much by dropping PC2 since it contributes the least to the variation in the data set.



Figure 6.5: **Determining the Principal Compnent of a sample dataset.** It can be seen that in a change in axis to follow the first principal component (right), it is possible to explain most of the variation in the samle dataset (left). Source: [Powell, 2020]

6.3.2.1 Mathematical Explanation Of PCA

Note: The basic statistics/mathematics required to understand this section is shown in Section A.1. Please read this if you are not familiar with any of the terms below.

The mathematics behind PCA consists of first calculating the covariance matrix - a $n \times n$ matrix outlining how strongly each variable changes with every other. Using this we can calculate both the eigenvalues and eigenvectors of the matrix ⁶. This can be done using a computational package such as numpy or scipy [Oliphant, 2006; Jones et al., 01].

We can now sort the eigenvector columns by influence using their eigenvalues—this way a feature dataset can be produced by removing vectors of low importance. The final feature dataset can now be transposed and multiplied by the transpose of the original dataset. This results in an output dataset containing each principal component of the desired dimension.

6.3.3 t-Distributed Stochastic Neighbor Embedding (t-SNE)

t-SNE is an algorithm designed with visualisation in mind [Maaten and Hinton, 2008]. Rather than representing the data through a series of linear transformations, t-SNE uses local relationships to create a low-dimensional mapping, much in the same way as a fully connected force graph, as shown in Figure 6.6. This allows the ability to capture non-linear structures in the data which cannot be accomplished through linear mapping methods (e.g. PCA).



Figure 6.6: **Representing the t-SNE algorithm as a fully connected force graph.** Here each node is attached to every other node. Nodes with a strong relationship are pulled closer together than those with a weaker one.

⁶These need to be unit vectors, although most packages already do this out of the box.

The algorithm itself can be simplified into two parts, (see A and B below), and is described in Subsubsection 6.3.3.1.

- A. Create a probability distribution which dictates relationships between neighbouring points
- B. Recreate a lower-dimensional space following the probability distribution established in A.

The main reason t-SNE produces good results is that it can handle the 'crowding problem' very well. The crowding problem is a product of the 'curse of dimensionality' [Indyk and Motwani, 1998]. In a high dimensional space, the surface of a sphere will grow much quicker than one in a lower dimension space. This means that the higher dimension spaces will have more points at a medium distance from a certain point, Figure 6.7. When we map our data into a lower dimension, data will try to gather at its medium distance, resulting in a more 'squashed', and thus crowded, output.



Figure 6.7: An example of how the curse of dimensionality affects the mapping of points a certian distance from eachother.

6.3.3.1 Mathematical Explanation Of t-SNE

In the original paper [Maaten and Hinton, 2008], the algorithm is described using the etymologic dissection of its name (described below).

(i) Step 1

First we begin with Stochastic Neigbour Embedding (SNE) - the distribution across neighbouring datapoints in our high dimension space. This is done by converting the high dimensional Euclidian distances between points into conditional probabilities representing their similarity:

$$p_{ij} = \frac{\exp(-\|x_i - x_j\|^2 / 2\sigma_i^2)}{\sum_{k \neq l} \exp(-\|x_k - x_l\|^2 / 2\sigma_i^2)}$$
(6.3)

Here $p_{i|j}$ is the conditional probability that x_i may pick x_j as a neighbour. This is proportional to the probability density of a Gaussian σ_i centered at x_i .

Perplexity

Since we want the number of neighbours of each point to be similar in number and prevent a single

point from having a disproportionate influence on the entire system we introduce a hyperparameter named *perplexity*. Perplexity works by ensuring that σ_i is small for points in densely populated areas and large for spare ones and can be thought of as a scale of the number of neighbours considered for any one point in the system. Generally, values between 5 and 50 are considered to give good results, with larger perplexities taking global features into account, and by consequence smaller ones, local features [Maaten and Hinton, 2008].

(ii) Step 2

Now a probability distribution describing the relationship between points has been formulated, we wish to express this as a low dimensional mapping of our inputs X in terms of our output dimensions Y. Naturally, we would want to make the low dimensional mapping represent a similar (Gaussian) distribution as in Step 1. However, it often causes issues presented by the 'overcrowding problem', Subsection 6.3.3, as the gaussian has a 'short tail', and thus nearby points are likely to be pushed together. A solution to this is the student t-distribution which has a longer tail ⁷:

$$q_{i|j} = \frac{(1 + \|y_i - y_j\|^2)^{-1}}{\sum_{k \neq l} (1 + \|y_k - y_l\|^2)^{-1}}$$
(6.4)

233

Note: The definition and explination of the Student t-distribution is given in Section A.2.

The optimisation of this equation is achieved through the use of gradient decent⁸ on the Kullback-Leibler divergence ?? between distributions p and q. Here the gradient is used to apply an attractive and repulsive force on the items⁹.

6.3.4 PCA vs t-SNE, A Quick Comparison.

PCA has been around for much longer than t-SNE, and its uses are well established within the scientific community. In essence, an example of this gives by Wyche et al. [2015] where mechanisms can be separated into different pathways (on account of the underlying chemistry) and Turanyi and Tomlin [2015] where sensitivity analysis is used within mechanism reduction. It is fast, simple and easy to use and very intuitive. The PCA algorithm works by creating a lower-dimensional embedding which

 $^{^{7}}$ The distribution employed is a t-distribution with only one degree of freedom and is identical to the Cauchy distribution

⁸Gradient Decent - an optimisation algorithm used to minimise a function by iteratively moving in the direction of the steepest descent. Gradient descent is used to find local minima and is defined by the negative of the gradient of the system. Its primary usage in machine learning is the updating of parameters (coefficients in linear regression and weight in neural networks).

 $^{^{9}\}mathrm{A}$ positive gradient signifies attraction, while a negative one corresponds to repulsion.

best preserves the overall variance of the dataset. Clusters created from the algorithm are grouped in ways, such that they retain the highest variance of the data.

The main drawback of PCA is that it is a linear projection. If our data happened to be in a 'swiss roll' (spiral) pattern, we would not be able to 'unroll' it. The reason for this is that the PCA algorithm works by viewing the data from different perspectives, much like casting a shadow from various directions. With such an example, there is no one way we can do this that unfurls the spiral.

t-SNE, on the other hand, is a relatively new method [Maaten and Hinton, 2008]. Its greatest asset is that linear projections do not limit it. Although more computationally intensive for large datasets, t-SNE produces visibly cleaner results. Unlike in PCA, t-SNE cannot be trained on additional data at a later point; however, the output clusters are more visually distinct (they have less of an overlap). Much like in a force graph, the output from t-SNE is scale-invariant. This means that while the location of clusters in a PCA reduced representation has an attributable quality, those produced by t-SNE will not necessarily contain the same information.

A box model run representative of the chemistry within Beijing was used to compare the differences between PCA and t-SNE. The aim is to classify the diurnal profiles of each species concentration (much like the cosine similarity in Subsubsection 5.7.1.2). Diurnal profiles were extracted on the third day of a spun up model initialised with initial conditions representative of the chemistry within the Beijing environment (Table 4.4). These were then standardised and converted into temporal vectors for use in the algorithms.

Figure 6.8 shows the output of both dimensionality reduction algorithms on the dataset. Different colours represent the location of clusters of similar diurnal profiles. A higher dispersion between clusters and species overlap is seen within the PCA output, Figure 6.8a. This makes it harder to distinguish species from each other or other groups around them. Since the distance between clusters within t-SNE does not hold the same mathematical meaning as PCA, the algorithm can provide a better distribution of points, creating better-defined clusters, Figure 6.8b. The concentration profile shapes for each coloured group is shown in Figure 6.8c.



(a) PCA

(b) t-SNE



(c) t-SNE with cluster outlines.

Figure 6.8: Showing the difference between PCA and t-SNE clustering. These figures show the clustering of a set of standardized species concentration profiles (c) across two styles of dimensionality reduction: PCA (a) and t-SNE (b).

6.3.5 The Auto-Encoder (AE)

Auto-encoders are a subclass of neural networks with primary use in compressing data (dimensionality reduction). Rather than predicting a numerical output, AutoEncoders focus on the construction and deconstruction of data through the use of an encoder and decoder pair. The encoder takes an n-dimensional input and applies a compression, reducing it to the number of dimensions in the bottleneck layer. The reduced dataset is then reconstructed within the decoder. Such a process not only allows for an easy understanding of the error of the reduced data but can also be used in the filtration of

noisy or pixelated data [Leite et al., 2018; Dataman, 2019] and as an input to more complex machine learning models.



Figure 6.9: An example autoencoder structure which reduces a 16 dimensional input to 2. Draw with the aid of [Krizhevsky et al., 2012]

There are two features of an autoencoder that make it a particularly powerful tool. The first is the ability to sample the latent space using the decoder. The implications of this are that we can establish features that correspond to gaps between our data points - which can have its application if the data used is sparse or incomplete. Next comes the inherent non-linearity of the model. As an autoencoder is just a neural network, the amount of information passed through each link between layers is governed by an activation function. Should this activation function be linear, the reduced dimension will be much akin to a PCA decomposition. Where PCA reduces the dimensions of a dataset by discarding those with little effect on the variance, an autoencoder opts to combine it. Here the entirety of the dataset remains encoded within the links of the AE network. To decide how data flows along the edges of the network, a series of threshold (activation) functions are used for each layer. These are described in B.

6.3.5.1 Demonstration Of Non-Linear Activation Functions

To demonstrate the effect of these features, we take a sample isopleth of Methane, NOx and Ozone from 300 box model simulations and reduce it to two dimensions. This is then reconstructed back into three dimensions using the DR algorithms. Figure 6.10 shows the difference between the original dataset (Figure 6.10a) and that of the PCA (Figure 6.10b) and AutoEncoder (Figure 6.10c) reconstructions.

Here we see a loss in the non-linearity of the original data for the PCA reconstruction. However, the use of a non-linear (tanh) activation function within AutoEncoder produces a result much closer to the original. Use of a linear activation function, however, produces a similar result to the PCA algorithm.



Figure 6.10: Comparing the result of the 2D encoding and decoding of an Ozone-NOx-Methane isopleth. The original data (a) is reduced to two dimensions and then reconstructed back into 3D. This is done with Principal Component Analysis (b) and an AutoEncoder (c). The original isopleth is created using 300 simulations of different initial conditions: NOx (variable), Methane (variable) and Ozone (constant). These were designed using a latin hypercube and converted into a surface plot using Delaunay triangulation.

6.3.6 Node2Vec

Finally, Node2Vec is an embedding algorithm designed to generate vector representations of the nodes in a *undirected* and *unweighted* network. Although it can be used to reduce a complex network into a 2D vector (dimensionality reduction), for this work we shall only use it to generate a fingerprint for a species' position within a mechanism network graph - and then apply this as an input to the DR methods above. This method of input creation has been found more computationally efficient, by circumventing the need for expensive composition, in producing better predictions on network-related tasks compared to more classical methods such as PCA [Grover and Leskovec, 2019].



Figure 6.11: The process of converting a graph into a vector using Node2Vec. Source: [Cohen, 2018]

The process of converting the graph structure (Figure 6.11) into a numerical vector node embedding starts by taking a series of 2^{nd} order random walks. These describe the neighbourhood of a node in the form of a set of random walk paths, much in the same way words are dependent on their neighbours within a sentence- for example in the OH initiated degradation of isoprene in the MCM result in the following path along with the mechanism graph.:

$$ISOPRENE \to OH \to TISOPA \to ISOPBO_2 \to TISOPA \to \dots$$
(6.5)

This methodology allowed for the use of word2vec algorithm, converting the walk into a vector (Subsubsection 6.3.6.2)

6.3.6.1 Sentence Construction By Sampling Of A Network

The probability and path (as described above) depend both on a set of arguments, and a random $seed^{10}$ provided to the model.



Figure 6.12: Calculation of the random walk path. Source: [Grover and Leskovec, 2019]

Figure 6.12 shows the return and input parameters (p & q) determine how fast we explore the network and our probability to leave the neighbourhood. In a system, where the previous path is from t to v, we may calculate the probability of returning to t as 1/p, going to a mutual node connected between t and v as 1, and viewing a new node as 1/q. If q > 1 we have a high probability to end up at nodes close to t, and with q < 1 we are likely to explore other nodes. Additionally if we chose $p > \max q, 1$ we are less likely to return to an already visited node $(p < \min q, 1$ is likely to generate a backwards step). Since we wish to generate a 'local' view, but do not wish to return to t we select $q \ge 1$ and p > q our parameters as p = 2.0, q = 1.1. In the case of a weighted graph (something that we are *not*

 $^{^{10}}$ Computers can never generate truly random numbers. If we want reproducibility within models, the random number generator can be initiated with the same seed.

exploring within this chapter) the resultant *alpha* value calculated is further multiplied by the edge weight.

To generate the Node2Vec embeddings for each species, we use the python2 code provided by the original paper by Grover and Leskovec [2019] with a set of 50000 random walks, each of length 9 product/reaction generations. The reasoning behind this is that we have a large graph, with a power-law like structure (where species are often heavily connected, Chapter 4).

NOTE: This process takes over a week to compute (in serial), and then the binary file containing all walks in character form approaches 10 GB, for the complete MCM.

6.3.6.2 Word2Vec

Once we have constructed our random path 'sentences' (e.g. Equation 6.5), we can make use of Googles word2vec algorithm [Mikolov et al., 2013]. This is similar to an auto-encoder in many regards; however, the algorithm looks at neighbouring words (or species) in the corpus rather than learning word embeddings using reconstruction. This form of representation has found many uses beyond the realm of natural language processing. Some of these are objects, people, code, tiles, genes and graphs [Lynch, 2011; People2Vec, 2019; Alon et al., 2019; Jean et al., 2018; Du et al., 2019; Narayanan et al., 2017].

6.3.7 Summary Of Dimensionality Reduction Methods

There exist several methods of reducing a complex dataset into a smaller one. PCA is the simplest method to understand but is constrained to linear decompositions. AutoEncoders can have both a linear and non-linear response, based on the activation functions that they use, and t-SNE applies a non-linear grouping which mimics a complete force-directed graph.

Having defined each method, we next explain how they will be evaluated (Section 6.4), before applying them to the MCM in Equation 6.5.

6.4 Visualisation Of Clustering

In assessing the validity of clustered space, we require a level of exploratory data analysis. To reveal features of interest, we plot the reduced 2D dataset and apply interactivity coupled with a selection of visualisation techniques described below. This section outlines the different visualisation methods used.

6.4.1 Viewing The 2D Species Embeddings

Since the different DR algorithms return data on various scales, comparison between the outputs is not straightforward. To overcome this outputs in x and y are normalised (scaled between $\{0,1\}$), Subsection 6.3.1, before being plotted as a scatterplot.

6.4.2 Exposing Overlapping Data

If the nodes within a tight-knit cluster overlap, this can cause obfuscate the results and limit the user's ability to select them. As an initial test, node sizes can be reduced. However, this may often result in points too small to pick. Another solution is to create a force-directed graph where each point is strongly attracted to its initial position. Here we can apply collision detection, while still preserving the overall grouping of nodes within a cluster - a technique that was seen in Chapter 5.

6.4.3 Gooey Effect (Gaussian Blur)

Taking a quote from Reinhardt [1975]: "The more stuff in it, the busier the work of art, the worse it is. More is less. Less is more." and combining it with the work from Chapter 2, we realise that showing each species, when observing overall clusters just add unnecessary clutter to the images. Instead, since we are only interested in the clusters as a unit, a 'gooey effect' filter can be applied. This works by merging nearby points into a single water-like blob using a gaussian blur¹¹. Here since each point is allocated a colour, if a colour gradient exists, then there are multiple clusters occupying the same place. The aim of this is to reduce the cognitive load on the end-user by reducing the number of distinct objects that they need to take in.

6.4.4 Four Colours Theorem

When plotted, the number of clusters detected often exceeds the number of categorical colours available. In cartography, it has been noted that the colouring of neighbouring polygons should at most take four colours. This is the origin of the four colours theorem [Appel and Haken, 1976], of which a greedy implementation is applied.

The aim of this is to show item boundaries (for instance countries, or in our case clusters) while reducing ambiguity (if, say, two neighbours have the same colour). The algorithm uses the Delaunay tesselation from DataDrivenDocuments.js (d3js) [Bostock, 2012]. This partitions our plane into polygon-regions, each of which includes boundaries at the furthest distance from each point (Voronoi

cells) [Watson, 1981]. First, we chose a random cell and assigned it a colour. Next, all its neighbours are recursively iterated, giving them the lowest possible colour in a list, which does not match any of their neighbours. Although such a greedy approach does not produce an optimum result, it allows for the colouring of data with ≤ 5 distinct colours, as is shown in Figure 6.13.



Figure 6.13: An example 4 colour matching This uses the first implementation of the algorithm mentioned in Subsection 6.4.4. The greedy approach does not often find the optimum solution, which may result in 5 colours instead. Observable Notebook : Daniel Ellis [2019]

Having defined all the visualisation techniques we can now move on to explain the clustering algorithms which are used, and how 'goodness of fit' may be measured in the clustering context.

6.5 Cluster Evaluation

The previous section discussed methods of visualising the reduced data for use with interactive exploratory data analysis. In this section, we look at the use of vector clustering algorithms¹² (Subsection 6.5.1) to highlight groups in a 2D dataset, as well an automated method of assessing the quality of the clusters selected (Subsubsection 6.5.1.1) and feature extraction (Subsection 6.5.2).

6.5.1 Automated Selection Of Clusters

When it comes to clustering data points in a dataset, there exists a range of methods which may accomplish a task, Figure 6.14. Most often, the k-means [MacQueen, 1967] is used as it is fast and straightforward to understand. However, its linear method of partitioning cannot capture the splits

 $^{^{12}}$ Vector clustering is the grouping of data based on their proximity or density to other nearby points

between non-linear relationships of real data. The other problem is that an estimate for the number of expected clusters is required - something that is often unknown without prior understanding of the data. When this is the case, often it is easier to select the nodes with interactivity manually.

In contrast, density-based clustering techniques such as GMM ([Pedregosa et al., 2011a]) or DBSCAN ([Ester et al., 1996]) tend to be better at locating non-linear trends in the data. The DBSCAN algorithm assesses the distribution of data across a specific location. This allows clusters with a high density of datapoints to be located without the need for a predefined number as an input. Another method: OPTICS (Ordering Points To Identify the Clustering Structure) [Ankerst et al., 1999], shall be used¹³. This is an adaptation of the DBSCAN algorithm which does not require the specification of a minimum distance between points (for the density estimate)- instead, we specify a gradient for the distribution and the minimum number of points for a cluster to be classified.



Figure 6.14: A comparison of different clustering methods on a toy dataset. The plot shows the performance of several vector clustering algorithms in Scikit-Learn. Cluster algorithms are represented across the horizontal axis, and several types of datasets are across the vertical. Clustered groups are coloured. Source: [sklearn, 2019]

When deciding which algorithms to use, each algorithms' ability to partition non-linear data is considered. The first two rows of Figure 6.14 show data which cannot be partitioned linearly, here spectral, DBSCAN and optics are the only clustering algorithms to identify both correctly. It is for this reason that we shall look at these for the remainder of the chapter.

 $^{^{13}}$ If using Python 2, the library for this needs to be extracted from the sci-kit-learn library for python3 package and altered to run with the previous version. (See copy in attached code.)

In selecting a value for the results section, several clustering algorithms, with a wide range of input parameters, are run. From these, the simulation with the best silhouette coefficient (Subsubsection 6.5.1.1) is taken.

6.5.1.1 Clustering (Silhouette) Coefficient

The silhouette measure is a tool used for accessing the validity of a set of clusters. Here each cluster is represented as a silhouette, based on the comparison of its tightness and separation. To calculate the silhoette coefficient we look at the intra-cluster a and the mean inter-cluster¹⁴ distance b. The silhouette cluster can then be described using [Rousseeuw, 1987; Pedregosa et al., 2011b]:

$$s(i) = \frac{b(i) - a(i)}{\max a(i), b(i)}$$
(6.6)

This gives a value $-1 \leq s(i) \leq 1$. Values near zero suggest overlapping clusters, 1 - dense, wellseparated clusters and negative values indicate that a sample may have been incorrectly classified. In using this method, we can get an overview of how well individual objects lie within their assigned cluster.

6.5.2 Feature Extraction

Upon establishing a set of DR datasets, and their groups (the clusters of species they contain), it is important to evaluate what input features they represent. Rather than doing this manually, we make use of Random Forests - described below.

6.5.2.1 Random Forests

Random forests [Breiman, 2001], are a subset of ML algorithms called ensemble learning. This means that they train a large number of decision trees, each on a random subset of the original features. A decision tree is a tree formed from a series of conditionals¹⁵, much like a perceptron network (Subsubsection 4.6.1.2) with binary activation functions. Random forests introduce a level of additional randomness by selecting only a subset on which to create each decision tree. This may introduce a higher bias, but lowers the overall model variance, which creates a better (more robust) model. Such methods have been applied to replacing the computationally expensive process of chemistry integration of GEOS-Chem (a global 3D model of tropospheric chemistry) [Keller and Evans, 2019] and the

¹⁴Inside and between different clusters.

 $^{^{15}\}mathrm{Questions}$ with a True/False answer

prediction of global sea-surface iodine based on observations coupled with sea-surface temperature, depth, and salinity [Sherwen et al., 2019].

6.5.2.2 Calculating Importance Using Random Forests

Since random forests are in essence a collection of decision trees, it is possible to generate a 'decision tree aggregate' to visualise the ensemble structure of the random forest [Ellis and Sherwen, 2019] (Figure 6.15). Alternatively, if all that is required is the relative importance of each feature, the RandomForestClassifier from [Pedregosa et al., 2011b] provides a quick and easy way of understanding which features matter, [Géron, 2017]. This works by aggregating the weighted nodes which use a particular feature using the number of samples and then scales the result to 1. We use this method to access the overall importance of features within each DR output and identify the differences between clusters.



Figure 6.15: A decision tree aggregate from a random forest plotted with the Epiphyte version of the TreeSurgeon program [Ellis and Sherwen, 2019]. The data originates from Sherwen et al. [2019] and the importance of Temperature (blue), Depth (orange) and Chlorophyll a (green). It is shown that all models create their first split based on the temperature (is it >21 degrees). In the case it is (right branch) the sea depth is seen as the most important variable to test (is it deeper than 26m). This sort of split allows us to get a feel for which (if any) properties are dominant in partitioning the data.

NOTE: The only downside is that Random Forests are in themselves ML techniques which also need to be evaluated. To do this, as they are simply being used as indicators of cluster properties which we are to explore further, we can initiate a collection of 300 random Forest classifiers, from which we take the median. A sort of ensemble learning from an ensemble.

6.6 Results

There exist many methods of defining the chemical structure of species within the MCM. This section evaluates the different structural representations (Section 6.2) and the ability of DR algorithms to separate the chemical space within which these lie into a two-dimensional scatterplot.

6.6.1 Visual Overview

Explorative data analysis involves a degree of figure interactivity. Chapter 2 described the importance of visualisation in employing the cognitive pattern functions of the human brain, and Chapter 3 explained the importance of having an evenly distributed data points to aid the understanding of graphs. This subsection combines the two ideas in the usage of dimensionality reduction to exploit patterns within a dataset.

Using the techniques in Section 6.4, we explore the visual distribution of different dimensionality reduction methods across all input types. This subsection explores the spatial distribution of groups (blobs) in the 2D dimensionally reduced dataset. The colours represent the automatically calculated clusters (which are further explored analytically in Subsection 6.6.2). This is then built upon using three case studies, where individual cluster distributions are compared (Subsection 6.6.3).

An autoencoder will produce near-identical results to the PCA algorithm, using linear activation functions. Them ain difference is that rather than discarding components which represent little variance, the neural network (autoencoder) combines their values when deciding on the output. Although ordinal data (Figure 6.17 b,f) still produce regular patterns, the non-linear (tanh) activation functions result in a greater separation between data points for the SMILES dataset (Figure 6.17g).

Unlike PCA and AE, t-SNE does not contain any inherent meaning behind the spatial positioning of points. Instead, it provides a non-linear grouping of points through a graph-like force-directed model (all items are connected with the weights of the links decided by their relationship values). This results in the most visually pleasing output die to clusters increased separation. This property also makes it the easiest to visually isolate clusters from their neighbours (Figure 6.18), making it a useful tool for interactive data exploration or explaining groups within a figure.



Figure 6.16: Comparing clusters for all inputs after a reduction to 2 dimensions using Principal Component analysis. Each graph has undergone several clustering algorithms under a range of parameters. The result with the best silhouette coefficient has been chosen. Colours follow the greedy four colour theorem and are there only to indicate the contrast between cluster boundaries.



Figure 6.17: Comparing clusters for all inputs after a reduction to 2 dimensions using an AutoEncoder. Each graph has undergone several clustering algorithms under a range of parameters. The result with the best silhouette coefficient has been chosen. Colours follow the greedy four colour theorem and are there only to indicate the contrast between cluster boundaries.



Figure 6.18: Comparing clusters for all inputs after a reduction to 2 dimensions using t-SNE. Each graph has undergone several clustering algorithms under a range of parameters. The result with the best silhouette coefficient has been chosen. Colours follow the greedy four colour theorem and are there only to indicate the contrast between cluster boundaries.

6.6.2 Mathematical Cluster Analysis

Subsection 6.6.1 explored the visual appeal of the plotted clusters. For large systems, the selection of many clustering algorithm outputs is impractical, so the analysis of DR methods has been automated (Subsection 6.5.1). Similarly, we can once again apply the silhouette coefficient (Subsubsection 6.5.1.1) to compare the best output for each input and DR algorithm.

Outputs for the number of groups the DR output has been clustered in and its corresponding silhouette coefficient are shown in Tables 6.2 - 6.4. These shall be discussed by name rather than being referenced each time for ease in reading. Inputs for each algorithm are ranked in order of their silhouette coefficient (the closer to 1 the value, the better the clustering).

Ordinal inputs such as functional groups or the protocol categories consistently rank the highest within each algorithm. This is because the algorithms only have to identify the permutations of each category and classify the species into these. It is noted that the t-SNE silhouette coefficient for these is 30% lower than for PCA and AE algorithms. However, the number of groups located is also greatly reduced from 140 to 106. This suggests that the vector clustering algorithms have tried to combine data points into a group, which has come at a cost to the silhouette value.

Next, the SMILES strings are ranked highly for both AE and t-SNE algorithms. Visually this agrees with Figure 6.17(g) and Figure 6.18(g), where these are much better separated than Figure 6.16(g). The PCA ranking is almost half of other algorithms, but contains a much smaller number of 'clusters'. This can be attributed to its periodic lattice-like spacing of points which are not conducive to producing good vector clustering groups in an unsupervised algorithm.

MACCS keys, although providing a modest silhouette score across most DR algorithms, often only have two or three clusters. The reason for this is that dimensionality reduction of these often separate species into two groups, those with Nitrogen elements and those without (Figure 6.19). This is because many of the questions on which the MACCS fingerprint is based concern themselves with Nitrogen species [rdkit, 2019]. Group composition and cluster analysis is discussed in Subsection 6.6.3 and Subsection 6.6.4.



Figure 6.19: The individual decomposition of clusters in Figure 6.16(c) This shows that the main difference between the two clusters is the existence of Nitrogen elements within Nitrate and Peroxyacetyl Nitrate (PAN) groups. The table on the right acts as a key for the colours and shows the overall importance of each feature in separating an item into the various clusters (using an ensemble of decision trees).

Similar to SMILES strings molecular quantum numbers result in plots consisting of regular rows of data with like-properties (Figure 6.16-6.18(d)) - making it difficult for the clustering algorithms to select each group correctly. This property makes it suitable for usage in the field of chemical informatics where molecules with similar properties are desired [Arús-Pous et al., 2019], but less so for establishing an overarching categorisation of species within a mechanism.

Finally, the graph-based fingerprint input is consistently the lowest scoring in the silhouette coefficient. Visually this can be attributed to the mismatching of cluster number to the number of visually separate clusters (Figure 6.16-6.18(a)). The random distribution of these visual 'specs' suggests that the custom graph-fingerprint is not the most informative input for use with DR and clustering algorithms.

Overall combining the visual representation of selected clusters (Subsection 6.6.1) with the silhouette scores, show that the clustering algorithms struggle with the prediction of correct groups. Although the silhouette coefficient is a useful metric for determining the density of points around a cluster, it should be used with the aid of other metrics if selecting the best results from an automated clustering process. Additionally, the automatic clustering process can include feedback from the goodness of fit metrics, which can allow for the finer tuning of clustering algorithm parameters in future.

DR	input	silhouette	groups
PCA	fngroups	0.9122	141
PCA	protocol	0.8761	149
PCA	node2vec	0.8569	3
PCA	maccs	0.6563	2
PCA	mqn	0.4041	8
PCA	smiles	0.3648	6
PCA	fingerprints	0.3529	6
PCA	spec	0.3364	6

Table 6.2: The inputs to the PCA dimensionality reduction algorithm sorted by the best obtained silhoette coefficient.

DR	input	silhouette	groups
AE	fngroups	0.9251	140
AE	protocol	0.9095	28
AE	maccs	0.6671	3
AE	smiles	0.6313	6
AE	node2vec	0.6253	2
AE	mqn	0.6096	2
AE	spec	0.6062	3
AE	fingerprints	0.4145	4

Table 6.3: The inputs to the AutoEncoder dimensionality reduction algorithm sorted by the best obtained silhoette coefficient.

DR	input	silhouette	groups
t-SNE	fngroups	0.7524	105
t-SNE	protocol	0.6012	74
t-SNE	smiles	0.5360	16
t-SNE	maccs	0.4418	3
t-SNE	node2vec	0.4359	6
t-SNE	spec	0.3781	35
t-SNE	mqn	0.3684	8
t-SNE	fingerprints	0.3539	6

Table 6.4: The inputs to the t-SNE dimensionality reduction algorithm sorted by the best obtained silhoette coefficient.

6.6.3 Feature Selection Comparison

The previous subsection assessed how well DR algorithms were able to separate the chemistry of a mechanism into distinct, well-defined clusters. Now the content of each of those groups is looked at, comparing them to the functional groups most responsible for the variation within the 2D compression of a chemical mechanism. Importance of each functional group in explaining cluster composition is obtained using a random forest classifier (Figures 6.20-6.22).

Comparing all three DR algorithms, we see that the shape (and this group importance) is persistent between each input across all algorithms. This means that although values may differ, the same functional groups are important between each DR algorithm - indicating that this is a property of the input style and not the type of dimensionality reduction technique selected.

In establishing that the input format is responsible for the splitting of clusters into groups, we look at what the relationship of these is. In this section, the characters (a-h) refer to **all** the corresponding subplots in Figures 6.20-6.22.

In the Common Representative Intermediates (CRI) mechanism, the ratio of C-H and C-O bonds are used to lump species with the same oxidation capacity [Jenkin et al., 2008]. This makes the ratio of carbons to oxygens an essential defining characteristic between them. The number of Oxygens and Carbons is a consistently important feature for all input styles (with the exception of the gecko protocol categories (f)). Here the number of carbons or oxygens does not fit any of the reaction branches meaning that there is no way for the dimensionality reduction algorithm to know these. Instead, Alcohol and Carbonyl groups are seen as the most important in separating the chemistry, which may be due to the number of species undergoing each type of reaction.

Species names (h) have prefixes (e.g. Cxx) and suffices (e.g., -OOH, -NO₃, -O₂, -OL) which allow an easy way for a user to distinguish the types of species, but also the DR algorithms. These show the second-best separation of aromatic species, most likely attributed to the standard naming convention, e.g. ('BENZ', 'PIN', 'TMB'). Similarly, the SMILES string (g) shows a human-readable representation of the structure of the molecule. Since this is explicitly defined, the SMILES input provides the highest uniformity in group types when analysing cluster composition. As aromatic compounds are represented using a lower case 'c', this makes them easy to distinguish (especially in the case of AE Figure 6.21g).

As was touched on in Subsection 6.6.2 the MACCS input consists of a series of logical questions about a species structure. Since many of those questions regard the existence of a Nitrogen atom, data was separated species with a Nitrate or PAN group, and those without. In making a series of decisions on which cluster a species falls under, this largest most recurring branch for the RandomForestClassifier (imagine of temperature in Figure 6.15) falls under the existence of a Nitrate group.

The main inconsistency between clusters and DR algorithms comes from the Node2Vec embedding (e) - much of which can be explained by the poor performance of the DR and clustering algorithms of separating the chemistry into groups (see plots in Subsection 6.6.1). Subsection 6.6.4 continues this analysis by comparing output with < 3 clusters each against the graph plots presented in this subsection. The content of individual groupings is explored for an output with multiple clusters.



Figure 6.20: Comparing feature importance for PCA clusters. Importance ranges are trimmed at 40% for comparison. Some categories may contain values greater than this. All bars sum to 100%.



Figure 6.21: Comparing feature importance for AE clusters. Importance ranges are trimmed at 40% for comparison. Some categories may contain values greater than this. All bars sum to 100%.



Figure 6.22: Comparing feature importance for t-SNE clusters. Importance ranges are trimmed at 40% for comparison. Some categories may contain values greater than this. All bars sum to 100%.

6.6.4 Cluster Comparison

In this final subsection, we look at the composition of different clusters within the dimensionally reduced dataset. We begin by looking at the simplest cells from Subsection 6.6.1 - ones which only contain two or three cluster groups (Subsection 6.6.5) and then move on to explore three examples showing multiple clusters (Subsubsection 6.6.5.1).

6.6.5 Bi / Tri Cluster Groups

Using the DR output where only two/three groups are located by the clustering algorithms we have (Figure 6.23 and Figure 6.24). In exploring the MACCS key input for the PCA and t-SNE DR algorithms (Figure 6.23) we find that for the cumulative importance bar charts we know that the existence of Nitrates is vital in the split determining which group a species falls into. This manifests itself as having a single cluster containing PAN and Nitrate species, with others not. In the t-SNE plot (Figure 6.23b) we see that there exists a third group which is missing both Aldehyde and PAN functionalisation for each species. This is shown by the teal colour in Figure 6.18c and resides between the Nitrogen-containing and Nitrogen-deficient groups.

Figure 6.24 shows the comparison of the Node2Vec embedding using PCA and the AE DR algorithms. In Figure 6.16e and Figure 6.17e, it is seen that these are generally not separated into well-partitioned clusters. Both groups consist of one large cluster (shown by the second bar chart of each row which contains all functional groups) and one or two fragment ones. In exploring the AE plot (Figure 6.24b), it is seen that as part of the cumulative plot (right), the -OOH functional group is an important separatory factor since the smaller of the two groups does not contain any species which contain a hydroperoxy functional group. In the PCA plot, although providing different cumulative results, again shows species within the smaller groups not containing any RO, RCO₃, OOH, ONO₂ or OOH functional groups. This can potentially be due to the graph structure, where the random walker (which generates the Node2Vec embedding) has become trapped by a group of non-oxidised species.



Figure 6.23: Comparing individual clusters between MACCS for PCA and t-SNE algorithm output. The bar chart to the right is the cumelative chart which represents the splits in deciding the cluster a species falls into from Subsection 6.6.3. Unlabeled bar charts to the left represent the partitioning of species within an individual cluster.



(b) AE

Figure 6.24: Comparing individual clusters between Node2Vec for PCA and t-SNE algorithm output. The bar chart to the right is the cumulative chart which represents the splits in deciding the cluster a species falls into from Subsection 6.6.3. Unlabeled bar charts to the left represent the partitioning of species within an individual cluster.
6.6.5.1 Multicluster Groups

Next, we observe several multi clustered examples from all DR algorithms. We start with Figure 6.25 where the PCA algorithm has generated an extensive collection of points in an area. Unsurprisingly the clustering algorithm has failed to identify separate groups (as there is only one), and instead partitioned the data into six groups. Although not ideal, this still has its use in determining how species have been partitioned temporally. Here we find that nitrogen-containing species are positioned on the right side of the graph, and aromatic species are in the bottom half. RO₂ containing species span the entirety of Figure 6.25, but increase towards the left direction. This shows that although PCA was unable to separate the chemical species of the MCM into groups from the fingerprint input, it still presents patterns within the data such that the arrangement of groups can be seen when querying specific areas.



Figure 6.25: Case Study 1: PCA graph-fingerprint We compare the functional group distribution for individual clusters within the PCA 2D representation of the graph-fingerprint input.

Next, we explore the AE algorithm. Here the smiles input separates the data into more defined groups. This becomes apparent with specific functional groups only appearing within individual clusters (Figure 6.26) - as opposed to the gradients seen in Figure 6.25. Here all the aromatic species are contained within the central (pink cluster). Similarly, the green and brown clusters do not contain any nitrates, and most of the PAN species are within the grey cluster in the top right. The better separation of clusters aids in the identification of groups (by the automated vector clustering algorithm) as well as highlighting the process undergone within the DR algorithm to partition the data.



Figure 6.26: Case Study 1: AE SMILES We compare the functional group distribution for individual clusters within the AE 2D representation of the SMILES input.

261

Finally, we explore the t-SNE DR algorithm using the MQN inputs (Figure 6.27). Previously it is found that the t-SNE produces the most well-defined clusters, the cost of which comes from losing information about group similarity encoded in the distance separating them. Here RCO_3 containing species are located in the central turquoise cluster, PAN species in the top right (lime) cluster and the aromatic species are all in the pink cluster at the bottom. The brown and purple clusters (top left) contain no carbonyl groups, and the purple, grey (top left) and orange (mid-right) contain not Nitrates. The t-SNE provides the best spatial distribution of groups. However, inspecting cluster colours visually suggests that the automatical vector clustering algorithm has not necessarily located the best combination of groups. As was suggested in Subsection 6.6.2 a more dynamic method of tuning the clustering algorithm hyperparameters may result in better cluster selection - which may better separate each cluster's species category.



Figure 6.27: Case Study 1: t-SNE MQN. We compare the functional group distribution for individual clusters within the t-SNE 2D representation of the Mollecular Quantum Number fingerprint.

6.7 Conclusions

This chapter aims to tie up the research presented in Chapter 2-5 in preparation for future using Graph Convoluted Neural Networks [Kipf and Welling, 2016] to classify and even predict (generate) mechanisms (or at least attempt to). In Chapter 3 we showed that networks are a useful representation for the relational nature of species within the atmosphere, and then applied several mathematical techniques to it Chapter 4-5. This chapter looked at simplifying chemical structure used by many dimensionality reduction algorithms and using visualisation and computational algorithms to assess their ability at partitioning species into similar groups.

It was found that t-Distributed Stochastic Neighbor Embedding (a graph-based DR technique) provided the best results for species separation and visualisation. This provides a non-linear mapping of the relationships between items and does not omit any data (PCA) or require the selection of layers and activation functions (AE). It does, however, lose information about cluster similarity based on distance and cannot be used to compress (encode/decode) information - although neither of those features is required for our use.

Additionally, several possible inputs for each dimensionality reduction algorithm were used. Since machines cannot understand the meaning of words, these are different representations of species structure we can put into a machine-learning algorithm to inform it of a species. Out of the non-ordinal inputs, it was found that tokenised SMILES strings and the Molecular Quantum Number fingerprint produced 2D visualisations with the best separation between clusters. Other inputs either caused large groups of overlapping nodes or had a strong dependency on a single species or functional group (e.g. MACCS and Nitrates).

It is suggested that within future work, the t-SNE dimensionality reduction algorithm is used if trying to visualise the different groups of a complex dataset. In the case of any machine learning algorithms, the MQN fingerprint should be used for generic examples, unless a specific feature is required from the input. If instead a more complex/larger input is required, it is also possible to apply any of the DR algorithms discussed to simplify it. Here the use of an AutoEncoder is suggested as the encoderdecoder pair allows for the testing of a trained model (as well as the ability to explore the embedded space). Although this can be achieved through the use of principal component analysis, this does not handle non-linear relationships and cannot be trained on a further dataset if one becomes available.

Bibliography

- Alon, U., Zilberstein, M., Levy, O., and Yahav, E. (2019). Code2Vec: Learning Distributed Representations Of Code. http://dl.acm.org/citation.cfm?doid=3302515.3290353.
- Anderson, C. (2008). The End Of Theory: The Data Deluge Makes The Scientific Method Obsolete. online. http://www.wired.com/print/science/discoveries/magazine/16-0.
- Ankerst, M., Breunig, M. M., Kriegel, H.-P., and Sander, J. (1999). Optics: Ordering points to identify the clustering structure. SIGMOD Rec., 28(2):49–60. https://doi.org/10.1145/304181.304187.
- Appel, K. and Haken, W. (1976). Every planar map is four colorable. Bull. Amer. Math. Soc., 82(5):711–712. https://projecteuclid.org:443/euclid.bams/1183538218.
- Arús-Pous, J., Blaschke, T., Ulander, S., Reymond, J.-L., Chen, H., and Engkvist, O. (2019). Exploring The Gdb-13 Chemical Space Using Deep Generative Models. *Journal of cheminformatics*, 11(1):20. http://dx.doi.org/10.1186/s13321-019-0341-z.
- Aumont, B., Szopa, S., and Madronich, S. (2005). Modelling the evolution of organic carbon during its gas-phase tropospheric oxidation: Development of an explicit model based on a self generating approach. Atmospheric Chemistry and Physics, 5(9):2497–2517. https://www.atmos-chem-phys. net/5/2497/2005/.
- Baillargeon, R. and Carey, S. (2012). Core cognition and beyond: The acquisition of physical and numerical knowledge. *Early childhood development and later outcome*.
- Bostock, M. (2012). D3.js data-driven documents. http://d3js.org/.
- Box, G. E. P. (1976). Science And Statistics. Journal of the American Statistical Association, 71(356):791–799. https://www.tandfonline.com/doi/abs/10.1080/01621459.1976.10480949.
- Breiman, L. (2001). Random Forests. *Machine learning*, 45(1):5–32. https://doi.org/10.1023/A: 1010933404324.
- Cohen, E. (2018). Node2Vec: Embeddings For Graph Data. https://towardsdatascience.com/ node2vec-embeddings-for-graph-data-32a866340fef.
- Daniel Ellis (2019). D3-Fourcolour Voronoi. https://observablehq.com/@wolfiex/ d3-fourcolour-voronoi.
- Dataman (2019). Convolutional Autoencoders For Image Noise Reduction. https://towardsdatascience.com/convolutional-autoencoders-for-image-noise-reduction-32fce9fc1763.

- Descartes, R. and Lafleur, L. J. (1960). *Meditations On First Philosophy*. Bobbs-Merrill New York. http://selfpace.uconn.edu/class/percep/DescartesMeditations.pdf.
- Du, J., Jia, P., Dai, Y., Tao, C., Zhao, Z., and Zhi, D. (2019). Gene2Vec: Distributed Representation Of Genes Based On Co-Expression. *BMC genomics*, 20(Suppl 1):82. http://dx.doi.org/10.1186/ s12864-018-5370-x.
- Durant, J. L., Leland, B. A., Henry, D. R., and Nourse, J. G. (2002). Reoptimization Of Mdl Keys For Use In Drug Discovery. Journal of chemical information and computer sciences, 42(6):1273–1280. https://www.ncbi.nlm.nih.gov/pubmed/12444722.
- Ellis, D. (2019). Chemical Kinetic Interactions Cover Image. https://s100.copyright.com/ AppDispatchServlet?startPage=i&publisherName=Wiley&publication=kin&contentID=10.1002% 2Fkin.21180&endPage=i&title=Cover+Image%2C+Volume+50%2C+Issue+6.
- Ellis, D. and Sherwen, T. (2019). Wolfiex/treesurgeon: Wollemia. https://doi.org/10.5281/zenodo. 3346817.
- Ester, M., peter Kriegel, H., Sander, J., and Xu, X. (1996). A density-based algorithm for discovering clusters in large spatial databases with noise. *AAAI Press*, pages 226–231.
- F.R.S., K. P. (1901). Liii. on lines and planes of closest fit to systems of points in space. The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science, 2(11):559–572. https://doi.org/10.1080/14786440109462720.
- Géron, A. (2017). Hands-On Machine Learning With Scikit-Learn And Tensorflow: Concepts, Tools, And Techniques To Build Intelligent Systems. O'Reilly Media. https://books.google.co.uk/books? id=khpYDgAAQBAJ.
- Grover, A. and Leskovec, J. (2019). Node2vec: Scalable feature learning for networks. *online*. Accessed: 2019-10-21.
- Hamadache, M. and Lee, D. (2017). Principal Component Analysis Based Signal-To-Noise Ratio Improvement For Inchoate Faulty Signals: Application To Ball Bearing Fault Detection. International journal of control, automation, and systems, 15(2):506–517. https://doi.org/10.1007/ s12555-015-0196-7.
- Heller, S., McNaught, A., Stein, S., Tchekhovskoi, D., and Pletnev (2013). Inchi The Worldwide Chemical Structure Identifier Standard. *Journal of cheminformatics*, 5(1):7. http://dx.doi.org/10. 1186/1758-2946-5-7.

- Hernandez, W. and Mendez, А. (2018).Application Of Principal Component Analysis To Image Compression. Göksel, Т., Grow-In editor, *Statistics* Sets Growing Demand Statistics. InTech. http://www. inqDataandforintechopen.com/books/statistics-growing-data-sets-and-growing-demand-for-statistics/ application-of-principal-component-analysis-to-image-compression.
- Hotelling, H. (1933). Analysis of a complex of statistical variables into principal components. Journal of Educational Psychology, 24(6):417–441. https://doi.org/10.1037%2Fh0071325.
- Indyk, P. and Motwani, R. (1998). Approximate nearest neighbors: Towards removing the curse of dimensionality. In *Proceedings of the Thirtieth Annual ACM Symposium on Theory of Computing*, STOC '98, page 604–613, New York, NY, USA. Association for Computing Machinery. https: //doi.org/10.1145/276698.276876.
- Jean, N., Wang, S., Samar, A., Azzari, G., Lobell, D., and Ermon, S. (2018). Tile2Vec: Unsupervised Representation Learning For Spatially Distributed Data. online. http://arxiv.org/abs/1805.02855.
- Jenkin, M., Watson, L., Utembe, S., and Shallcross, D. (2008). A common representative intermediates (cri) mechanism for voc degradation. part 1: Gas phase mechanism development. Atmospheric Environment, 42(31):7185 – 7195. http://www.sciencedirect.com/science/article/pii/ S1352231008006742.
- Jolliffe, I. T. and Cadima, J. (2016). Principal Component Analysis: A Review And Recent Developments. Philosophical transactions. Series A, Mathematical, physical, and engineering sciences, 374(2065):20150202. http://dx.doi.org/10.1098/rsta.2015.0202.
- Jones, E., Oliphant, T., Peterson, P., et al. (2001–). Scipy: Open source scientific tools for Python. http://www.scipy.org/.
- Keller, C. A. and Evans, M. J. (2019). Application of random forest regression to the calculation of gas-phase chemistry within the geos-chem chemistry model v10. *Geoscientific Model Development*, 12(3):1209–1225. https://www.geosci-model-dev.net/12/1209/2019/.
- Kipf, T. N. and Welling, M. (2016). Semi-supervised classification with graph convolutional networks. CoRR, abs/1609.02907. http://arxiv.org/abs/1609.02907.
- Krizhevsky, A., Sutskever, I., and Hinton, G. E. (2012). Imagenet Classification With Deep Convolutional Neural Networks. In Pereira, F., Burges, C. J. C., Bottou, L., and Weinberger, K. Q., editors, Advances in Neural Information Processing Systems 25, pages 1097–1105. Curran Associates, Inc. http://papers.nips.cc/paper/ 4824-imagenet-classification-with-deep-convolutional-neural-networks.pdf.

- Landrum, G., Tosco, P., Kelley, B., sriniker, gedeck, NadineSchneider, Vianello, R., Dalke, A., Cole, B., AlexanderSavelyev, Turk, S., Ric, Swain, M., Vaucher, A., N, D., Wójcikowski, M., Pahl, A., JP, strets123, JLVarjo, O'Boyle, N., Berenger, F., Fuller, P., Jensen, J. H., Sforna, G., DoliathGavid, Cosgrove, D., Nowotka, M., Leswing, K., and van Santen, J. (2019). Rdkit 2019-03-2 (q1 2019) release. https://doi.org/10.5281/zenodo.2864247.
- Leite, N. M. N., Pereira, E. T., Gurjão, E. C., and Veloso, L. R. (2018). Deep convolutional autoencoder for eeg noise filtering. In 2018 IEEE International Conference on Bioinformatics and Biomedicine (BIBM), pages 2605–2612.
- Lynch, H. (2011). Infant Places, Spaces And Objects: Exploring The Physical In Learning Environments For Infants Under Two. PhD thesis, -. http://dx.doi.org/10.21427/D73W37.
- Maaten, L. v. d. and Hinton, G. (2008). Visualizing Data Using T-Sne. Journal of machine learning research: JMLR, 9(Nov):2579–2605. http://www.jmlr.org/papers/v9/vandermaaten08a.html.
- MacQueen, J. (1967). Some methods for classification and analysis of multivariate observations. In Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability, Volume 1: Statistics, pages 281–297, Berkeley, Calif. University of California Press. https://projecteuclid.org/ euclid.bsmsp/1200512992.
- MDL (1984). Maccs-ii. MDL Information SystemsSymyx (MDL).
- Mikolov, T., Chen, K., Corrado, G., and Dean, J. (2013). Efficient Estimation Of Word Representations In Vector Space. *online*. http://arxiv.org/abs/1301.3781.
- Morozov, A. (2016). Modelling biological evolution: Linking mathematical theories with empirical realities. *Journal of Theoretical Biology*, 405:1 4. http://www.sciencedirect.com/science/article/pii/S0022519316301849.
- Narayanan, A., Chandramohan, M., Venkatesan, R., Chen, L., Liu, Y., and Jaiswal, S. (2017). Graph2vec: Learning distributed representations of graphs. *CoRR*, abs/1707.05005. http://arxiv. org/abs/1707.05005.
- Nguyen, K. T., Blum, L. C., van Deursen, R., and Reymond, J.-L. (2009). Classification Of Organic Molecules By Molecular Quantum Numbers. *ChemMedChem*, 4(11):1803–1805. http://dx.doi.org/ 10.1002/cmdc.200900317.
- Noble, C. E. (1957). Human Trial-And-Error Learning. *Psychological reports*, 3(2):377–398. https://doi.org/10.2466/pr0.1957.3.h.377.
- Oliphant, T. (2006). Guide to numpy. https://docs.scipy.org/doc/_static/numpybook.pdf.

- Oliveira, B., Pereira, F., de AraÃojo, R., and Ramos, M. (2006). The hydrogen bond strength: New proposals to evaluate the intermolecular interaction using dft calculations and the aim theory. *Chemical Physics Letters*, 427(1):181 – 184. http://www.sciencedirect.com/science/article/pii/ S000926140600861X.
- Parsons, J., Holmes, J. B., Rojas, J. M., Tsai, J., and Strauss, C. E. M. (2005). Practical Conversion From Torsion Space To Cartesian Space For In Silico Protein Synthesis. *Journal of computational chemistry*, 26(10):1063–1068. http://dx.doi.org/10.1002/jcc.20237.
- Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., and Duchesnay, E. (2011a). Scikit-learn: Machine learning in python. J. Mach. Learn. Res., 12:2825–2830. http://dl.acm.org/citation.cfm?id=1953048.2078195.
- Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., and Duchesnay, E. (2011b). Scikit-Learn: Machine Learning In Python . Journal of Machine Learning Research, 12:2825–2830.
- People2Vec (2019). People2Vec. http://people2vec.org/.
- Powell, V. (2020). Principal Component Analysis Explained Visually. http://setosa.io/ev/principal-component-analysis/.
- Probst, D. and Reymond, J.-L. (2018). Smilesdrawer: Parsing And Drawing Smiles-Encoded Molecular Structures Using Client-Side Javascript. Journal of chemical information and modeling, 58(1):1–7. http://dx.doi.org/10.1021/acs.jcim.7b00425.
- rdkit (2019). Rdkit. https://github.com/rdkit/rdkit/blob/24f1737839c9302489cadc473d8d9196ad9187b4/ rdkit/Chem/MACCSkeys.py.
- Reinhardt, A. (1975). Art-As-Art: The Selected Writings Of Ad Reinhardt. Documents of 20th-century art. Viking Press. https://books.google.co.uk/books?id=zyK4AAAAIAAJ.
- Roberts, R. (1989). Serendipity: Accidental Discoveries In Science. Wiley Science Editions. Wiley. https://books.google.co.uk/books?id=hf57X0s4aPwC.
- Rousseeuw, P. J. (1987). Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. Journal of Computational and Applied Mathematics, 20:53–65. http://www.sciencedirect. com/science/article/pii/0377042787901257.
- Roweis, S. T. and Saul, L. K. (2000). Nonlinear dimensionality reduction by locally linear embedding. Science, 290(5500):2323–2326. https://science.sciencemag.org/content/290/5500/2323.

- Sherwen, T., Chance, R. J., Tinel, L., Ellis, D., Evans, M. J., and Carpenter, L. J. (2019). A machinelearning-based global sea-surface iodide distribution. *Earth System Science Data*, 11(3):1239–1262. https://www.earth-syst-sci-data.net/11/1239/2019/.
- sklearn (2019). Comparing Different Clustering Algorithms On Toy Datasets Scikit-Learn 0.21.3 Documentation. https://scikit-learn.org/stable/auto_examples/cluster/plot_cluster_comparison. html.
- Spahn, V., Del Vecchio, G., Labuz, D., Rodriguez-Gaztelumendi, A., Massaly, N., Temp, J., Durmaz, V., Sabri, P., Reidelbach, M., Machelska, H., Weber, M., and Stein, C. (2017). A Nontoxic Pain Killer Designed By Modeling Of Pathological Receptor Conformations. *Science*, 355(6328):966–969. http://dx.doi.org/10.1126/science.aai8636.
- T. Leube, B., Inglis, K., J. Carrington, E., and Sharp, P. (2018). Lithium transport in li 4.4 m 0.4 m 0.6 s 4 (m = al 3+ , ga 3+ and m Â= ge 4+ , sn 4+): Combined crystallographic, conductivity, solid state nmr and computational studies. *Chemistry of Materials*, 30.
- Turanyi, T. and Tomlin, A. (2015). Analysis Of Kinetic Reaction Mechanisms. Springer. http: //eprints.whiterose.ac.uk/84294/.
- Wang, S.-G. and Schwarz, W. H. E. (2009). Icon Of Chemistry: The Periodic System Of Chemical Elements In The New Century. Angewandte Chemie, 48(19):3404–3415. http://dx.doi.org/10.1002/ anie.200800827.
- Watson, D. F. (1981). Computing The N-Dimensional Delaunay Tessellation With Application To Voronoi Polytopes*. The Computer Journal, 24(2):167–172. https://doi.org/10.1093/comjnl/24.2. 167.
- Weininger, D. (1988). Smiles, A Chemical Language And Information System. 1. Introduction To Methodology And Encoding Rules. Journal of chemical information and computer sciences, 28(1):31–36. https://pubs.acs.org/doi/abs/10.1021/ci00057a005.
- Wyche, K. P., Monks, P. S., Smallbone, K. L., Hamilton, J. F., Alfarra, M. R., Rickard, A. R., McFiggans, G. B., Jenkin, M. E., Bloss, W. J., Ryan, A. C., Hewitt, C. N., and MacKenzie, A. R. (2015). Mapping gas-phase organic reactivity and concomitant secondary organic aerosol formation: Chemometric dimension reduction techniques for the deconvolution of complex atmospheric data sets. Atmospheric Chemistry and Physics, 15(14):8077–8100. https://www.atmos-chem-phys.net/ 15/8077/2015/.
- Yu-ChenLo (2018). Machine learning in chemoinformatics and drug discovery. Drug Discovery Today, 23(8):1538 – 1546. http://www.sciencedirect.com/science/article/pii/S1359644617304695.

Chapter 7

Conclusions and Future Work

"So Long, and Thanks for All the Fish"

- Douglas Adams, HitchHikers Guide to the Galaxy - Part 4

7.1 Conclusions

The topic of changing climate has been a prominent talking point in the last decade [IPCC, 2013]. Anthropogenic activities starting with the industrial revolution have served to increase the amount of heat retained by the earth (radiative forcing). Similarly, the use of CFCs has damaged the protective layer of ozone in the atmosphere [Bais et al., 2018], and dangerous levels of air quality have produced an increase in respiratory distress of living organisms. Although we have guidelines and policies to determine the acceptable levels of pollutants, it is not uncommon for these to be unregulated or broken - for example, more than 95% of the EU urban population were exposed to concentrations higher than the WHO regulations in [EEA, 2018].

To prevent further irreversible harm, both physical and environmental, we must mitigate any further damage. The problem is, however, that this is not as small of a feat in itself. Taking the production of ground-level ozone as an example, the complex interplay between emissions and production within the earth system may produce two scenarios where the identical concentrations of a chemical species can result in the production or the loss of the secondary pollutant based entirely on the chemical regime it is in. For example, Fitzky et al. [2019] discusses the role of urban vegetation, and how trees can both reduce [Hardin and Jensen, 2007] and greatly contribute to the formation [Jenkin et al., 2015] of ozone - all while the shading provided from tree canopies could also influence the amount of radiation and its production [Yli-Pelkonen et al., 2018]. As we try to better represent the processes that govern the physical world, the larger and more complex our models become. This means an important balance must be struck, whereupon the 'selected'¹ tool must be both robust, accurate within reason and computationally efficient.

With the development of construction protocols, the automatic generation of very large and comprehensive chemical mechanisms is now a possibility, [Aumont et al., 2014]. This, however, presents problems which are both cognitive and computational. This thesis has explored the use of modern techniques in visualisation, reduction and machine learning in an attempt to address the above problem.

7.2 Results

In attempting to optimise the information transfer between computational models and the reader, we start by understanding human evolution and how an increase in neocortex size led to the ability (and necessity) to communicate large numbers between many people. The inial method for doing this was through the use of language, storytelling and pitograms. This sort of external information 'sharing' proved pivotal for the propagation of ideas, and ultimately the creation of technology and scientific advancement. Similarly, it is seen that even in the present-day setting, the use of narrative and selected metaphors can enrich the user's ability to navigate data, and instil a personal aspect to which they can relate to. When applied to atmospheric chemical mechanisms, the resultant output is a node-link representation of reactions, where species are analogous to people or items, and reactions (relationships) the links holding them together.

¹In atmospheric chemistry and climate change it is common to compare the results of several different models and mechanisms when studying something new. This style of 'ensemble' modelling provides a good way to check that things are working whilst eliminating some of the errors presented by individual simulations.

We then encode additional information into the visualisation, first syntactically using element design and colour, then semantically by setting the node positions and line lengths based on an additional property. The latter of which was achieved by a simple physical system similar to treating nodes as like charged magnets (they repulse) and links as springs pulling them back together. This force-directed graph structure (a subset of the sociograph class) alleviates many of the traditional difficulties of manually outlining a species degradation pathways. The push-pull physics nature of force-directed graphs makes them effortless to understand while allowing for additional information (such as the rate of reaction) to be embedded within the network shape all whilst being able to juxtapose different subsets or mechanisms within the same visual space (e.g. Figure 2.16).

At the limits of perceivable (visual acuity) and physical resolution, we were able to translate the graphical network structure into a purely computational one. Here we are able to perform temporal analysis of the state of a mechanism within a simulation by taking a series of static 'snapshots' or aggregating the data. This mathematical approach not only gives us information on the number of reactions of a species but also its importance within the system. Similarly in looking at where the flow of information within the network we can determine bottlenecks and controlling points, whereupon a small change to a one chemical species can have a significant effect on a large number of others. This type of analysis helps us to identify important areas to study, especially in the context of policy and air quality studies.

The computational graph can be further leveraged to categorise the type of network a mechanism represents, For example, we see that the Master Chemical Mechanism presents a sparse structure with the many highly connected (small-world) and hierarchical features. This is a pattern commonly found in real-world graphs and other chemical mechanisms, [Watts and Strogatz, 1998; Jacob and Lapkin, 2018].

The classification and ranking of species their modular structure allow us to apply several graph-based clustering techniques. Rather looking at the proximity and distribution of data within space, these techniques often navigate the links of a network, locating areas of high connectivity between species - thus forming a clique, module or cluster (depending on the field of study). This form of analysis not only highlights structural patterns from the network shape (e.g. Figure 5.5) but can also be used to access the suitability of combined species within mechanism lumping.

Finally, in preparation for future research, the use of different species structure representations was run through a selection of dimensionality reduction algorithms. Here different representations were reduced to two dimensions and shown in a x - y scatterplot. The analysis showed t-SNE reduced data was the most aesthetically pleasing, as it provides better separation between clustered groups. Additionally, the type of representation had a significant effect on the type of features which were outlined by each DR algorithm. This highlights the importance of careful selection regarding input data when training a computational model. Out of the methods assessed it is suggested that the molecular quantum number or tokenised SMILES stings are used in any future works.

7.3 General Overview

Although no definitive improvements over existing methods have been found, the wide reaches of the study suggest that graphs, visual representations and machine learning have their place in the field of atmospheric chemistry. Although they may not replace current 'tried and tested' solutions, they have been shown to produce similar and agreeable results and demonstrated the pattern-finding abilities of computational models (for data analysis).

Where these methods come into their own is by demonstrating a more user-friendly approach to model diagnostics, mechanism comparison and change perturbation within a large complex system. Presenting chemistry in such a way enables us to successfully communicate what is going on in a way that policymakers and the general public can understand. This in itself can go a long way into the prevention and mitigation of the global problems described at the start of this thesis.

7.4 Future Work

When discussing future projects relating to this work, there are two apparent avenues which should be explored. The first lies in applying this work to better communicate issues of air quality, whilst the second focuses more on the use of graph neural networks to generate dynamic mechanisms based on user requirements. These are outlined below.

7.4.1 Policy and Communication

As was described previously, one of the more successful parts of this project has been the communication of atmospheric chemistry in a visually intuitive way. Building on this, it would be highly bene to create an 'immersive' user-controllable box model GUI which policymakers and students can adjust, whilst watching the chemical graph dynamically change based on the user regime the chemistry falls into at that point in time. This will go a long way into educating people about the complexities of the atmosphere and how a small change may have a large effect based on circumstances/conditions.

7.4.2 Dynamic Box Model Emulation

Except for long-range transport, much of the chemistry which occurs within different regions of the earth is constrained by the surrounding environment. It would be useful to develop an automatically adapting mechanism based on its position within the earth system - whereupon the number of species and calculations is adjusted in accordance to location, elevation and time of day (photolysis). This will allow global and regional models to provide higher quality results - e.g. by computing high (chemical) resolution runs within urban and surrounding areas whilst removing the same computational overhead for isolated and rural grid boxes.

With the newly emerging age of 'big data', the fields of data analysis and graph theory are everimproving. An example of this is the release of graph convolutional neural networks in 2016 - this is a neural network which takes into consideration not only its inputs but also the relationships between them. If we could get a neural network to learn the protocols for mechanism construction, and then simulate a box model output based on this, the idea of an adaptive mechanism may have potential. This would nicely tie in the visual, mathematical and ML aspects of this thesis.

Reproducability

The code used within this thesis is provided 'as is' within the relevant repositories. There will be an attempt to make it more presentable and fully documented within the near future, but this has not yet happened. For many of the tasks, it is possible to download a clean repository and implement any relevant changes yourself.

The Box Model

Most of the work in this thesis relies on the use of the DSMACC Box model [Emmerson and Evans, 2009]. To reproduce it the specific code I have used can be found in [Ellis, 2020], however, any box model which allows you to extract both the fluxes and Jacobian matrix may be used.

Photolysis Calculations

Photolysis rates are calculated with version 5.2 of the Tropospheric and Ultraviolet and Visible codebase. Photolysis rates are calculated once at the start of each box model run and then interpolated with the use of cubic splines to provide the values required throughout the day. This can be located at [Bräuer, 2020], Photolysis rates within the J array correspond to the lines outlined in ./INPUTS/MCMTUV and are hard-wired within the ./MCMvXX.inc include files.

The Master Chemical Mechanism

For the work, we have made use of various versions of the master chemical mechanism [Rickard, 2020]. Different versions of this and its reduced component (CRI) can be obtained from the MCM website: mcm.york.ac.uk. Alternatively, the KPP presentation of all the mechanisms I have used is located within the ./mechanisms folder in the DSMACC repository.

Kinetic Pre-Processor

To transpose the chemical mechanism into a usable format, the Kinetic Pre-Processor rewrites the human-readable first-order ordinary differential equations into FORTRAN95 code. The version of this originates from FlexChem - the KPP rewrite used in GEOSChem (KPP 2.3.01). This is located at https://github.com/wolfiex/kpp_2.3.01_gc/

ML libraries

Simple processing tasks as clustering, PCA and t-SNE generally make use of the Scikit-Learn package [Pedregosa et al., 2011]. Graph Layouts such as TSNET and Mercator can be found in https://github.com/wolfiex/tsNET and https://github.com/networkgeometry/mercator.

The AutoEncoder code can be found within the DSMACC repository at https://github.com/wolfiex/DSMACC-testing/blob/master/dsmacc/examples/rate_ae.py and the Graph AutoEncoder at https://github.com/tkipf/gae.

Although not documented, this thesis aimed to work up to the use of a graph convolutional network such as the one in https://github.com/wolfiex/gcn.

Chemial representation and Molecular Keys

Chemical species representation for SMILES and INCHI strings are taken directly from the MCM. Additional conversions into MACCS and MQN keys make use of the RDKIT python package: [Landrum et al., 2019].

Observation and model run reproducibility

To reproduce the results made from field campaigns it is possible to extract the data directly from the Centre for Environmental Data Analysis. The four field campaigns used are provided below.

- https://catalogue.ceda.ac.uk/uuid/648246d2bdc7460b8159a8f9daee7844
- https://catalogue.ceda.ac.uk/uuid/81892deb2dd5e7f0d26b9c587af45f3d
- https://catalogue.ceda.ac.uk/uuid/a457d9715f3c4bc295ef975932e491d9
- https://catalogue.ceda.ac.uk/uuid/cee49a1f044b79d5413b7a0282467508

Once downloaded, these are wrangled into the initial conditions CSV format for the use in model runs - some of which are spun up to a steady state based on the user's preference and aim of the study.

Non-observational runs are initiated through the use of a Latin hypercube format to provide a random assortment of initial concentrations within a pre-defined limit. An example of the output of the initial condition for one run of these can be found in https://github.com/wolfiex/DSMACC-testing/blob/master/InitCons/lhs_spinup.csv.

Bibliography

- Aumont, B., Hodzic, A., La, S., Camredon, M., Lannuque, V., Lee-Taylor, J. M., and Madronich, S. (2014). Assessment Of The Gecko-A Modeling Tool And Simplified 3D Model Parameterizations For Soa Formation. *online*, 2014:A23Q–04. https://ui.adsabs.harvard.edu/abs/2014AGUFM.A23Q. .04A.
- Bais, A. F., Lucas, R. M., Bornman, J. F., Williamson, C. E., Sulzberger, B., Austin, A. T., Wilson, S. R., Andrady, A. L., Bernhard, G., and McKenzie (2018). Environmental Effects Of Ozone Depletion, Uv Radiation And Interactions With Climate Change: Unep Environmental Effects Assessment Panel, Update 2017. Photochemical & photobiological sciences: Official journal of the European Photochemistry Association and the European Society for Photobiology, 17(2):127–179. http://dx.doi.org/10.1039/c7pp90043k.
- Bräuer, P. (2020). TUV 5.2X DSMACC. https://github.com/pb866/TUV DSMACC.
- EEA (2018). Air Quality In Europe 2018. https://www.eea.europa.eu/publications/ air-quality-in-europe-2018.
- Ellis, D. (2020). DSMACC-Testing. https://github.com/wolfiex/DSMACC-testing.
- Emmerson, K. M. and Evans, M. J. (2009). Comparison of tropospheric gas-phase chemistry schemes for use within global models. *Atmospheric Chemistry and Physics*, 9(5):1831–1845. https://www.atmos-chem-phys.net/9/1831/2009/.

- Fitzky, A. C., Sandén, H., Karl, T., Fares, S., Calfapietra, C., Grote, R., Saunier, A., and Rewald, B. (2019). The interplay between ozone and urban vegetation—bvoc emissions, ozone deposition, and tree ecophysiology. *Frontiers in Forests and Global Change*, 2:50. https://www.frontiersin.org/ article/10.3389/ffgc.2019.00050.
- Hardin, P. J. and Jensen, R. R. (2007). The effect of urban leaf area on summertime urban surface kinetic temperatures: A terre haute case study. Urban Forestry and Urban Greening, 6(2):63 – 72. http://www.sciencedirect.com/science/article/pii/S1618866707000180.
- IPCC (2013). Climate Change 2013: The Physical Science Basis. Contribution Of Working Group I To The Fifth Assessment Report Of The Intergovernmental Panel On Climate Change. Cambridge University Press, Cambridge, United Kingdom and New York, NY, USA. www.climatechange2013. org.
- Jacob, P.-M. and Lapkin, A. (2018). Statistics of the network of organic chemistry. *React. Chem. Eng.*, 3:102–118. http://dx.doi.org/10.1039/C7RE00129K.
- Jenkin, M. E., Young, J. C., and Rickard, A. R. (2015). The mcm v3.3.1 degradation scheme for isoprene. Atmospheric Chemistry and Physics, 15(20):11433–11459. https://www.atmos-chem-phys. net/15/11433/2015/.
- Landrum, G., Tosco, P., Kelley, B., sriniker, gedeck, NadineSchneider, Vianello, R., Dalke, A., Cole, B., AlexanderSavelyev, Turk, S., Ric, Swain, M., Vaucher, A., N, D., Wójcikowski, M., Pahl, A., JP, strets123, JLVarjo, O'Boyle, N., Berenger, F., Fuller, P., Jensen, J. H., Sforna, G., DoliathGavid, Cosgrove, D., Nowotka, M., Leswing, K., and van Santen, J. (2019). Rdkit 2019-03-2 (q1 2019) release. https://doi.org/10.5281/zenodo.2864247.
- Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., and Duchesnay, E. (2011). Scikit-Learn: Machine Learning In Python . Journal of Machine Learning Research, 12:2825–2830.
- Rickard, A. (2020). MCM Website. http://mcm.york.ac.uk/.
- Watts, D. J. and Strogatz, S. H. (1998). Collective Dynamics Of 'Small-World' Networks. Nature, 393(6684):440–442. http://dx.doi.org/10.1038/30918.
- Yli-Pelkonen, V., Viippola, V., Rantalainen, A.-L., Zheng, J., and Setälä, H. (2018). The impact of urban trees on concentrations of pahs and other gaseous air pollutants in yanji, northeast china. Atmospheric Environment, 192:151 – 159. http://www.sciencedirect.com/science/article/ pii/S135223101830582X.

Appendices

Appendix A

Supplementary Mathematics

A.1 PCA

A.1.1 Statistics

Firstly we define the variance:

$$\sigma = \frac{\sum_{i=1}^{N} (X - \mu_X)(X - \mu_X)}{n - 1}$$
(A.1)

where X is the dataset, μ the mean and n the number of data points.

If we wish to then compare dataset X with dataset Y we may use the covariance:

$$cov(X,Y) = \frac{\sum_{i=1}^{N} (X - \mu_X)(Y - \mu_Y)}{n - 1}$$
(A.2)

For n distinct variables we may construct an $n \times n$ matrix containing $n!/(n-2)! \times 2$ different combinations of covariences:

$$C = \begin{pmatrix} \sigma_X & cov(X,Y) & cov(X,Z) & \cdots & cov(X,n) \\ cov(Y,X) & \sigma_Y & cov(Y,Z) & \cdots & cov(Y,n) \\ cov(Z,X) & cov(Z,Y) & \sigma_Z & \cdots & cov(Z,n) \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ cov(n,X) & cov(n,Y) & cov(n,Z) & \cdots & \sigma_n \end{pmatrix}$$

A.1.2 Matrices and Eigenvectors

An eigenvector is a vector \mathbf{v} , that when operated on by a given operator produces a scalar multiple of itself (Equation A.3) - this scalar multiple is called the eigenvalue λ . Eigenvectors can only be found for square matrices and are perpendicular to the matrix regardless of their dimension. A $n \times n$ matrix

will produce n eigenvectors. Conventionally these are scaled to unity, which may be done by dividing the eigenvector by the Pythagorean distance of each element.

$$C\mathbf{v} = \lambda \mathbf{v} \tag{A.3}$$

An example of an eigenvector/value pair is shown in the following equations:

$$\begin{pmatrix} 2 & 3 \\ 2 & 1 \end{pmatrix} \times \begin{pmatrix} 3 \\ 2 \end{pmatrix} = \mathbf{4} \begin{pmatrix} 3 \\ 2 \end{pmatrix}$$
(A.4)

One property of the eigenvalue/eigenvector pair is that the square matrix acts as a transformation on the eigenvector. This means that we may treat the eigenvector as a direction from the origin, whose magnitude we can scale. The eigenvalue remains to scale independent and is the same value as before:

$$\begin{pmatrix} 2 & 3 \\ 2 & 1 \end{pmatrix} \times \begin{pmatrix} 6 \\ 4 \end{pmatrix} = 4 \begin{pmatrix} 6 \\ 4 \end{pmatrix}$$
(A.5)

A.2 t-SNE

A.2.1 Student t distribution

Created by William Gosset and published under the pseudonym student ¹?.

The distribution consists of a family of continuous probability distributions which may be used when the sample size is small, and the standard deviation is unknown. The curve itself resembles that of a normal distribution, just with a shorter amplitude and greater full width at half maximum (FWHM).

A.2.1.1 t-Score

Much like the z-score mentioned earlier [ref standardiz], t-scores also convert individual values to a standard form. This is generally used when you do not know the population standard deviation (often due to having too few data points). At greater than 30 datapoints this resembles the equation of the z-score, and will often give you the same result.

$$t(x_i) = \frac{x_i - \mu_x}{S_{sample}/\sqrt{n}} \tag{A.6}$$

 $^{^{1}}$ At the time Gosset was employed by Guinness Breweries in Dublin. This meant that chemists were forbidden from publishing their findings. After explaining that his mathematical and philosophical conclusions were of no use to competing breweries, he was finally allowed to publish under the pseudonym 'student'. This was mainly to avoid difficulties with the rest of the staff.

Appendix B

Neural Network Activation Functions

B.1 Binary Step

This is a simple threshold function. If the input is above the threshold, the message is passed on. This makes it efficient, but unable to classify a single input into multiple categories. This can be likened to a yes|no decision tree.

$$f(x) = \begin{cases} 1, & \text{if } x < threshold \\ 0, & \text{otherwise} \end{cases}$$
(B.1)
Binary Step

Figure B.1: Binary Step activation function.

B.2 Linear

This produces a signal proportional to the input multiplied by the weight of each neuron. It is an improvement over the step function as it allows for multiple outputs. It does, however, mean that we are unable to use backpropagation (gradient descent) to train the model. In addition to not being able to improve a model, all the layers in the neural network collapse into a single layer. This means that the final layer will always be a linear function of the first layer. This eliminates all the merits which may be gained from deep learning. A neural network with a linear activation function is simply a linear regression model.

$$f(x) = m(x) \tag{B.2}$$

Figure B.2: Linear activation function.

B.3 Sigmoid / Logistic

The first of the non-linear activation functions, the sigmoid activation function has a smooth gradient providing smooth output values which are bound between 1 and 0, normalising the output of each neuron. The main disadvantage is that falls foul the vanishing gradient problem - for extreme values of x there is close to no change in the prediction. This may result in either early termination of the training or a slow training cycle in obtaining adequate precision. The activations are computationally expensive, and the outputs are not zero centred.

$$f(x) = 1/1 + e^x$$
 (B.3)



Figure B.3: Sigmoid activation function.

B.4 Hyperbolic Tangent

Much like the sigmoid function in both advantages and disadvantages. The hyperbolic tangent function provides a smooth curve which is zero centred. It is, however, computationally expensive and suffers from the vanishing gradient problem.

$$f(x) = \frac{e^x - e^{-x}}{e^x + e^{-x}}$$
(B.4)



Figure B.4: Tanh activation function.

B.5 Rectified Linear Unit

A commonly used activation for large deep neural networks, due to its computational efficiency and quick convergence. It is non-linear although it appears like a linear function, and allows for backpropagation. It does, however, suffer from the dying ReLU problem - when inputs tend to zero or below, the gradient of the function becomes zero and the network cannot perform backpropagation to learn.

$$f(x) = \begin{cases} 0, & \text{if } x < threshold \\ x, & \text{otherwise} \end{cases}$$
(B.5)



Figure B.5: ReLU activation function.

B.6 Swish

The Swish Activation is a self-gated activation function discovered by google (https://arxiv.org/abs/1710.05941v1). Is has been shown to perform better than ReLU at a similar level of computational effidiency and generates results just under 1% more accurate.

$$f(x) = x/1 - e^{-x}$$
(B.6)
Swish

Figure B.6: Swish activation function.

B.7 A note on backpropagation

As it has not been explicitly explained before backpropagation is an algorithm used to train neural networks. The derivative (or gradient) of an activation function is important in the use of backpropagation. Here the model weights are adjusted, and improved, by tracing back all the connections in the network, suggesting an optimal weight of each neuron.

Appendix C

Graphs and Networks

C.1 Heavily Labeled Citation Graph



Figure C.1: The labelled co-author network - as referenced in Chapter 4

C.2 Centrality on the UK rail network.

The data in the following section was extracted from OpenStreetData using overpass turbo https: //overpass-turbo.eu. Here ways within the geographic information system mapping (GIS) format are represented as paths between locations (i.e. a graph). Following some simple processing, the distance of each section was calculated, and a weighted graph represented the UK rail network was generated. This was then used to form a rudimentary analysis of the graph structure using centrality metrics (shown below).

NOTE:

Since the network is extracted from a GIS data file, nodes within the rail network include not only stations but also switches, routing nodes and crossings. Although this can be filtered, the iterative reconstruction of a graph for the entire UK is a lengthy process - one which I am unable to do at the time of writing (there are 90011 nodes which make up the entirety of the UK rail network, most of which need to be removed, and their links rerouted).



(c) Betweenness Centrality.

(d) **PageRank Centrality.** Colours are groups from the modularity clustering algorithm which partitions the network into highly connected areas or 'cliques'.

Appendix D

Miscellaneous

D.1 Correspondence with Mike Jenkin

Mike Jenkin 11th September 2019

Note on naming conventions in the CRI mechanism

The lumped or "common" species in the CRI mechanism are, by definition, used to represent a set of real species with different structures and properties. The criterion for lumping is the maximum number of NO-to-NO2 conversions (i.e. maximum number of ozone molecules) that the subsequent degradation can produce - and lumped species can, therefore, represent a large number of real species with different structures and properties.

In later expansions of the mechanism, the chemistry for species such as isoprene and terpenes defined intermediates that are representative of more restricted sets of real species. For these, it is possible to relate them to more restricted sets of MCM species that are the main contributors.

Although I tried to be logical in naming, the mechanism was developed over many years with little or no funding and may therefore not be fully transparent and foolproof throughout. However, I think quite a lot of the naming is logical, as expanded on below.

1) The numbers in most of the species names (the "CRI index") are the number of NO-to-NO2 conversions that can result from the subsequent OH-initiated NO-propagated chemistry. For radical termination products (e.g. hydroperoxides formed from RO2 + HO2 and nitrates formed from RO2 + NO), this is a grey area, and the number is, therefore, the same as that for the precursor RO2 radical. In these cases, it is simply a convenient label.

2) There are a number of series of peroxy radicals, which are denoted RNxxO2, RIxxO2, RAxxO2, RExxO2, RUxxO2, RTNxxO2, RTXxxO2. These represent peroxy radicals with different structural features or formed from different types of precursor, as indicated below. Occasionally, extra peroxy radicals with the same CRI index are included by inserting a letter after the index (e.g. RNxxAO2) to increase flexibility of the mechanism. Peroxy radicals formed specifically from addition of NO3 to an alkene/diene are prefixed by "N".

RNxx02: These were originally representative of peroxy radicals formed from linear or "n-" alkanes and their carbonyl products. They are also used for peroxy radicals formed from slightly-branched precursors (e.g. 2-methylhexane), and are formed as a convenient default intermediate with the correct CRI index in the latter stages of degradation of other precursor classes.

RIxxO2: These were originally representative of peroxy radicals formed from branched or "i-" alkanes and their carbonyl products, but tend to be used only for smaller branched precursors that can produce acetone as a major product from their subsequent degradation. This is because acetone is a particularly unreactive carbonyl, the formation of which can interrupt the ozone formation processes under typical regional-scale photochemical episode conditions in north-west Europe.

RAxxO2: These peroxy radicals are formed from the addition of OH to aromatic compounds, and are complex bicyclic structures containing a peroxide bridge (e.g. like BZBIPERO2 in the MCM).

RExxO2: These peroxy radicals are formed from ether degradation, and allow the formation of unreactive formate ester products to be represented.

RUxxO2: These peroxy radicals are formed from degradation of conjugated dienes (currently only isoprene and 1,3-butadiene). Those formed initially (e.g. RU14O2) contain allyl functionalities (i.e. a specific unsaturated linkage), although the terminology is also used for some peroxy radicals formed from subsequently-formed unsaturated products.

Related to this, the species CRU1402 and TRU1402 in the EMEP variant of CRI v2.2 (described in https://doi.org/10.1016/j.atmosenv.2019.05.055) were specifically introduced to represent the cis- and trans- isomers required for the Peeters (LIM) reaction framework. CRU1402 represents CISOPA02 and CISOPC02 in MCM v3.3.1 and TRU1402 represents ISOPA02 and ISOPC02 in MCM v3.3.1. However, CRI v2.2 itself uses a different approach where the chemistry is represented by a conditions-dependent rate coefficient for the single peroxy radical, RU1402.

RTNxxO2: This terminology is used for peroxy radicals formed from monoterpenes containing an endocyclic double bond. This is currently limited to α -pinene in CRI, although the original idea was that the mechanism could be used as a surrogate for other endocyclic monoterpenes by simply adding new sets of initiation reactions. RTXxxO2: This terminology is used for peroxy radicals formed from monoterpenes containing an exocyclic double bond. This is currently limited to β -pinene in CRI, although the original idea was that the mechanism could be used as a surrogate for other exocyclic monoterpenes by simply adding new sets of initiation reactions.

Finally, the species DHPR1202 in CRI v2.2 is a peroxy radical containing two hydroperoxy groups. Again, it is required for representation of the Peeters (LIM) mechanism, and is representative of the species C53602 and C53702 in MCM v3.3.1 (these species being referred to as "di-HPCARPs" by Peeters et al., 2014: https://doi.org/10.1021/jp5033146).

3) Hydroperoxides formed the reactions of HO2 with the above peroxy radicals have "OOH" in place of "O2". Nitrates formed the reactions of NO with the above peroxy radicals have "NO3" in place of "O2".

4) There are a number of series of carbonyl compounds, which are denoted CARBxx, UCARBxx, UDCARBxx, TNCARBxx and TXCARBxx.

CARBxx: These are used to represent carbonyls and hydroxycarbonyls. Occasionally, extra carbonyls/hydroxycarbonyls with the same CRI index are included by inserting a letter after the index (e.g. CARBxxA) to increase the flexibility of the mechanism.

Related to this, the species DHPCARB9 in CRI v2.2 is a carbonyl containing two hydroperoxy groups. Again, it is required for representation of the Peeters (LIM) mechanism, and is representative of the species DHPMEK and DHPMPAL in MCM v3.3.1 in MCM v3.3.1.

UCARBxx: This terminology is used for unsaturated carbonyls/hydroxycarbonyls, formed for example from isoprene (although one of the main ones, UCARB10, has been "unlumped" into MVK and MACR in the EMEP CRI v2.2 variant).

Related to this, the species HPUCARB12 in CRI v2.2 is an unsaturated carbonyl containing a hydroperoxy group. Again, it is required for representation of the Peeters (LIM) mechanism, and is representative of the species C5HPALD1 and C5HPALD2 in MCM v3.3.1.

UDCARBxx: This terminology is used for unsaturated dicarbonyls, formed from aromatics.

TNCARBxx and TXCARBxx: This terminology is used for carbonyl compounds, formed from monoterpenes with endocyclic and exocyclic double bonds, respectively.



D.2 Functional Groups

Appendix E

Chapter Keywords

This section uses the Term Frequency Inverse Document Frequency to determine the keywords of each chapter - a technique which has been described in Chapter 4. Text size corresponds to the importance of each word.

E.1 Introduction

OZONE MODEL RO ATMOSPHERE CHEMISTRY SPECIES AGO NUMERICAL. ATMOSPHERIC AIR EARTH HO CHEMICAL DT CL.TMATE. CONCENTRATIONS OH HOX NNO PLANET OXYGEN RADICAL NOX MECHANISM CYCLE YEARS CHANGE POLLUTION TIMESCALES KM SOLVERS HNO GCM TIME NITROGEN SYSTEM REACTION RANGE INCREASE RESULTED FUNDAMENTALS EMISSIONS ENERGY RISE LEAD NIGHT TROPOSPHERE KNOWN THESTS DEVELOPMENT HUMANKIND SINK DEVCYCLE HYDROXYL PARIS RAPID GEAR CHINANOX RUNGE GEOS KUTTA HOMO BILLION ROSENBROCK TRANSPORTED HOWEVER METHODS POSSIBLE UNDERSTANDING BOX IPCC PARTS MODELS REPRESENTATION MANY REACTIONS METHOD DIFFERENT SCIENCE DECREASE STARTED HUMANITY STIFF GAS PPTV REVOLUTION POWER SOLVED EVENTUALLY EQUATIONS SCIENTIFIC BATE SOURCE REACT CURRENT CELLS DSMACC FORWARDS GLOBAL ANTHROPOGENIC

E.2 Visualisation and its use in understanding Complex Data

SPECIES ARC FTGURES MCM REACTIONS CHORD OH NUMBER ARCS GROUPS DATA MECHANISM PROTOCOL STORYTELLING METAPHOR DIFFERENT HO INFORMATION NETWORK SOCIAL FUNCTIONAL CHEMISTRY VISUALISATION CRI GRAPH REPRESENTATION REACTION TWO SHOWS DIAGRAM TRUNK KBPAN NIGHTINGALE SHOW PLOT BRANCHES DESIGN ITEMS GROUP PANS CR FLOWCHART FEATURES CHEMICAL MAY TIME PROCESS TREE CONTAIN ABILITY SOURCE METAPHORS COMPOSITE EVENTS RED MANY REPRESENT COMPLEX DIAGRAMS THICK BECK SOCIOGRAPHS CAVE TEPHI GOSSIP FAMILIARITY ENABLE HYDROXIDE IDEAS KNOWLEDGE NEW REPRESENTED OFTEN US RELATIONSHIPS POSSIBLE SEVERAL LINES HV BIG LIMITED CHANGES SEEN ALTHOUGH SHOWN STRUCTURE METHODS LIKE INTUITIVE CONNECTED EXIST REASON HUMANS ROOH CARBONS NARRATIVE EFFECTIVE BLUE MAIN #XIS

E.3 Applying Visual Analytics to the Atmospheric Chemistry Network

LAYOUT NODES FIGURES EDGE NODE REPRESENTATION SPECIES EDGES GRAPHS CONFLUENT LAYOUTS MERC DESIGN FORCEDIRECTED MERCATOR DENSITY ALGORITHM INFORMATION SEMANTIC NETWORK DISTRIBUTION MECHANISM MAY REACTIONS OPENORD HU CHEMISTRY ANGLE FORCE CROSSING ROUTING BEZIER YIFAN STRUCTURE VISUALISATION APHH POSSIBLE DATA CHEMICAL MCM BUNDLING TSNET SHOWS AREA CM DIFFERENT BUTANE DRAWING QUADTREE FORCEATLAS ATLAS DEGREE LINKS BEIJING ENERGY PROCESS SYNTACTIC CURVES ONE REPRESENT USER ITEMIZE REPRESENTING CLUTTER ADDITION EXAMPLE VISUAL TSNE CONF BEIJINGTEST ALTHOUGH MANY SHOWN NUMBER SYSTEM SINCE RESOLUTION BEST SHAPES METHANE DIRECTED OFTEN HIGH NETWORKS SHAPE CARBON CURVEDEDGE ALLSAMPLES FILL EVETRACK METHODS ORTHOGONAL INTERACTIVITY POINTS INTERACTION ORANGE FOUND LARGE DIRECTION

E.4 Chemical model diagnostics using graph theory and metrics.

ECTES NETWORK GRAPH NODE NAPINBO PAPERS PAGERANK JACOBIAN CLOSENESS MCM FIGURES EQNARRAY SUM METRICS ALGORITHM DATA MATRIX METRIC MLPREGRESSOR BETWEENNESS FREQUENCY RESULTS MAY MONTH CITATION DOCUMENT CONCENTRATION GOOGLE FLUX NODES VALUE NUMBER CHEMISTRY OH MECHANISM TOTAL CO OBSERVATIONAL PAGE RANK INFLUENCE VALUES AUTHORS IDF MODEL PERCEPTRON ANALYSIS MLP INFORMATION STRUCTURE TFIDE TIME SIMULATION POSSIBLE EDGE LINKS IMPORTANCE NAPINBOOH BEND DATASET US DIFFERENT TAB CONCENTRATIONS LINE BEIJING ONE EXAMPLE SEEN CHEMICAL IMPORTANT EVERY PAPER SEGMENT LONDON CHANGE METHOD AUTHOR CENTER MAIN SHOWING RANGE TRADITIONAL BMATRIX ARTICLES MANY SMALL OFTEN WORD SINCE SOURCE DEGREE NET METHODS PRODUCTION PREDICTED PRECURSORS FONT INVERSI

E.5 Using Graph Clustering And Natural Language Processing To Aid Mechanism Reduction.

MECHANTSM LIFETIME INFOMAP MAY CLUSTERING HOCH NETWORK GRAPH COSINE CRT ALGORITHM REACTIONS NODES SIMILAR CO CLUSTERS LUMPED CHEMICAL MAGNITUDE DENSITY STRUCTURE NUMBER METRIC RESULTS GROUPS LUMPING HUFFMAN NRI ALLUVIAL EUCLIDEAN REDUCTION DISTANCE PEROXY PLOT ENSEMBLE REACTION MCM MATRIX POSSIBLE GROUP PAIRS DYNAMICS PREFIX WORST RO CLUSTER APPEAR HIERARCHICAL JACOBIAN SIMULATION METHOD TOGETHER CHEMISTRY PRODUCE SEEN LUMPPAIR RI DIEK PROFILES OOH TIME TEMPORAL DIFFERENT TWO REPRESENTATIVE LEVEL BEST WORKS GROUPED TASK FLOW VECTOR NUMBERS SINCE CODE PROCESS LINKS USEFUL CARBONYL DATA IMAP TRAPPED MODULES WALKER II LIFETIMES PINENE PHOTOLYTIC QSSA PECOH MORIG IML HUFF EUCLID LOUVAIN REDUNDANT EXPERIMENT VALUES REACT
E.6 Computational Learning of Species Structure using Visualisation and Vector Clustering

PCA SPECIES CLUSTERS TSNE DATA SMILES GROUPS DR ALGORITHM VEC **CLUSTER** ALGORITHMS DATASET DIMENSIONALITY FUNCTIONAL CLUSTERING INPUT STRUCTURE GRAPH GROUP NODE NUMBER REDUCTION MACCS SILHOUETTE POINTS RANDOM METHODS AE PRINCIPAL OUTPUT ACTIVATION STRING OUTPUTS AUTOENCODER COLOURS VECTOR COEFFICIENT FEATURE DIFFERENT TWO IMPORTANCE FINGERPRINTS LINEAR SINCE QUANTUM TABLES LANDSCAPE INPUTS BEST DISTRIBUTION DIMENSIONS ST ALTHOUGH FEATURES CHEMICAL KEYS TEX NONLINEAR MOLECULAR ORIGINAL OFTEN MAY POSSIBLE RESULT EMBEDDING COMPARING FORESTS MCM MECHANISM ONE MUCH PROCESS FINGERPRINT MQN PARAMETERS REPRESENTING CHEMISTRY COLOUR SET RESULTS SHOWS THREE PCAVIS TSNEVIS VARIANCE GREEDY CASE ANALYSIS METHOD STRINGS CONTAIN FUNCTIONS MAPPING COMPONENT PROBABILITY MATRIX DESCRIBED SPACE REPRESENT

E.7 Conclusions and Future Work

CHEMICAL MECHANISM SPECIES MODEL DATA WOLFIEX CEDA CATALOGUE UUID NETWORK GRAPH THESIS BASED MECHANISMS BOX COMPUTATIONAL PHOTOLYSIS UK WHILST ANALYSIS INFORMATION EARTH DSMACC EXAMPLE FOUND WAY DIFFERENT WORK CHEMISTRY COMMUNICATE REPOSITORY KPP FUTURE RESULTS FIELD STUDY NEURAL USER RUNS OUTLINED CHANGE SYSTEM STRUCTURE MODELS LARGE ONE MASTER REGIME DSMACCTESTING ENVIRONMENTAL RELEVANT REPRODUCABILITY CAMPAIGNS CSV CONVOLUTIONAL PHYSICAL QUALITY ADDITIONAL CODE TYPE OZONE ATMOSPHERIC RATES LOCATED NUMBER RUN REPRESENTATION VISUAL MANY MAY MCM US ALTHOUGH PRODUCTION SETTING SURROUNDING TSNET RESOLUTION POLICYMAKERS MERCATOR KEYS CALCULATIONS ML SMILES SMALLWORLD VERSIONS RDKIT DOCUMENTED MUST AIR REPRESENTATIONS INITIAL OUTPUT EFFECT GRAPHS LINKS REDUCED INCREASE SIMILARLY TECHNIQUES