

UNIVERSITY OF SOUTHAMPTON
FACULTY OF NATURAL AND ENVIRONMENTAL SCIENCES
School of Chemistry

Crystal Structure Prediction of Organic Semiconductors

by

Josh E. Campbell

Thesis for the degree of Doctor of Philosophy

May 2017

UNIVERSITY OF SOUTHAMPTON

ABSTRACT

FACULTY OF NATURAL AND ENVIRONMENTAL SCIENCES

School of Chemistry

Doctor of Philosophy

CRYSTAL STRUCTURE PREDICTION OF ORGANIC SEMICONDUCTORS

by Josh E. Campbell

This thesis presents the use of crystal structure prediction (CSP) in the evaluation and design of novel organic semiconductors. Heteroatom substitution into common organic semiconductors (pentacene in this thesis) offers a way of modulating their crystal packing and electronic properties. Initially CSP was performed on six human designed molecules and the charge mobility of their predicted crystal structures was calculated. The packing landscapes changed significantly from the unsubstituted pentacene. We found that five nitrogen atoms led to a landscape showing a range of packing motifs, while seven nitrogen atoms favours the adoption of sheet-like motifs. Substitution patterns expected to result in the highest mobilities were found to perform worse than assumed, showing the importance of tuning both molecular electronic properties and crystal engineering. A genetic algorithm was then developed to generate new nitrogen substituted pentacenes. A population members fitness was calculated using two molecular properties important for electron transport in organic semiconductors. Five runs of the genetic algorithm gave 12 promising candidates for CSP and mobility calculations. The packing landscapes were similar to those of the seven nitrogen substituted human designed molecules. One genetic algorithm molecule showed a high number of high mobility structures close to the global minimum, making this molecule an attractive target for synthesis. Extensions to include CSP within the fitness function of the genetic algorithm represents possible future work. Additional work included the design and testing of structure generator for the generation of trial crystal structures during a CSP. The novel structure generator performed well in locating the experimental structures of three test molecules and was used in the group's submission to the 6th blind test, of which one molecule is also presented here. The experimental structure of this molecule was located in lists ranked by lattice energy and free energy, though the free energy list ranked the experimental structure as the global minimum.

Contents

Declaration of Authorship	xvii
Acknowledgements	xix
Abbreviations	xxi
1 Introduction	1
2 Molecular Crystals and Crystal Structure Prediction	5
2.1 Crystals and Polymorphism	5
2.2 Crystal structure prediction	10
2.2.1 Molecular model	12
2.2.2 Crystal structure generation	12
2.2.2.1 Systematic grid searches	13
2.2.2.2 Random searches	13
2.2.2.3 Genetic Algorithms	13
2.2.3 Ranking Structures	14
2.3 Intermolecular Interactions and Force Field Methods	15
2.3.1 Long range intermolecular interactions	15
2.3.2 Short range intermolecular interactions	16
2.3.3 Relation to physical contributions to the intermolecular interaction	17
2.3.4 Functional forms	18
2.3.5 CSP methods used in this thesis	20
2.4 Progress in CSP	21
2.4.1 The blind tests	21
2.4.2 Flexibility in CSP	28
2.4.3 Beyond classical methods	29
2.5 Conclusions	29
3 Organic Semiconductors	31
3.1 Introduction	31
3.2 Organic Semiconductors	32
3.2.1 Charge transport	33
3.2.2 Marcus theory, reorganisation energy and transfer integrals	36
3.2.3 PAHs and Pentacene	38
3.2.4 Charge transport calculations in this thesis	42
3.2.5 Conclusions	44

4	Machine Learning and Genetic Algorithms	45
4.1	Machine Learning and Genetic Algorithms	45
4.1.0.1	Support Vector Machines	50
4.1.0.2	Linear Regression	52
4.1.0.3	Artificial Neural Networks	53
4.1.0.4	k -Nearest Neighbours	56
4.2	Genetic Algorithms	58
4.2.1	GA operators and concepts	59
4.2.2	Selection methods	60
4.2.3	Crossover and mutation	64
4.2.4	Other operators and parameters	65
4.2.5	When to use a GA	67
4.2.6	Brief GA history	67
4.3	Conclusions	69
5	Structure Generation and Blind Test	71
5.1	Introduction	71
5.2	Structure generator	71
5.2.1	Initial testing of the structure generator	76
5.2.2	Quinacridone and full testing of rigid structure generation	81
5.2.2.1	Lattice Energy Minimisation and Clustering	83
5.2.2.2	Results and discussion	84
5.2.2.3	Full searches	90
5.3	CSP of a blind test molecule	93
5.4	Conclusions	100
5.5	Additional figures	102
6	Azapentacene Crystal Structure Prediction	105
6.1	Introduction	105
6.2	Molecules studied	106
6.2.1	Validation molecules	106
6.2.2	Hypothetical molecules	108
6.3	Methods	109
6.3.1	CSP	109
6.3.2	Classification of Predicted Crystal Structures	110
6.3.3	Mobility calculations	111
6.4	Results and discussion	112
6.4.1	CSP	112
6.4.1.1	Validation molecules	112
6.4.1.2	Hypothetical azapentacenes	115
6.4.2	Charge mobility of predicted structures	118
6.5	Conclusions	120
7	Azapentacene Genetic Algorithm	123
7.1	Introduction	123
7.2	The GA	124
7.2.1	Encoding	124

7.2.2	Generating a population	126
7.2.3	Calculating fitness	128
7.2.4	Selection methods	129
7.2.5	Crossover	130
7.3	GA testing	136
7.3.1	Population size	137
7.4	Conclusions	143
8	Genetic Algorithm Production Runs	145
8.1	Introduction	145
8.2	Improvements to the GA	145
8.3	Production runs of our GA	150
8.3.1	GA results	150
8.3.2	GA molecules CSP and mobility calculations	159
8.3.3	Results and discussion	160
8.4	GA extensions	166
8.5	Conclusions	170
8.6	Additional figures	171
9	Conclusions	175
9.1	Azapentacene CSP and GA	175
9.2	Other CSP work	179
	References	181

List of Figures

2.1	The three known polymorphs of benzamide (CSD refcode BZAMID), form I is the stable polymorph, form II metastable discovered in 2005 and form III the silky needles first seen in 1823	6
2.2	The two poster children for polymorphism. The lack of knowledge of an additional ritonavir polymorph led to the withdrawal of the drug. ROY currently has the most known polymorphs.	7
2.3	Four organic semiconducting molecules with large mobility differences between polymorphs	8
2.4	Two pigment molecules known the exhibit polymorphism, with some particular polymorphs being much more useful as pigments	9
2.5	Two explosive molecules known the exhibit polymorphism	10
2.6	The steps in a crystal structure prediction (QM = quantum mechanical, FF = force field	11
2.7	Lennard-Jones potential	18
2.8	Buckingham potential	19
3.1	The first five acenes	32
3.2	Some of the benchmark organic semiconductors	34
3.3	Definition of the internal reorganisation energy for an electron-transfer reaction.	35
3.4	The four most common packing motifs in PAHs	38
3.5	(a) Herringbone packing of pentacene with edge to face interactions highlighted, (b) altered herringbone packing seen upon perfluorination of pentacene.	39
3.6	Changes in packing as the silyl substituent becomes larger	41
4.1	Schematic representation of a support vector machine in two dimensions. H_1 does not separate the classes correctly, H_2 does but with a very small margin and H_3 separates them with the maximum margin.	51
4.2	Structure of a basic NN.	54
4.3	Structure of one neuron, where g is the transfer function.	55
4.4	How the k NN works for classifying (or predicting) the colour of a rectangle. When $k=1$ the green query will be classified as belonging to the red class, when $k=3$ it is again red by a 2-1 vote, when $k=5$ it will be classified as blue by 3-2.	57
4.5	Structure of a basic GA.	59
4.6	Example of the selection of a single individual using FPS.	61

4.7	Example of selection of multiple members of the population using SUS. F/N is the width of the pointers. $re \in [0, F/N)$ the initial "spin" to choose where to begin selecting.	62
4.8	In this example, population member A would dominate the members chosen for crossover and negatively impact the diversity of the next generation. With rank selection A still has a higher chance to be selected but the fitness variance has been masked.	63
4.9	The three common crossover schemes seen in GAs, parent solutions on the left, new children on the right.	65
4.10	Mutation operator acting on a bitstring	65
4.11	The structure of Fraser's evolution program, containing the foundation of today's GAs	68
5.1	Molecular projections onto the lattice vectors, used to define the sampling range for unit cell lengths. The directions of the three lattice vectors, $l_{1,2,3}$ are shown, and the molecular projections of two quinacridone molecules are shown onto lattice vector l_2 . Thin lines show the projection of the edges of the van der Waals radii of each atom onto the lattice vector. Bold red and blue lines show the molecular shadows onto l_2	74
5.2	SAT test for molecular overlap. SAT prescribes the vectors upon which to project the vertices of the convex hulls when testing polytopes for their overlap in space. An example for the cage molecule CC1 is shown with convex hulls overlaid on the molecular geometry. In the geometry shown there is a vector upon which the "shadows" of their hulls, the blue and red vectors, do not overlap. If they did overlap, the set of overlapping blue and red vectors would determine the minimum displacement necessary to separate them in the direction of that vector.	75
5.3	7A one of the azapentacenes studied in other parts of the thesis.	76
5.4	Typical flat cell produced.	76
5.5	Distribution of ldp from the CSD (5.34) (search restricted to organic molecules in a triclinic space group)	77
5.6	Initial ldp distribution of structure generator for 2000 structures in $P1$ for molecule 7A.	78
5.7	Plot showing the differences between sampling functions	79
5.8	Linear sampling ldp distribution (7A $P1$)	80
5.9	Arcsine sampling ldp distribution (7A $P1$)	80
5.10	The three molecules chosen to test the structure generator.	82
5.11	The number of unique crystal structures within 15 kJ/mol of the global minimum (60 kJ/mol for CC1) for all searches and all space groups, displayed as a function of the total number of successfully energy minimised structures.	86
5.12	The average lattice energy of the ten lowest structures is shown as function of the number of minimised structures. The solid line indicates the energy of the single lowest energy structure, where the colour matches the legend. For CC1 the data had converged at 60 minimisations so is truncated.	87
5.13	Bar charts showing the frequency with which each low energy structure is located. The lowest 10 are shown, with number of hits on the vertical axis and relative energy difference from the global minimum on the horizontal. Each of the five searches appear alongside each other.	88

5.14	Hits to the 10 lowest ranked crystal structures of quinacridone $P\bar{1}$. Each point represents a minimisation from a trial structure showing where the trial structure was generated in the Sobol sequence.	89
5.15	Lattice energy versus density plots for quinacridone. Each point is a unique minimum in the lattice energy surface upto 15 kJ/mol from the global minimum. Two searches were performed for quinacridone with the basis set used for the calculation of the multipoles in parentheses. For 6-311G** the α polymorph appeared too high in the set to be plotted with all other matches to experimental structures circled.	91
5.16	The five molecules chosen for the 6th blind test of CSP.	95
5.17	The two XXII conformations chosen for CSP.	96
5.18	Lattice energy landscape of XXII, plotting relative lattice energy versus density. The lowest 100 structures are shown. Labels refer to the space group number.	98
5.19	Lattice energy landscape of XXII. Plotting relative free energy (T=300K) versus density. The lowest 100 structures are shown and labels refer to the space group number.	99
5.20	Overlay of the global minima (green) from our free energy ranked list of submitted structures with the experimental structure of XXII (RSMD ₃₀ = 0.292 Å).	100
5.21	Hits to the 10 lowest ranked crystal structures of quinacridone $P\bar{1}$. Each point represents a minimisation from a trial structure showing where the trial structure was generated in the Sobol sequence.	102
5.22	Hits to the 10 lowest ranked crystal structures of quinacridone $P2_1/c$. Each point represents a minimisation from a trial structure showing where the trial structure was generated in the Sobol sequence.	103
5.23	Hits to the 10 lowest ranked crystal structures of quinacridone $Z' = 2 P\bar{1}$. Each point represents a minimisation from a trial structure showing where the trial structure was generated in the Sobol sequence.	104
6.1	The two validation molecules chosen for our study.	107
6.2	Crystal packing of the two validation molecules.	108
6.3	The hypothetical molecules chosen for this study (hydrogen omitted).	109
6.4	The four packing types seen in crystal structures of polycyclic aromatic hydrocarbons: a) herringbone; b) sandwich herringbone; c) γ and d) sheet (β).	110
6.5	Overlays of the global minima (green) from our validation CSPs with the experimental structures for TT (RSMD ₃₀ = 0.355 Å) and pentacene (bulk, RSMD ₃₀ = 0.393 Å).	112
6.6	Structure–energy landscapes for the predicted crystal structures of pentacene. (a) and (b) are the low–energy (within 15 kJ/mol above the predicted global minimum) lattice energy landscapes for pentacene and TT, respectively, where each point is coded with respect to its crystal packing type.	113
6.7	Structure–energy–mobility landscapes for the predicted crystal structures of azapentacenes with 5N. Colouring and the size of circles on the right–hand–side correspond to the magnitudes of calculated electron mobilities in cm ² /Vs.	115

6.8	Structure–energy–mobility landscapes for the predicted crystal structures of azapentacene with 7N, mobilities are given in cm^2/Vs	116
6.9	Sheet and γ herringbone directing interactions in the 3rd lowest 5A structure and the 5B global minimum	117
6.10	Sheet motifs and packing seen in the global minima of 7A,7B and 7C, showing triangular hydrogen bonding connecting parallel sheets and varying levels of offset.	118
7.1	Two potential encoding schemes for the GA on pentacene and 7A, (a) and (c) are SMILES strings, (b) and (d) are InChi key strings.	125
7.2	The input and output of the mutator function used to generate members of the initial population.	127
7.3	Ten members of a randomly generated initial population of 5N substituted pentacenes.	128
7.4	The atoms highlighted for the atomic site crossover and their position in the SMILES string.	130
7.5	Two azapentacene molecules with their crossover genomes.	131
7.6	Single-point crossover for the two azapentacene molecules from Fig 7.5.	131
7.7	Two-point crossover for the two azapentacene molecules from Fig 7.5.	132
7.8	Uniform crossover for the two azapentacene molecules from Fig 7.5.	133
7.9	Highlighted atoms showing where rings are separated in the ring crossover method.	134
7.10	Two molecules fragmented for crossover, * represents dummy atoms that are placed into the fragments where the molecule was cut.	135
7.11	(a) shows how dummy atoms are moved to create branched molecules, (b) shows some typical branched molecules produced.	136
7.12	The number of generations until pentacene is hit for all population sizes and runs.	138
7.13	The number of unique molecules until pentacene is hit for all population sizes and runs.	139
7.14	The minima located in our the searches of minimising the HOMO/LUMO energy difference and minimising the LUMO energy, values are in eV.	142
8.1	A plot of electron affinity versus reorganisation energy for all molecules encountered in a run of our GA. Points are coloured according the the amount of N atoms present in the molecule.	148
8.2	The best molecules for our search from 10 bins of 0.1 eV width across our preferred range of electron affinities (3.0 - 4.0 eV). The legend of each molecule is its reorganisation energy in eV.	149
8.3	A plot showing electron affinity versus fitness (as calculated from 8.2) for the molecules sampled in our initial run.	150
8.4	GA run 1 top 10 molecules. The number given below for each molecule is its fitness in eV (see equation 8.2).	152
8.5	GA run 2 top 10 molecules. The number given below for each molecule is its fitness in eV (see equation 8.2).	152
8.6	GA run 3 top 10 molecules. The number given below for each molecule is its fitness in eV (see equation 8.2).	153
8.7	GA run 4 top 10 molecules. The number given below for each molecule is its fitness in eV (see equation 8.2).	153

8.8	GA run 5 top 10 molecules. The number given below for each molecule is its fitness in eV (see equation 8.2).	154
8.9	The clustered best 10 molecules from all of our GA runs. The number below each molecule is the molecules fitness in eV (equation 8.2).	154
8.10	A plot showing electron affinity versus reorganisation energy for the clustered results of our 5 runs, circular points are linear molecules, triangles branched.	156
8.11	A zoomed-in version of Fig 8.10 showing where our top 10 molecules are found.	157
8.12	The fitness of the branched molecules sampled during the GA against their electron affinity. Most branched molecules have too low an electron affinity (<3.0 eV) and are penalised by our fitness function.	158
8.13	Two molecules expected to be sampled by the GA but were not in our top 10. They were located in the GA but fell just outside our electron affinity lower bound. The number under the molecule is their electron affinity in eV.	159
8.13	Structure–energy–mobility landscapes for the predicted crystal structures of GAMol1, GAMol2 and GAMol4, the number after GAMol refers to the rank of the molecule from the fitness function. Colouring and the size of circles on the right–hand–side correspond to the magnitudes of calculated electron mobilities.	162
8.14	Three common packing motifs seen across our predicted structures.	163
8.15	Structure–energy–mobility landscapes for the predicted crystal structures of GAMol6 and GAMol12, the number after GAMol refers to the rank of the molecule from the fitness function. Colouring and the size of circles on the right–hand–side correspond to the magnitudes of calculated electron mobilities.	164
8.16	Three of the high mobility crystal structures predicted for GAMol12.	165
8.17	The clustered best 10 molecules from all of our hexacene runs. The number underneath is the molecules fitness in eV.	167
8.18	A plot showing electron affinity versus reorganisation energy for the clustered results of our five hexacene runs, circular points are linear molecules, triangles branched.	168
8.19	A zoomed in version of Fig 8.18 showing where our top 10 molecules are found.	169
8.20	Structure–energy–mobility landscapes for the predicted crystal structures of GAMol3, GAMol5 and GAMol7, the number after GAMol refers to the rank of the molecule from the fitness function. Colouring and the size of circles on the right–hand–side correspond to the magnitudes of calculated electron mobilities.	171
8.21	Structure–energy–mobility landscapes for the predicted crystal structures of GAMol8, GAMol9 and GAMol10, the number after GAMol refers to the rank of the molecule from the fitness function. Colouring and the size of circles on the right–hand–side correspond to the magnitudes of calculated electron mobilities.	172

8.22 Structure–energy–mobility landscapes for the predicted crystal structures of GAMol11, the number after GAMol refers to the rank of the molecule from the fitness function. Colouring and the size of circles on the right–hand–side correspond to the magnitudes of calculated electron mobilities	173
--	-----

List of Tables

2.1	Molecular structures of the molecules chosen in the blind test	22
3.1	The main differences between band and hopping transport	35
4.1	The three common types of learning used to train machine learning algorithms.	47
4.2	Desired outputs of machine learning and the common algorithms used for these outputs.	49
5.1	Number of trial structures required to generate 10000 accepted crystal structures (50000 for $Z' = 2$) for each system. $Z' = 1$ unless otherwise stated. The number in parentheses is the number of accepted structures that lead to a successful lattice energy minimization.	84
5.2	Matches from the full CSP to experimentally determined structures of the observed polymorphs. RMSD ₃₀ is the deviation in atomic positions of a cluster of 30 molecules taken from predicted and experimental structures, not including hydrogen atoms. Non quinacridone results are also included. CC1 (<i>R3</i>) was generated in the <i>P1</i> spacegroup, which reduces to <i>R3</i> on account of intramolecular symmetry, hence the cell angles differ at the second decimal place. The experimental structures of CC1 also contained residual solvent, which was removed for purposes of comparison. All structures were converted to their reduced unit cell for comparison. Å and degrees are used throughout.	93
5.3	Matches from the full CSP to experimentally determined structures of the observed polymorphs. RMSD ₃₀ is the deviation in atomic positions of a cluster of 30 molecules taken from predicted and experimental structures, not including hydrogen atoms. Å and degrees are used throughout.	100
6.1	Lattice parameters (vectors in Å, angles in ° and d spacing of the three pentacene polymorphs in addition to the experimentally observed TT structure.	107
6.2	Matches from the full CSP to experimentally determined structures of the observed polymorphs. RMSD ₃₀ is the deviation in atomic positions of a cluster of 30 molecules taken from predicted and unoptimised experimental structures, not including hydrogen atoms. All structures were converted to their reduced unit cell for comparison. Å and degrees are used throughout.	113

6.3	Summary of the charge transport parameters for the azapentacene molecules investigated: λ_e is electron reorganisation energies, calculated at B3LYP/6-31G** level of theory. μ_{\max} is the maximum predicted electron mobility among the predicted crystal structures. $\Delta E(\mu_{\max})$ is the lattice energy gap between the crystal structure with the highest charge mobility to the predicted global minimum. $\langle\mu\rangle$ is the ensemble-averaged electron mobility across all crystals with calculated mobilities. μ_{gm} is the mobility of the global minimum.	119
7.1	Showing the decreasing amount of unique molecules for each population sizes longest run.	140
7.2	The averaged results for each population size for atomic site crossover method. GTP is generations until pentacene, MTP unique molecules until pentacene, CTP calculations until pentacene.	140
7.3	The averaged results for each population size for the ring crossover method. GTP is generations until pentacene, MTP unique molecules until pentacene, CTP calculations until pentacene.	141
7.4	The averaged results for each population size for the ring crossover method, using HOMO/LUMO energy difference as the fitness. GTM is generations until the minimum and MTM unique molecules until the minimum. . . .	141
7.5	The averaged results for each population size for the ring crossover method, using LUMO energy as the fitness. GTM is generations until the minimum and MTM unique molecules until the minimum.	142
8.1	A table showing the Tanimoto similarity scores for the six azapentacene molecules studied in Chapter 6.	146
8.2	The number of unique molecules sampled per run for our five runs. . . .	151
8.3	Summary of the charge transport parameters for the azapentacene molecules investigated: λ_- is electron reorganisation energies, calculated at B3LYP/6-311G** level of theory. μ_{\max} is the maximum predicted electron mobility among the predicted crystal structures. $\Delta E(\mu_{\max})$ is the lattice energy gap between the crystal structure with the highest charge mobility to the predicted global minimum. $\langle\mu\rangle$ is the ensemble-averaged electron mobility across all crystals with calculated mobilities.	166

Declaration of Authorship

I, **Josh E. Campbell** , declare that the thesis entitled *Crystal Structure Prediction of Organic Semiconductors* and the work presented in the thesis are both my own, and have been generated by me as the result of my own original research. I confirm that:

- this work was done wholly or mainly while in candidature for a research degree at this University;
- where any part of this thesis has previously been submitted for a degree or any other qualification at this University or any other institution, this has been clearly stated;
- where I have consulted the published work of others, this is always clearly attributed;
- where I have quoted from the work of others, the source is always given. With the exception of such quotations, this thesis is entirely my own work;
- I have acknowledged all main sources of help;
- where the thesis is based on work done by myself jointly with others, I have made clear exactly what was done by others and what I have contributed myself;
- parts of this work have been published as: Report on the sixth blind test of organic crystal structure prediction methods. *Acta Crystallographica Section B: Structural Science, Crystal Engineering and Materials*, 72(4), pp.439-459 and Convergence properties of crystal structure prediction by quasi-random sampling. *Journal of chemical theory and computation*, 12(2), pp.910-924.

Signed:.....

Date:.....

Acknowledgements

I would like to thank my supervisor, Professor Graeme Day, for his encouragement, advice, support and most importantly patience over the course of my Ph.D and help in completing this Thesis. In addition I would like to thank members of the Day group who have helped me over the years specifically Pete and Dave for helping me find my feet after I started. Personal acknowledgements go to Maya, Ron and Otis, who have been there for me every time I have needed help.

Abbreviations

CSP	Crystal Structure Prediction
GA	Genetic Algorithm
OS	Organic Semiconductor
CCDC	Cambridge Crystallographic Data Centre
OLED	Organic Light Emitting Diode
OFET	Organic Field Effect Transistor
QM	Quantum Mechanical
FF	Force Field
DFT	Density Functional Theory
DMA	Distributed Multipole Analysis
AIM	Atoms In Molecules
MEP	Molecular Electrostatic Potential
DFT-D	Dispersion corrected DFT
LUMO	Lowest Unoccupied Molecular Orbital
HOMO	Highest Occupied Molecular Orbital
PAH	Polycyclic Aromatic Hydrocarbon
FDE	Frozen Density Embedding
ADF	Amsterdam Density Functional
SVM	Support Vector Machines
OLS	Ordinary Least Squares
NN	Neural Network
kNN	k-Nearest Neighbours
FPS	Fitness Proportionate Selection
SUS	Stochastic Universal Sampling
SAT	Separating Axis Theorem
TVP	Target Volume Parameter
MVP	Maximum Volume Parameter
LDP	Lattice Deformation Parameter
GLEE	Global Lattice Energy Explorer
TT	Tetraazatetracene

Chapter 1

Introduction

The overall aims of the work presented in this thesis is the application of crystal structure prediction methods for the discovery of promising molecules for organic semiconductor applications. The crystal structure of any molecule plays an important role in determining the properties of the solid state. The arrangement of molecular units in the crystal is determined by the intermolecular interactions present. These interactions (such as hydrogen bonding) are in turn a product of the electron density of the molecule. The electron densities of different molecules interact, in either a repulsive or attractive manner, until an equilibrium geometry is reached. For some systems there will be a clear "winner", the global lattice energy minimum significantly lower in energy than any other possible structures. However it is much more likely that there will be other structures close to the global minimum (within a few kJ/mol¹) which could be possible polymorphs^{2;3}. Polymorphism was defined in 1965 by McCrone as "a solid crystalline phase of a given compound resulting from the possibility of at least two different arrangements of the molecules of that compound in the solid state"⁴. If the molecule can also adopt different conformations within the different crystal structures, it is then an example of conformational polymorphism⁵.

A study of available crystal structure databases estimates that half of all organic molecules exhibit polymorphism⁶. As many bulk properties arise from the crystal structure, different polymorphs of the same molecule can exhibit wildly different properties. The most well known example is Ritonavir⁷, an inhibitor of the HIV protease. A late appearing polymorph drastically changed the solubility (due to the different free energies of the polymorphs) leading to a recall and investigation⁸. Polymorphism is also important in organic semiconductors, with the three known polymorphs of pentacene all display different charge carrier mobilities⁹.

It can be seen knowing the possible polymorphs of a molecule is of great importance to many industries and this is where crystal structure prediction (CSP) enters. CSP methods aim to predict the possible crystal structures of a molecule from first principles

and rely on the computation or approximation of the electron density which allows calculation of the lattice energy. The description of the interactions differs with the method chosen, but generally CSP involves the search for the global minimum in the lattice energy. This is usually a pure thermodynamic model at 0K: the static lattice energy, disregarding free energy and disregarding the kinetics of crystal growth. CSP can help reassure experimentalists that the polymorph they have is the thermodynamically stable one, or suggest alternatives that could be produced by different experiments. The ultimate aim of CSP is to predict a crystal structure *ab initio*. In practical terms this means beginning from a molecular diagram. The question "Are crystal structures predictable" was posed nearly 20 years ago¹⁰. The answer then was "No" but progress has been made, enough to upgrade to "Maybe"¹¹ in 2003. Chapter 2 contains a more in depth analysis of the improvements to CSP methods since their inception and the phenomenon of polymorphism. Effectively sampling the crystal energy landscape is an important part of the CSP process and work on testing a new structure generator is presented in Chapter 5 along with a CSP study of a molecule from the 2015 CCDC blind test.

Organic semiconductors (OS) are the driving force behind the development of organic electronics. The ability to deposit organic films on a wide variety of substrates has led to flexible displays, printable circuits and plastic solar cells.^{12;13} Major electronic companies such as Samsung, LG and Sony offer OLED (organic light emitting diode) displays powered by OFETs (organic field effect transistors). Organic semiconductors can be broadly split into two categories: small molecules (usually polycyclic aromatic hydrocarbons and their derivatives) and conjugated polymers. Molecules such as pentacene and rubrene are the focus of intense research due to performance approaching that of inorganic semiconductors¹⁴. A review of the growth of the OS field and the parameters governing charge transport can be found in Chapter 3. Charge mobility can be tuned by the rational design of novel molecules. In Chapter 6, the CSP of six designed molecules (and two validation molecules) is presented and their charge mobilities analysed.

Designing molecules by hand can quickly become a drawn-out process as the parameter space becomes larger and more properties are to be taken into account. This is where the third part of this thesis becomes relevant; the use of a genetic algorithm to design and optimise novel organic semiconductors. Genetic algorithms (GAs) have found use in many fields that rely on global optimisation¹⁵. GAs mimic the process of natural evolution such as inheritance, mutation, selection and crossover. A population of possible solutions is generated and then scored by some fitness function. Each candidate solution has a set of characteristics (its chromosome) which can be mutated and altered. Once ranked and selected fit solutions are paired up. These two solutions are then combined in a way to produce two child solutions. All the new child solutions then make up a new population ready to be evaluated for fitness. The GA then continues until some finishing condition is met and a minimum is found. A review of GA performance and key concepts

is found in Chapter 4. The design and application of our GA is presented in Chapter 7. Convergence and performance were measured by optimising a range of molecular properties with known minima. Chapter 8 contains the application of the GA to OS design. Properties needed for good charge mobility were optimised and fittest molecules taken for CSP. The introduction of CSP into the fitness function is also presented there.

Chapter 2

Molecular Crystals and Crystal Structure Prediction

2.1 Crystals and Polymorphism

The simplest unit in chemistry is the atom. Atoms form bonds through orbital overlaps which allow electrons to be shared creating strong, predictable bonding patterns. The understanding of how atoms bond allows the design of new molecules by synthetic chemists.

Molecular crystals are formed from the interactions of molecules. Molecular orbital overlap and the interactions between the electron densities of the molecules allows the formation of motifs which then can pack to produce solid forms. A crystal differs from other solids in that it shows long range order, where the motifs are repeated by symmetry operators, the key symmetry in a crystal is translational symmetry. The bonds between molecules are much weaker than the ones between atoms, for example the covalent bond between oxygen and hydrogen is about 25 times stronger than the hydrogen bonding between water molecules. This comparative weakness leads to the unpredictability of many intermolecular interactions. In a molecule with many hydrogen bond donors and acceptors it can be difficult choosing the most likely motif to form. Thus chemists looking to engineer crystals find more difficulties than their molecule building counterparts.

With many possible intermolecular interactions molecules can pack in many different ways, if this leads to multiple crystal structures of the same molecule, the molecule exhibits polymorphism, with each crystal structure referred to as a polymorph. There are two types of polymorphism, packing and conformational. A packing polymorph is where the molecule maintains its conformation across the polymorphs, but packs in different ways in each. A conformational polymorph involves a change in the molecular geometry between polymorphs.

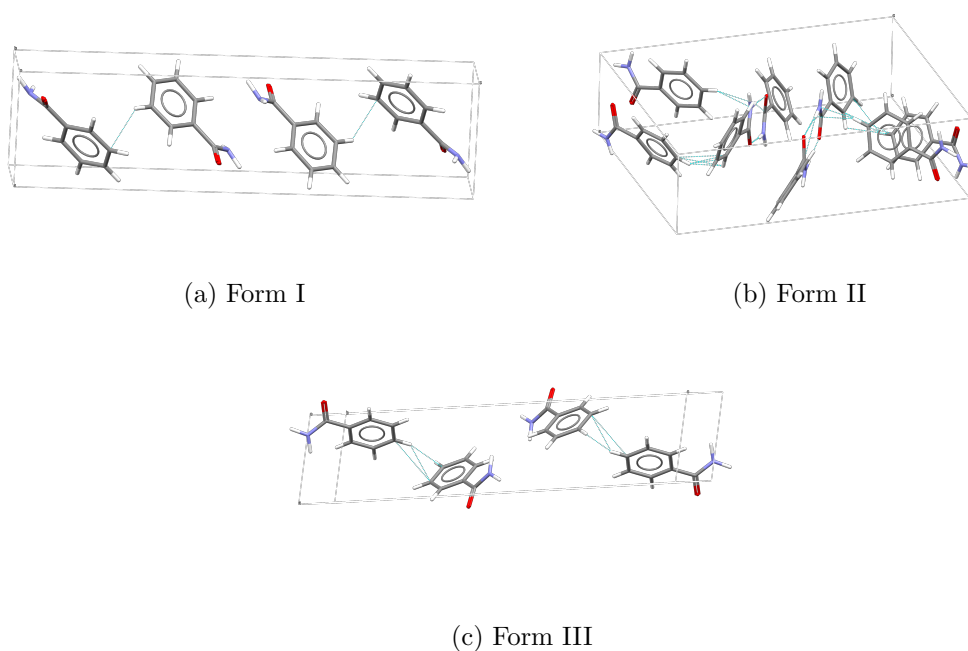


Figure 2.1: The three known polymorphs of benzamide (CSD refcode BZAMID), form I is the stable polymorph, form II metastable discovered in 2005 and form III the silky needles first seen in 1823

With rigid molecules only packing polymorphs are possible, but it is perfectly possible for flexible molecules to have both packing and conformational polymorphs including crystals where the molecule may adopt different conformations within the same structure.

Polymorphism was defined in 1965 by McCrone⁴ but was first discovered in 1832¹⁶. Two German chemists upon studying the cooling of boiling benzamide witnessed at first, the formation of silky needles, which were then slowly replaced by rhombic crystals. Demonstrating the struggles to identify and reproduce polymorphs the stable form was first characterised in 1959¹⁷, a new metastable form in 2005¹⁸ and finally the silky needles were seen again in 2007¹⁹. The differences between the polymorphs is primarily in the π - π interactions and can be seen in Fig 2.1. Benzamide is far from the only example of latent polymorphism, in 2006 a new polymorph of maleic acid was discovered 124 years after the stable polymorph was studied²⁰. This new polymorph was accessed by attempted co-crystal formation and dissolution; in a similar vein two new polymorphs of 1,3,5-trinitrobenzene were characterised with the use of an additive in the crystallisation 125 years after the molecule was first used as an explosive²¹.

The incidence of polymorphism can be hard to pin down. For much of the 20th century resolving a crystal structure was laborious work. Data collection could take days, while growing a good enough single crystal could take weeks. Structures with large amounts of disorder and those with more than one molecule in the asymmetric unit were intractable.

Combined, this meant the characterisation of only the most stable polymorph was justifiable. Cruz and Bernstein evaluated datasets from the Cambridge Structural Database (CSD), pharmaceutical companies, the European Pharmacopoeia and McCrone's early studies to arrive at a minimum polymorphism occurrence of 50%⁵, with the caveat that it could be much higher. As McCrone stated, "every compound has different polymorphic forms, and that, in general, the number of forms known for a given compound is proportional to the time and money spent in research on that compound"⁴.

The poster children of polymorphism are Ritonavir²² (Fig 2.2a) and ROY²³ (Fig 2.2b). Only one crystal form of Ritonavir was identified during its development, but after two years on the market some tablets began to fail dissolution tests. It was discovered that the original crystal structure would convert on contact with a new low solubility polymorph due to the inactive forms lower energy. The new polymorph was quickly discovered to have a much reduced solubility, resulting in the reformulation of the drug. ROY (named for its red, orange and yellow polymorphs) currently holds the record for the most observed polymorphic forms in the Cambridge Structural Database (CSD), the structures of seven polymorphs are known and three more have been observed, but have unknown structures²⁴.

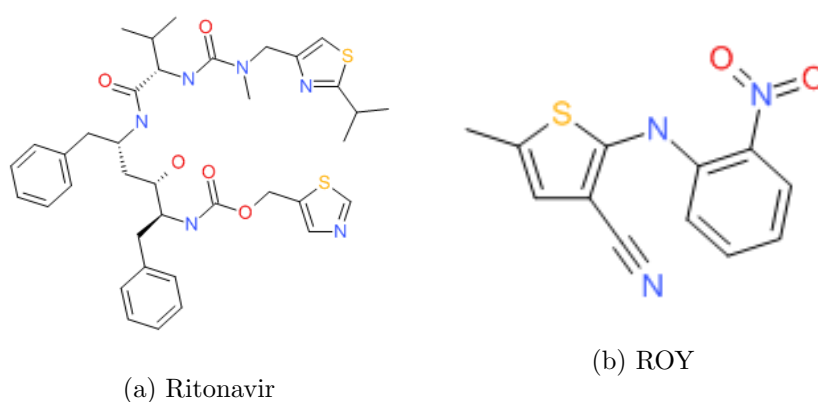


Figure 2.2: The two poster children for polymorphism. The lack of knowledge of an additional ritonavir polymorph led to the withdrawal of the drug. ROY currently has the most known polymorphs.

Polymorphism is relevant to all industries that rely on crystalline products, as different polymorphs can exhibit radically different properties. Typically industries such as pharmaceuticals, pigments/dyes, organic electronics and explosives will be actively engaged in polymorph screening. Many properties of the bulk material arise from molecular positions and orientations in the crystal. For example, the differences in solubilities between polymorphs is a product of the difference in free energy between the crystal structures. As seen from the example of ritonavir, polymorph screening is very important in the pharmaceutical industry. For a drug to act in a patient it must first make it into the patient and as mentioned above polymorphs may have wildly different dissolution rates. A simpler drug example is paracetamol. Paracetamol has three known

polymorphs, however the most stable form has unfavourable compaction and tableting properties requiring the need for large amounts of excipient²⁵. The metastable second form however has desirable properties for tableting but is far too unstable. Eventually a co-crystallisation study²⁶ found a co-crystal that combined the structural features (and therefore tableting properties) of form II and the stability of form I.

In organic semiconductors electron mobility in part depends on the level of molecular wavefunction overlap between molecules and the modification of crystal structures is an important part of the design of new semiconducting materials. 6,13-bis(triisopropylsilyl)ethynyl)pentacene (TIPS-pentacene Fig 2.3a) has three known thin film polymorphs with even small changes in packing showing a change in mobility of many orders of magnitude²⁷. Other polymorphic semiconductors that show large differences in mobility between polymorphs include rubrene²⁸(Fig 2.3b), tetrathiafulvalene²⁹ (THF, Fig 2.3c) and contorted hexabenzocoronene³⁰ (CHBC, Fig 2.3d). The interplay of crystal structure and charge mobility will be discussed in depth in Chapter 3.

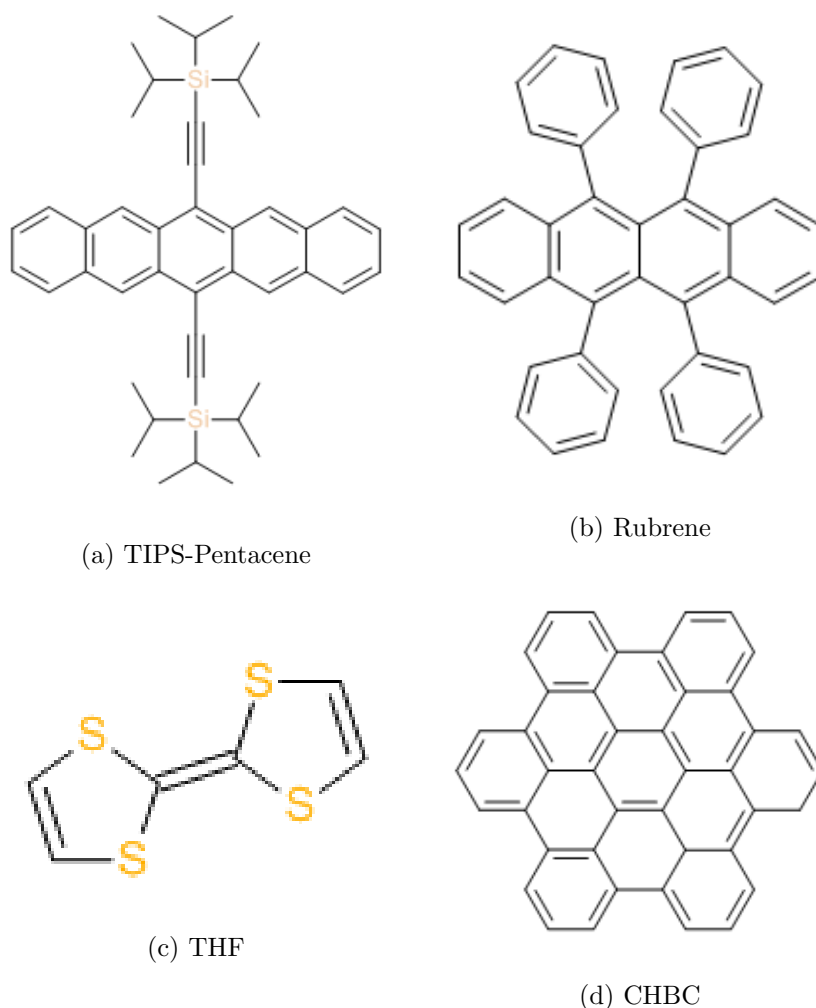


Figure 2.3: Four organic semiconducting molecules with large mobility differences between polymorphs

The performance of pigment molecules also depends on their crystal structure, including hue and photostability³¹. Naphtol red (Fig 2.4a) an important pigment in automotive paints and coatings has several polymorphs. When produced industrially an important step in the process is the conversion from the α form to the γ form. The γ form is much denser, which is an attractive property for pigments as this makes them more resistant to photodegradation³². Other polymorphic pigments include quinacridone (Fig 2.4b). Quinacridone and its derivatives comprise a large high performance pigment family. Quinacridone itself has four known polymorphs with the β and γ forms more prized for industrial applications. Quinacridone and its polymorphs will be discussed in more detail in Chapter 5.

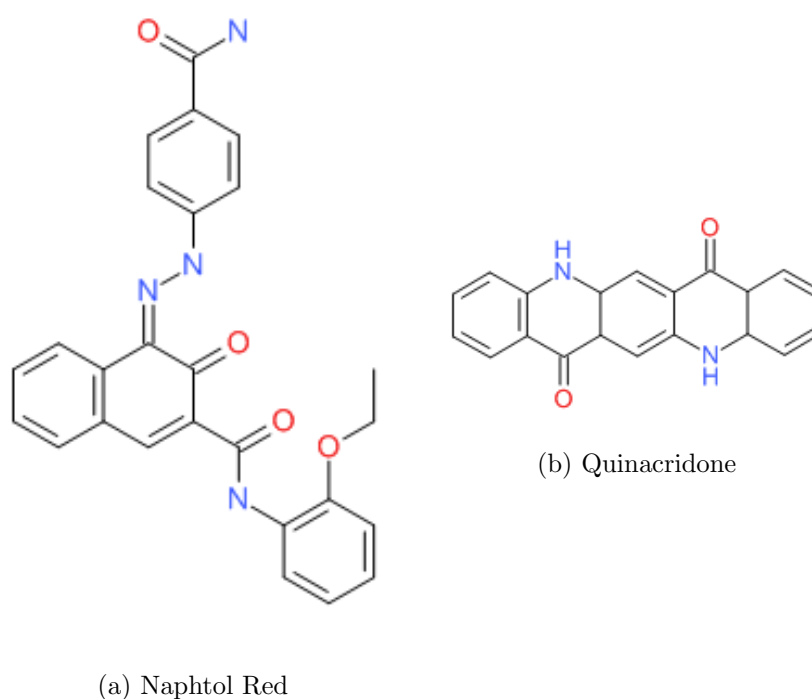


Figure 2.4: Two pigment molecules known to exhibit polymorphism, with some particular polymorphs being much more useful as pigments

Many energetic materials also exhibit polymorphism. Octogen (also known as HMX, Fig 2.5a), one of the most widely studied nitramine explosives has four known polymorphs. Density as expected is an important factor in explosive performance; the more material you can pack into the charge the larger the explosion will be. This influences the impact sensitivity of the polymorphs. A lower value means the material is more sensitive to impacts, an undesirable property in the production of the explosive. The density of the polymorphs follows $\beta > \alpha > \gamma > \delta$, while the impact sensitivity is reversed $\delta > \gamma > \alpha > \beta$ ³³. Trinitrotoluene's (Fig 2.5b) crystallisation and polymorphism was poorly understood until the late 70s (when published in a classified report³⁴, still not declassified to this day) and the lattice parameters were not published in full until 1997³⁵.

TNT exists in two crystalline forms which are very similar to one another, and indeed transformation from the metastable to stable has been shown suggesting the metastable form is a product of stacking defects when the stable form crystallises³⁶.

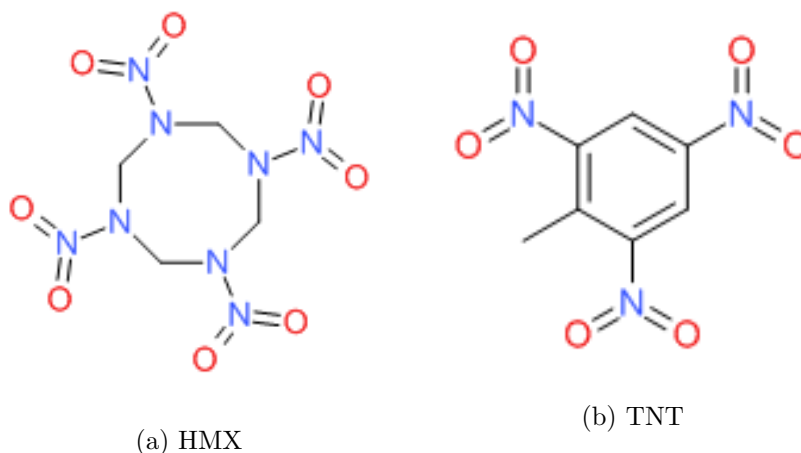


Figure 2.5: Two explosive molecules known to exhibit polymorphism

2.2 Crystal structure prediction

The importance of polymorphism shows the need to be able to predict possible crystal structures of a molecule. Knowing possible polymorphs before crystallisation allows the calculation of properties of interest and opens up the possibility of computer guided design of materials. The long-term goal of crystal structure prediction (CSP) is to propose all possible crystal structures of a molecule, given no information apart from the structural formula.

The main steps in *ab initio* crystal structure prediction can be seen in Fig 2.6. At its most basic, CSP can be split into three main steps:

1. the generation of an optimised 3D structure of the molecule
2. the generation of all the possible crystal structures for the structure of the molecule
3. the ranking of structures (usually by lattice energy), evaluation of properties and inspection of the lowest energy structures (including comparison to observed structures)

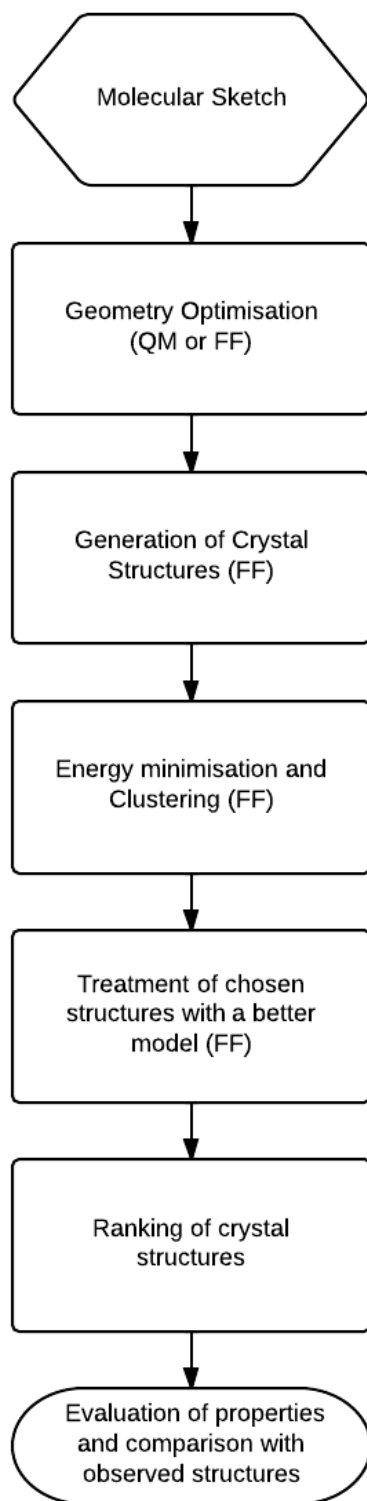


Figure 2.6: The steps in a crystal structure prediction (QM = quantum mechanical, FF = force field)

2.2.1 Molecular model

Ab initio crystal structure prediction begins with a 3D model of the atom positions and connectivity of the molecule. This is usually geometry optimised either with a force field (FF) or quantum mechanical (QM) methods. For most small rigid molecules, no further information needs to be supplied. Crystal packing forces do not usually have a magnitude big enough to affect the intramolecular properties (such as bond lengths and angles) of the molecule. So the geometry of the isolated molecule is a good place to start.

Molecular flexibility introduces difficulties into the prediction. Most CSP software packages were designed to treat rigid molecules, and allow relaxation of the geometry only when the final lattice energy is evaluated. The changes in molecular geometry in this step are small as the molecule is already in a crystal environment, so conformational energy barriers are unlikely to be crossed. This means for molecules with more than one distinct conformation, each must be treated separately, i.e. the entire process is repeated for each unique equilibrium conformation. FF methods that include torsional terms can be used to model flexibility. However the energies calculated are usually much less accurate than those from QM methods. Usually there are many crystal structures predicted within a small range (a few kJ/mol) of the global minimum so small changes in these relative energies can radically alter the ranking of crystal structures².

2.2.2 Crystal structure generation

The next step after the generation of our molecular model is to explore as much of the lattice energy surface as possible generating thousands of structures. Crystal structures of rigid molecules have six degrees of freedom in the unit cell dimensions (lengths and angles) and another six (orientation and position) per molecule in the unit cell. Even with the rigid body approximation the size of the problem quickly grows with the number of molecules and becomes unmanageable. However 80% of all known organic molecular crystal structures have one (or less) molecule in the asymmetric unit and are found in one of six space groups ($P2_1/c$, $P\bar{1}$, $P2_1$, $P2_12_12_1$, $P1$ and $C2/c$)³⁷. This simplifies the problem greatly, as searching certain space groups places limitations on the positions of the molecules (through the symmetry operations that generate other molecules) and the dimensions of the unit cell (certain space groups set limits on unit cell angles or lengths). This does usually mean space groups are searched independently but this is not a problem. Splitting the search up reduces the overall cost when compared to a full search using no symmetry or space group information. However 10%³⁸ of molecules crystallize with a $Z' > 1$ (Z' being the number of molecules in the asymmetric unit of a crystal structure) and there are 230 possible space groups to search. Many algorithms exist that can perform these searches and there are three common "families" of methods.

A much more detailed analysis on structure generation will be presented in Chapter 5 with the development and testing of a new structure generator.

2.2.2.1 Systematic grid searches

Systematic grid searches are useful for structures with fewer degrees of freedom³⁹. The surface is split into a grid and parameter(s) are varied by a set amount. But as dimensionality increases so does the computational expense of this method. While all methods show an increase in computational expense with increasing dimensionality, grid-based methods have a very steep increase. Also the use of such a method usually requires the number of grid points to be chosen *a priori*, which can introduce problems in deciding computational resources. The parameters that are searched vary between algorithms. PMC⁴⁰ and UPACK³⁹ search across the unit cell parameters and space groups. MOLPAK⁴¹ generates a coordination sphere around a central molecule, which has its orientation systematically varied to produce the densest packing. PROMET⁴² works in a similar way but systematically adds translational symmetry to generate stable molecular clusters.

2.2.2.2 Random searches

Random searches can deal with more complex problems but often they can miss low energy structures so need to be repeated many times. The use of quasi-random number generators can improve the searching of the space by ensuring a uniform distribution of points⁴³. Della Valle *et al*⁴⁴ used such a method to explore the crystal structures of pentacene. Karamertzanis⁴⁵ used Sobol sequences⁴⁶ to generate quasi-random lattice lengths and angles. Our in house structure generator also uses the Sobol sequence in a similar way⁴⁷. Other modifications on the random search include the use of Monte Carlo (MC) simulations. Random moves are made around the potential energy surface and the new crystal structure is accepted based on a probability that is related to the energy change associated with that move. A variation of this is found in the Materials Studio Polymorph Predictor^{48;49} software, which combines Monte Carlo with simulated annealing to sample the entire search space over many runs. CRYSTALG⁵⁰ uses a MC method that makes steps between families of conformations rather than individual structures allowing it to sample more of the low energy space than random sampling alone.

2.2.2.3 Genetic Algorithms

Genetic algorithms have found use in many fields that rely on global optimisation¹⁵ so it is not surprising that methods exist that use them in CSP. Genetic algorithms mimic the

process of natural evolution by assigning the descriptors of the crystal to genomes which then undergo operations designed to produce the best structure. MGAC⁵¹ minimises the lattice energy of the crystal and its genome is made up of molecular positions, orientations, unit cell angles and flexible dihedrals. The RANCEl program uses a genetic algorithm⁵² that compares atom-atom distance distributions of similar structures to the one of interest, then generates structures that best fit the distributions. USPEX⁵³ uses an *ab initio* or FF lattice energy calculation of the generated crystal as its fitness function and combines slices of each structure to generate children.

2.2.3 Ranking Structures

After the structures are generated, they are often clustered to remove duplicates and properties and energies are evaluated. There are many clustering methods such as COMPACK⁵⁴, radial distribution functions⁵⁵, comparison of reduced cells⁵⁶ and the comparison of simulated powder patterns⁵⁷. An in house method of clustering using radial distribution functions has been developed⁴⁷. Two methods of clustering will appear in this report. Clustering *via* radial distribution functions using our in house code, and COMPACK will be used for clustering and matching after final lattice energy minimisations. Space group searches and the first overall clustering will be made by our code and COMPACK for a final overall clustering, as it is possible some matches may be missed. COMPACK creates a cluster of molecules (the number is user defined) around a central reference molecule and interatomic distances are calculated between nearest neighbours. The set of interatomic distances describes the cluster. If the first molecule in both clusters match, another molecule is searched for in the cluster, but it must be connected to a match already found via one of the interatomic distances. COMPACK can miss matches, especially if the size of the molecular cluster is low or particularly anisotropic. $Z' = 2$ structures also present problems as the decision over which molecule in the asymmetric unit is the starting point for the cluster, and the clusters will differ depending on which one is chosen. The choice of method is important; it must be fast and be able to distinguish between similar structures.

The observed experimental structure will not always be the lowest in energy but hopefully should be close. Due to many possible crystal structures being found within a small range of the global minimum, the evaluation of energies must be as accurate as possible. Lattice energy minimisations occur at various stages throughout the CSP run. Often thousands of structures will be treated. The trade-off is one between speed and accuracy, and typically force field based methods are used. Usually a cheap method is used in initial structure generation and minimisation, good candidate structures are then refined using a more expensive force field. If the molecule is treated as rigid only intermolecular interactions need to be considered as the intramolecular energy will be

the same as when the molecule was first geometry optimised. Including molecular flexibility necessitates the inclusion of accurate intramolecular energies, which has seen some success when applied to moderately flexible molecules⁵⁸.

2.3 Intermolecular Interactions and Force Field Methods

Due to the need to calculate the lattice energy for possibly 100,000s of crystal structures, computationally cheap force field methods are the most widely used in CSP. The total lattice energy consists of two components, the intramolecular energy (U_{intra} , the energy of the bonds and torsions within a molecule) and the intermolecular energy (U_{inter} , the interactions between separate molecules in the crystal). The intermolecular potential can initially be written as the sum of atom-atom terms.

$$U_{MN} = \sum_{m \in M} \sum_{n \in N} u^{mn}(R^{mn}, \Omega^{mn}) \quad (2.1)$$

Where the sum is taken over atoms m of molecule M , atoms n of molecule N and the interatomic interaction, u^{mn} , is dependent on the separation, R^{mn} and the relative orientation, Ω^{mn} , of the atoms. This is the pairwise additive approximation. The total potential is the sum of all the two-body interactions in the system. The most common force field methods will include terms for the van der Waals forces and electrostatic interactions using the pairwise additive approximation.

$$U_{mn} = U^{vdW} + U^{el} \quad (2.2)$$

Then U_{latt} is the sum over all molecule-molecule interactions in the crystal.

$$U_{latt} = \frac{1}{2} \sum_{MN}^{N_{mol}} U_{MN} \quad (2.3)$$

Before discussing common functional forms it is helpful to examine the individual contributions to the potential, which can be split into long-range and short-range categories.

2.3.1 Long range intermolecular interactions

Long range terms are important when there is no significant overlap of the charge density of the individual molecules. The three most important long-range terms are electrostatics, dispersion and induction.

The electrostatic term arises from the classical interaction of the charge distributions of the individual molecules. It can be attractive or repulsive and is pairwise additive.

The simplest (and most popular) way of describing electrostatic interactions is through the use of isotropic point charges, centred on the atoms in the molecules. However, this method has its limitations as a set of atomic point charges is not a realistic description of the charge distribution of a molecule. While atomic point charges create an electrostatic potential they do not replicate important features of it, such as lone pairs. More accurate methods including Distributed Multipole Analysis (DMA)⁵⁹ and Atoms in Molecules (AIM)⁶⁰ have been developed to try and recover the anisotropic behaviour of the electrostatic field.

Dispersion is a non-classical attractive force that is produced from the correlated movement of electrons in interacting molecules. The charge density of a molecule is always in flux and the movements become correlated to favour lower energy configurations. Dispersion can be described at long ranges by a series in the intermolecular separation R :

$$U_{disp} = -C_6R^{-6} - C_7R^{-7} - C_8R^{-8} - \dots \quad (2.4)$$

But this is impractical for anything but the smallest molecules, as a true description of molecule-molecule interactions would require all orientation dependence to be captured in the C coefficients so, it is replaced by a sum of atom-atom terms of a similar form. The dispersion coefficient C depends on the type of atoms involved and their relative orientation though this is often ignored and the series truncated to just $-C_6R^{-6}$.

The induction term appears due to the reaction of a molecule to the electric field of all its neighbours and is always attractive. Induction arises from interactions between rotating permanent dipoles and from the polarizability of atoms and molecules (induced dipoles). These induced dipoles occur when one molecule with a permanent dipole repels another molecule's electrons. A molecule with permanent dipole can induce a dipole in a similar neighboring molecule and cause mutual attraction. The fields generated by neighbours may cancel out or reinforce so the induction term is non-additive. This makes computation of the term expensive and most potentials only include induction effects in an average way. There are other effects such as resonance and magnetic that can arise at long range but are not applicable to closed-shell molecules in their ground states. For molecules with unpaired electrons in the ground state these terms can become important.

2.3.2 Short range intermolecular interactions

Short-range interactions are important when the charge density of two molecules overlap and electron exchange becomes important. The most important term at short range is exchange-repulsion, which consists of an attractive and repulsive effect. The attraction

arises from the electrons being able to spread out over two molecules, which introduces more uncertainty in their positions and allows the energy to drop. This effect is far outweighed by the repulsive interaction. The Pauli exclusion principle forbids electrons with the same spin from occupying the same space simultaneously, which costs energy, leading to the term being repulsive overall. This term is often described with atom-atom terms such as in the Lennard-Jones (R^{-12}) or Buckingham ($\exp(-BR)$) potentials. The repulsive region of the Buckingham potential is described in the Born-Mayer form below, with the repulsion decaying exponentially.

$$u_{er} = Ae^{-BR} \quad (2.5)$$

Where A and B are suitable constants. Charge transfer also plays a role in the intermolecular interactions at short range as electron density can move from one molecule to another. This can be viewed as a separate effect but is the short range component of the induction energy. Charge transfer also decays exponentially with intermolecular separation so is described in the repulsion term of the Buckingham potential. It can be separated out of the overall induction term but this is not usually done.

2.3.3 Relation to physical contributions to the intermolecular interaction

While the splitting of forces into long and short range is a mathematical formalism, standard physical interactions arise from a combination of these forces. The van der Waals force is the sum of the interactions between molecules not created by the interaction of permanent charges or multipoles. It is present in all phases of matter and arises from a combination of the dispersion, exchange-repulsion and induction terms. Hydrogen bonding is the electrostatic attraction between polar groups that occurs when a hydrogen atom bound to a highly electronegative atom (such as N, O or F) experiences attraction to some other nearby highly electronegative atom. Hydrogen bonding is generally the strongest intermolecular interaction, their reliance on specific atoms means they are highly directional and arise from a combination of electrostatic, induction, dispersion, exchange-repulsion and charge transfer terms. π - π interactions arise from the influence of one aromatic system upon another, they are attractive and the size of the π system determines the strength of the interaction. Being similar to van der Waals they come from the same set of forces: dispersion, exchange-repulsion and induction (in polar π systems).

2.3.4 Functional forms

u^{vdW} is very large positive at short distances, has a minimum that corresponds to the atoms just being in contact and approaches zero as interatomic distances become large. A widely used function that meets these requirements is the Lennard-Jones potential⁶¹. The most common form is as seen below in equation 7. The shape of the potential at close distances can be seen in Fig 2.7.

$$u_{mn,vdW} = \frac{A}{R_{mn}^{12}} - \frac{B}{R_{mn}^6} \quad (2.6)$$

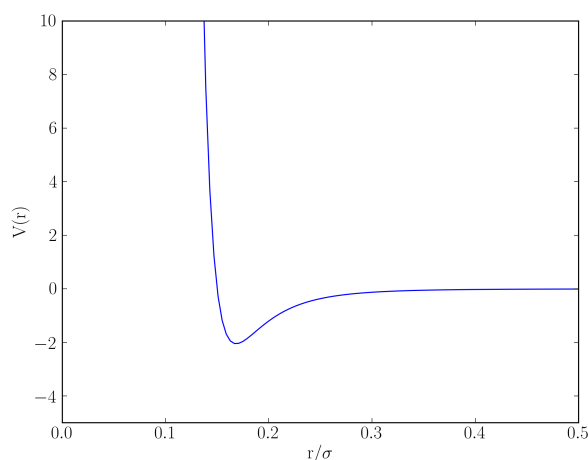


Figure 2.7: Lennard-Jones potential

A and B are constants that are fitted for the atom types present in the molecules, obtained from fitting to *ab initio* calculations or experimental data. The R^{12} term is chosen for computational convenience and in fact the repulsive region is better described as an exponential which features in the Buckingham potential (exp-6) (Eq 2.7)⁶². A plot of the Buckingham potential can be seen in Fig 2.8. The main drawback of the potential is apparent as the interatomic distances become small the potential will allow the nuclei to become strongly bonded together. However this modelling of the repulsive wall is more applicable to crystals where close contacts are common. There is also a increase in computational cost when compared to the Lennard-Jones functional.

$$u_{mn,vdW} = Ae^{-BR} - \frac{C}{R^6} \quad (2.7)$$

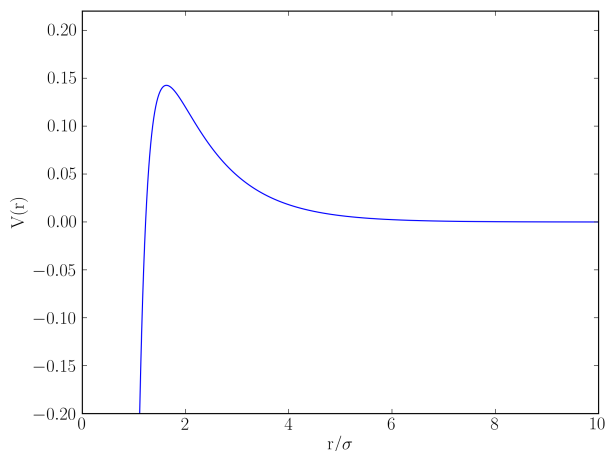


Figure 2.8: Buckingham potential

A, B and C are constants that are derived from empirical parametrisation. The parameters are fitted to reproduce data from crystal structures, sublimation enthalpies and vibrational frequencies.^{63;64} Empirical parametrisation does have limitations. It is important to choose the set that was fitted to molecules that are closest to the molecule of interest. Also, more general force fields cannot possibly include every atom type so molecules with uncommon bonding or atoms will not be reproduced well. However some of the many-body effects ignored in the pairwise approximation can be recovered in an averaged way, as well as parts of ignored interactions through parametrisation to experimental data.

u^{el} is the other non-bonded interaction that is considered in simple models. As mentioned above the most common model is *via* the assigning of point charges to atom centres. These point charges are isotropic (i.e the interactions only depend on distance) so any anisotropic character (lone pairs, π bonding) is poorly reproduced. The interaction between the point charges is given by the Coloumb potential as seen below.

$$u_{mn,el} = \frac{q_m q_n}{4\pi\epsilon_0 R_{mn}} \quad (2.8)$$

Where q are the point charges on atoms m and n , ϵ_0 is the dielectric constant and R the interatomic separation. Many force fields come with transferable charges that were used in the parameterisation but the easiest method of improving them is to use charges that reproduce the electrostatic potential (ESP) around the molecule. These can be obtained from a QM calculation using the CHELPG⁶⁵ scheme. Distributed multipole analysis⁵⁹ can further improve the description of the ESP. At each site (usually atomic centres) there are sets of multipoles which consist of charges, dipoles, quadrupoles and higher terms. DMA analysis features in the DMACRYS⁶⁶ program and has shown good results across a variety of problems^{67;68;69}. However atomic multipoles are highly

conformationally dependent so their use for flexible molecules is limited. The SCDS method of Gavezzotti⁷⁰ can also be used for higher accuracy. The QM calculated electron density around a molecule is split into points/pixels and a direct summation over them is used to calculate the energy terms. Each pixel is assigned a local polarizability using the average polarizability of the nearest atom. Repulsion energies are calculated as the overlap of the electron densities of the molecules, which is scaled proportionally by user defined parameters. The quality of this method depends strongly on the quality of the numerical integration, the original QM calculation and the choice of parameters for the repulsion energies and atomic polarizabilities.

2.3.5 CSP methods used in this thesis

The sections above lay the groundwork for the choices of methods used within this thesis. The first step in CSP is the molecular optimisation, U_{intra} is unimportant in rigid molecule CSP and the electron density must be calculated accurately for use with multipoles. A molecular model will be geometry optimised in the gas phase with density functional theory implemented in the Gaussian 09 code⁷¹, using the B3LYP functional⁷² and the 6-31G** basis set⁷³. The * signifies the addition of polarization functions on the atoms, with ** adding these functions to H and He as well. These additional functions can improve the accuracy of the calculation, by allowing for asymmetry in the atomic orbitals.

In addition to the functional form and choice of electrostatic method a forcefield with suitable parameters needs to be chosen. In crystals close contacts are the norm, so the Buckingham potential will be used. As forcefield parameters are usually non-transferable, a forcefield developed with molecular crystals in mind would be best. The W99 potential meets both of these prerequisites and is used for all atom-atom interactions^{74;75;76}. The full functional form of W99 is given as:

$$U_{ik,inter} = B\exp(-Cr_{jk}) - Ar_{jk}^{-6} + q_iq_kr_{jk}^{-1} \quad (2.9)$$

The first term is the exchange-repulsion energy, the second dispersion and the third electrostatics. W99 was parametrised to reproduce energies and geometries of known crystal structures. W99 was originally parametrised for use with atomic partial charges fitted to the molecular ESP but as mentioned above these are often non transferable and present problems with anisotropic features of the molecular ESP. DMA allows for the capture of these anisotropic features and will be used in place of the charges. The electron density is defined by Gaussian functions, which are a product of radial Gaussians and spherical harmonics. The molecular orbitals are then constructed as linear combinations of these atom-centred Gaussians. DMA⁵⁹ analyses the electron density matrix from any

wavefunction, though we will be using the density matrix from the initial molecular optimisation. The density matrix from Gaussian orbitals is given below:

$$p(\mathbf{r}) = \sum_{tu} p_{tu} \phi_t^A(\mathbf{r}) \phi_u^B(\mathbf{r}) \quad (2.10)$$

where ϕ_t^A is a Gaussian function centred at A. When represented by a Gaussian basis set the product of primitive functions is also a Gaussian. For example the product of two *s* functions can be represented by a point charge at the overlap point. The combination of a *s* and a *p* orbital has charge and dipole moments, while two *p* orbitals have charge, dipole and quadrupole. These overlap points are then moved back to atomic centres, giving a multipole series at each atom. Lattice energy minimisations using multipoles are possible in the program DMACRYS⁶⁶ which will be used for all lattice energy minimisations in this thesis. The multipoles themselves will be taken from the calculated electron density by GDMA⁷⁷.

2.4 Progress in CSP

2.4.1 The blind tests

The easiest way to gauge progress in the CSP field is to read the results of the blind tests^{78;79;80;81;82;83}. Hosted by the Cambridge Crystallographic Data Centre (CCDC), six blind tests have so far been conducted. Groups actively developing CSP methods are invited to make predictions on molecules before the experimental crystal structures become available. Molecules for the first three tests were limited to: (i) a small, rigid molecule less than 25 (20 for the first test) atoms consisting only of C,H,N and O; (ii) a small, rigid molecule with less common elements (halogens for example); and (iii) a small molecule with some conformational freedom. The three molecules in the first test⁷⁸ were also stipulated to be in common space groups (though no list of space groups was provided) and contain one molecule in the asymmetric unit ($Z' = 1$). 11 groups took part and a maximum of three predictions for each molecule was made. Methods included ones already touched upon in this section such as UPACK, MOLPAK, RANCEL and DMAREL⁸⁴ (the progenitor of DMACRYS). From this first test seven correct predictions were made, five with the global minimum matching experiment, however the metastable polymorph of the first compound was not found and only one correct prediction was made for the flexible molecule. The molecules chosen in all the blind tests and the categories they fall in can be seen in Table 2.4.1

Table 2.1: Molecular structures of the molecules chosen in the blind test

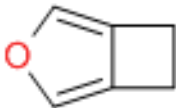
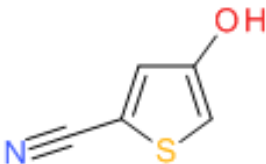
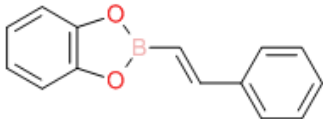
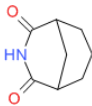

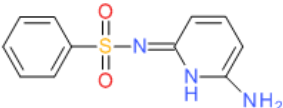
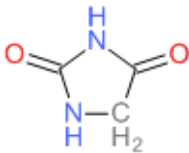
Molecular Structure	Category	Blind test
	i	1st
	ii	
	iii	
	i	2nd
	ii	
	iii	
	i	3rd
Continued on next page		

Table 2.1 – continued from previous page

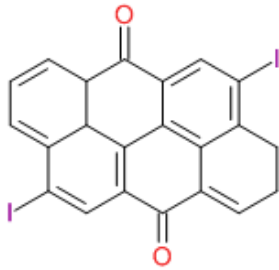
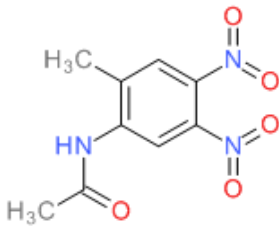
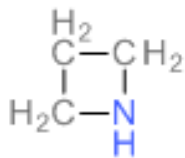
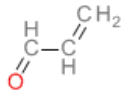
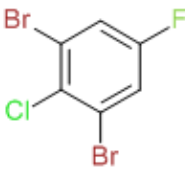
Molecular Structure	Category	Blind test
	ii	
	iii	
	i	4th
	ii	
	iiii	
Continued on next page		

Table 2.1 – continued from previous page

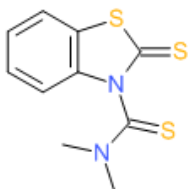
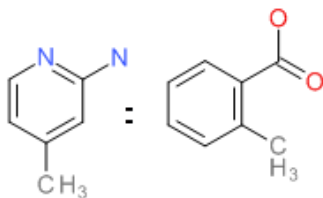
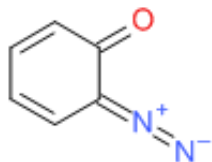
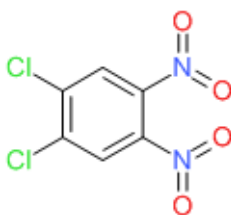

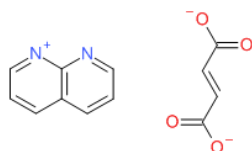
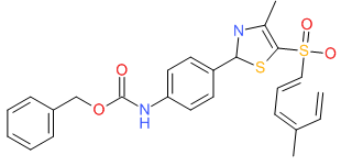
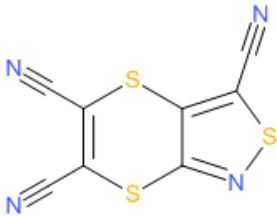
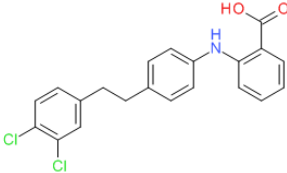
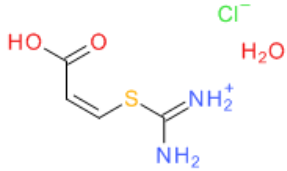
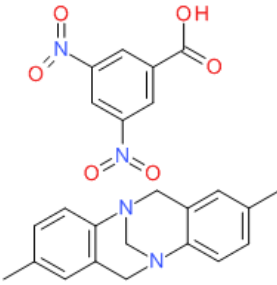
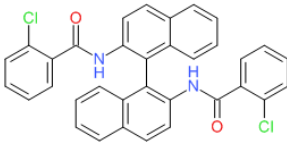
Molecular Structure	Category	Blind test
	iv	
	i	5th
	ii	
	iii	
	iv	5th
	v	
Continued on next page		

Table 2.1 – continued from previous page

Molecular Structure	Category	Blind test
	vi	6th
	i ₆	
	ii ₆	
	iii ₆	
	iv ₆	
	v ₆	

In the second test⁷⁹ 15 groups took part and six correct predictions were made. While the result is slightly worse than the first test, predictions for category (ii) improved.

However no successful predictions were made in category (iii). The third blind test⁸⁰ removed restrictions on the space groups being searched and allowed $Z' = 2$ structures. This introduced difficulties for the 18 groups participating but four had successful predictions for the molecule in category (i). Due to the unpublished crystal structure for (i) being released a second molecule was added in the category but no successful predictions were made. This was primarily due to the experimental structure crystallising with 2 molecules in the asymmetric unit. This reduced the number of participants who even had a chance of predicting the structure due to limitations in the search method or computing power. Only one successful prediction was recorded in (ii), as a result of the difficulties in modelling the anisotropic nature of iodine. Again no correct predictions were made for the flexible molecule. For the fourth blind test⁸¹, 14 groups took part and a 1:1 cocrystal was added to the list of molecules (category iv). For (i) a slight improvement was made over previous tests with three successful predictions in the participants' three submissions. Four successes were reported for (ii) with those groups reporting the correct structure as their global minimum. A marked improvement was seen in category (iii) with three groups predicting the experimental structure. The treatment of conformationally flexible molecules improved between tests 3 and 4, mainly through mapping the conformational energy surface using QM methods⁸¹. Two successful predictions were made of (iv) representing an improvement on the last time an attempt was made to predict a structure with two molecules in the asymmetric unit.

The 5th blind test⁸² introduced two new challenging categories: (v) a molecule with 4-8 internal degrees of freedom, 50-60 atoms and (vi) a polymorphic hydrate molecule that can fit into one of the other categories. Results in the first three categories were slightly down compared to the previous blind test with two successful predictions for (i), two for (ii) and one for (iii). For category (iv), a molecular salt was chosen which introduced some difficulties into the lattice energy ranking necessitating the comparison with similar molecules in the CSD. Nevertheless two groups successfully predicted the structure. Two groups also found success with category (v), both predicting the experimental structure as their global minimum. This marked the first time a structure of this complexity had been predicted under blind test conditions. No group matched the feat for category (vi), although four groups had exact matches within their extended list of predictions.

The most recent blind test⁸³ resulted in a change to the categories. Finding a polymorphic series of unpublished crystal structures only containing CHNO became difficult, so these were removed. The category containing co-crystals and salts was split resulting in five categories. Category i_6 now allows halogens in the accepted atoms, with the size of the molecule increased to 30 atoms. The second category (ii_6) became a partially flexible molecule with two to four internal degrees of freedom up to 40 atoms. iii_6 , a partially flexible molecule as in ii_6 but a salt. iv_6 multiple partially flexible molecules as a co-crystal or solvate and v_6 a molecule with four to eight flexible degrees of freedom and between 50-60 atoms.

As seen by the results of the previous blind tests the categories were challenging. However progress once again was made. The 6th blind test was the biggest so far, with 25 separate submissions and nine groups attempting every molecule. The first category saw 12 successful predictions (out of 21), one of which is included in Chapter 5. The second molecule caused some problems as it was discovered to have five polymorphs (A, B, C, D, E), two being $Z' = 2$ (C and E). 14 predictions were made at the three $Z' = 1$ polymorphs resulting in four predictions of A, eight of B and one of C. Four groups attempted the $Z' = 2$ polymorphs, resulting in one prediction of C and none of E. Nine attempts were made at prediction of the salt with one success but this was ranked 2nd in their submitted structures. Fourteen predictions for category iv₆ were made resulting in 5 groups locating the correct structure. The final flexible molecule was successfully predicted 3 times out of 14 attempts.

From the blind tests the limitations and strengths of CSP are readily apparent. For small, rigid molecules with well parametrized atoms CSP can be largely successful. However large, flexible molecules represent a major headache with very few successful predictions made. Cocrystals and structures with $Z' > 1$ are problematic as is modelling hydrates. Ions limit the transferability of potentials (by increasing the importance of induction and charge transfer terms). The results also allow the comparison of electrostatic models, with atomic multipoles doing slightly better than point charges, which often struggle to account for strongly directional interactions. The ranking of crystal structures was also of importance. Many groups predicted the correct structures but it was not ranked in their top three structures.

Using the same methodology on a range of molecules can also give insight into the performance of CSP. Day performed two CSP studies varying the potential⁸⁵ and electrostatic model⁸⁶. 50 organic, small, rigid molecules were chosen, occurring in the nine most common space groups with $Z' \leq 1$. Molecular geometry was optimised using DFT and simulated annealing generated trial structures. In the first study⁸⁵ two rigid molecule potentials were tested: W99⁶³ and FIT⁶⁴, which are both empirical potentials using the *exp-6* functional form for non-bonded interactions. Three flexible molecule potentials were also tested; Dreiding⁸⁷, an old but still widely used force field, CVFF⁸⁸ that uses the Lennard-Jones functional form (and is parametrised to *ab initio* data) and COMPASS⁸⁹ which uses molecular dynamics to include temperature in the parameter fitting. Out of a possible 62 known crystal structures 58 were correctly predicted. However, as seen from the blind tests, results for this type of molecule are usually encouraging. The W99 potential performed best overall. In the 2nd study⁸⁶, final lattice energy calculations were performed using two electrostatic models: an atomic point charge model (fitted to the MEP) and atomic multipoles from DMA. The use of multipoles improved the reliability of modelling hydrogen bonds significantly. 32 of 64 of observed crystal structures were found to be either global minima or within 0.5 kJ/mol of the lowest energy structure when atomic multipoles were used as the electrostatic model, compared

to 23 of 64 when using an atomic point charge electrostatic model.

2.4.2 Flexibility in CSP

Most CSP methodology is focused on rigid molecules. Flexibility introduces difficulties because the intramolecular (U_{intra}) energy needs to be taken into account when ranking structures. U_{intra} is the energy penalty associated with changing the gas phase conformation (relative to the *ab initio* minimum). However, this can be compensated by stabilising bonding motifs such as hydrogen bonds. The correct balance of intra and intermolecular forces can be difficult to achieve as the parameters for intra and intermolecular potentials are often derived in different studies.⁹⁰ The use of empirical intramolecular potentials can lead to large molecular distortions independent of the intermolecular model used⁹¹ and can often ignore the conformational effects on the intermolecular interactions.⁹²

Of course, inter and intramolecular contributions can be calculated with high accuracy *ab initio* calculations, or empirical intermolecular potentials combined with *ab initio* intramolecular terms^{93;45}. If a large conformational change is needed another *ab initio* calculation needs to be performed, allowing an accurate value for E_{intra} to be determined. If atomic multipoles are used they will also need recalculating as they are also highly conformationally dependent. While computationally expensive, this method has been used successfully in the predictions of glycol and glycerol^{93;94} as well as a series of monosaccharides⁹⁵ using UPACK³⁹. The program DMAflex⁹⁶ uses this hybrid approach, combining *ab initio* intramolecular calculations with a multipole description of the electrostatics. Only intramolecular degrees of freedom expected to change during crystal packing are optimised while others are fixed. While this cuts down on some of the expense it is still impractical for more than a few torsion angles and a few 10s of structures.⁹⁷ DMAflex has been used to successfully predict the structures of diastereomeric salt pairs⁹⁶, the steroid progesterone⁹⁸ and carboxylic acids⁹⁹ among others. Crystal Optimizer⁴⁵ develops on the ideas in both UPACK and DMAflex. The molecule under study is split into fragments and a database of E_{intra} is created for different torsion angles. If a specific torsion value falls between two already known points an interpolation/rotation is performed. If an interpolation cannot be performed an additional *ab initio* calculation is used.

Another approach begins with a search of the conformational space of the molecule. Using QM calculations E_{intra} is calculated as a function of the torsion angles of interest. Minima in conformational energy are then treated as rigid and used as the starting point of individual predictions. This method is particularly useful if the number of stable conformations is low and has been used in the prediction of a wide range of molecules.^{100;101;102} Obviously picking relevant torsions and the sampling ranges depends on chemical intuition but this can be assisted by using data from the Cambridge Structural Database.

There are tools such as Mogul¹⁰³ and ConQuest¹⁰⁴ that allow retrieval of geometric parameters such as bond lengths, angles and torsion angles. Brameld¹⁰⁵ showed that similar fragments often adopt the same conformation in different molecules. In the 5th blind test, CSD analysis played a part in both successful predictions of a molecule with 8 torsion angles.¹⁰⁶

2.4.3 Beyond classical methods

One way past the problems encountered in modelling flexible systems is to use quantum mechanical calculations for the entire crystal. Periodic DFT allows the electron density and positions of the atoms to be modelled together, eliminating the balance issues between intra and intermolecular contributions. The explicit modelling of the electron density also removes any reliance on atomic charges and allows electrons to move in response to the packing environment. These methods carry a high computational cost and are very sensitive to the exchange correlation functional.^{107;108} The inaccurate description of dispersive forces is the major weakness of DFT when applied to CSP as dispersion is a very important intermolecular term in molecular crystals. Dispersion corrected DFT (DFT-D) has been used in CSP where an empirical Van der Waals correction term was added to the DFT energy.¹⁰⁹ Many groups are involved in the development of their own dispersion corrections and are applying them to CSP and optimising crystals. This method combined with the use of tailor made force fields¹¹⁰, a force field (with parameters fitted from DFT-D) specific to each molecule was trialled in both the 4th and 5th blind tests where it performed extremely well. In the 6th blind test it predicted the structure of every molecule leading the researchers to say CSP has been "solved", however the rest of this thesis will show otherwise.

2.5 Conclusions

This chapter has summarised a variety of solutions to the prediction of possible crystal structures of a molecule. For the first step (molecular optimisation) DFT methods are the most common. Structure generation methods vary: random to pseudorandom generators are common, though other methods such as genetic algorithms and simulated annealing are also popular. Lattice energy minimisation is still most commonly performed with transferable force fields. Full DFT-D methods will become more common as the development of dispersion corrections continues and knowledge based methods still have a place in limiting the search space to relevant regions. Tailor made force fields have shown excellent success in recent blind tests, but their computational expense has limited their widespread adoption.

The results from the blind tests highlight two major problems; the increase in search cost with multiple molecules in the asymmetric unit and the treatment of molecular

flexibility. Co-crystals, hydrates and $Z' > 1$ structures increase the number of degrees of freedom rapidly, though as long as the components are rigid increases in computer power should open up these areas to fuller sampling. Molecular flexibility introduces the additional challenge of also searching conformational space, as multiple conformations may lead to low energy crystal structures. The effects of crystal packing on flexible torsions also needs to be taken into account. If the molecule is not rigid U_{intra} needs to be included in the energy ranking, a difficult proposition for current forcefields. This requires quantum mechanical methods to be used exclusively during the minimisation or some way of estimating the changes in conformational energy during a minimisation.

The next chapter focuses on the molecules CSP is applied to in this thesis, organic semiconductors.

Chapter 3

Organic Semiconductors

3.1 Introduction

The performance of any semiconducting device is measured by its mobility. That is, the speed at which charge carriers (electrons or holes) move through the structure of the semiconducting material in any direction. In order to have a net electrical current electrons must jump from completely filled levels to empty levels across the bandgap. If the bandgap is large, upon applying an external electric field at room temperature, there will be few electrons that have the necessary energy to jump from the valence band to the conduction band. Electrical conductivity can be described as;

$$\sigma = n \cdot \mu \cdot q \quad (3.1)$$

where σ is the conductivity, n the density of charge carriers, μ the mobility and q the elementary charge. To have conduction there must be charge carriers, which must be elevated to an excited state and therefore the conduction band (the LUMO level in organic semiconductors). To move the charge carriers an electric field must be applied. Charge mobility is the average speed of diffusion of the charge carriers (cm/s) as a function of applied electric field (V/cm) as below:

$$\mu = \frac{cm^2}{Vs} \quad (3.2)$$

For amorphous silicon μ is around 5 cm²/Vs, for organic semiconductors to compete with amorphous silicon they must have a μ of 1 cm²/Vs. The parameters governing the mobility of an organic semiconductor come from molecular electron properties, crystal structure and device fabrication. This chapter will introduce some of these parameters in more detail and how they can be tuned to improve the mobility.

3.2 Organic Semiconductors

Organic semiconductors are typically based on conjugated polycyclic aromatic hydrocarbons (PAHs). While the field can be split into polymers and molecular crystals both types require a conjugated aromatic backbone to allow charge to move. The π bonding present in the ring systems of these molecules allows the creating of a delocalised electron density above and below the plane of the molecule. The delocalisation allows charge carriers to move along the molecule. It is the interaction of neighbouring π -systems in a molecular crystal (or an aggregated solid) that leads to the movement of charge between molecules. The π -electrons are in the highest energy occupied orbitals and the lowest energy unoccupied orbitals are also π -orbitals, so that σ -orbitals and electrons can be ignored in models of charge transport. The larger the π -conjugation the better the delocalisation of these frontier orbitals and the easier movement of charge.

Crystalline organic semiconductors are usually one of the acenes, a series of linear fused benzene ring systems. Common members include benzene, naphthalene, anthracene, tetracene (Fig 3.1) and pentacene (Fig 3.2b). All have a general formula of $C_{4n+2}H_{2n+4}$, with $(4n+2)$ π -electrons. They are generally flat in the ground state, with the stability of the molecule decreasing as n increases; pentacene and hexacene for example can oxidise readily in light.

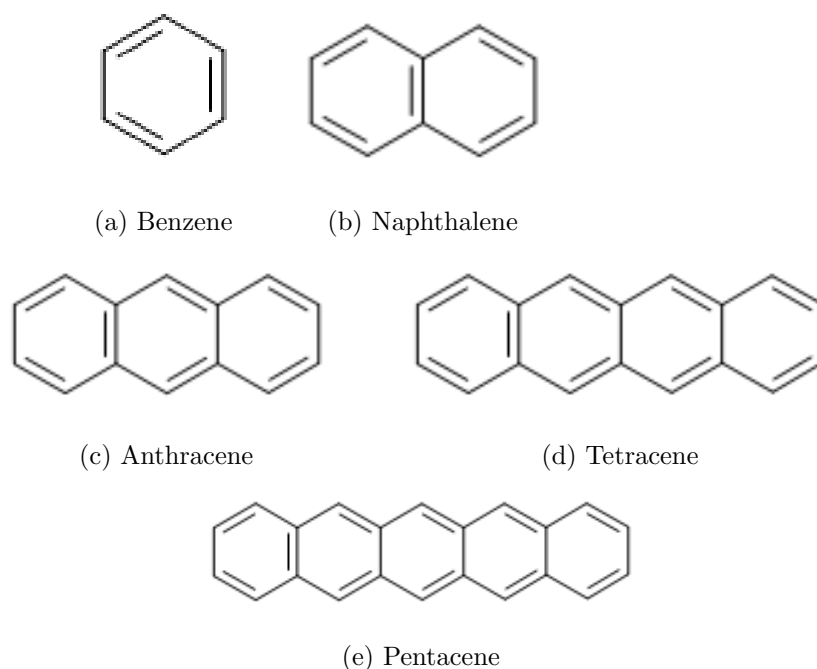


Figure 3.1: The first five acenes

While conductivity in graphite was known as early as 1947¹¹¹, the first reported results on molecular semiconductors were wide discotic molecules such as violanthrone in 1950⁷. The same authors then followed up by discovering that a large variety of PAH's formed

charge transfer complexes with iodine or bromine¹¹². This spurred the work of Kallman and Pope, who found that iodine could act as an electrode and inject holes into an organic crystal[?] . It was quickly realised that the rate of injection depended on the work function of the electrode and suitable metallic and semiconducting electrodes were found, kickstarting the development of organic light-emitting diodes[?] , organic field-effect transistors¹¹⁴ and organic solar cells[?] . Since these first discoveries of organic semiconductors, research in the field has exploded, with a search of the topic "organic semiconductors" returning over 180,000 hits through Web of Science (175,000 coming after 1991).

The benchmark for electronic devices, amorphous silicon, has a mobility of $1 \text{ cm}^2/\text{Vs}$ and is widely used in consumer electronics¹¹⁵. Intrinsic hole (a positive charge carrier) mobility in single crystals of rubrene (Fig 3.2c) has been measured at $20 \text{ cm}^2/\text{Vs}$ at room temperature¹¹⁶ with mobilities reaching $58 \text{ cm}^2/\text{Vs}$ in ultrapure pentacene¹¹⁷. In the realm of conducting polymers, the champions are diketopyrrolopyrrole (DPP, Fig 3.2d) containing polymers with hole mobilities of $8.2 \text{ cm}^2/\text{Vs}$ ¹¹⁸. Mobilities are lower for electron transporting materials, the crystals being represented by dicyanoperylene-3,4:9,10-bis(dicarboximide) (PDIF-CN2)¹¹⁹ and 5,7,12,14-tetrachloro-6,13-diazapentacene (TC-DAP)¹²⁰. Suitably functionalised DPP containing polymers again take the top spot¹¹⁸ but in both of these cases mobility does not exceed a few cm^2/Vs . These values pale in comparison to the typical values achieved by crystalline silicon ($500 \text{ cm}^2/\text{Vs}$) used in high speed electronics. If organic semiconductors are to become ubiquitous in our electronic devices and not just novelties they have a lot of ground to make up. Due to polymers inherent lack of crystallinity (and the subject of this project being the **crystal** structure prediction of organic semiconductors) this chapter will be focused on molecular crystals and specifically PAHs. The next section will discuss the different types of charge transport and the parameters affecting it.

3.2.1 Charge transport

For charge to move efficiently, the carriers (holes or electrons) must not be trapped or scattered. This depends on factors such as the electronic properties of the molecule, the crystal structure, how the device is processed and what conditions the device is operated at. Only the first two conditions are accessible to CSP (and relevant to this project). As mentioned above, most organic semiconductors are composed of π -conjugated systems and the availability of HOMO/LUMO levels allows the injection of charge carriers. An extended π -system also allows for the delocalisation of charge across a molecule. There are two distinct transport regimes: hopping and band.

In the band regime, the charge carrier wavefunction is delocalised over the entire system, meaning there is only a probability of finding the carrier at a given point. This results in coherent plane wave transport, in which the dominant factor is the electronic coupling

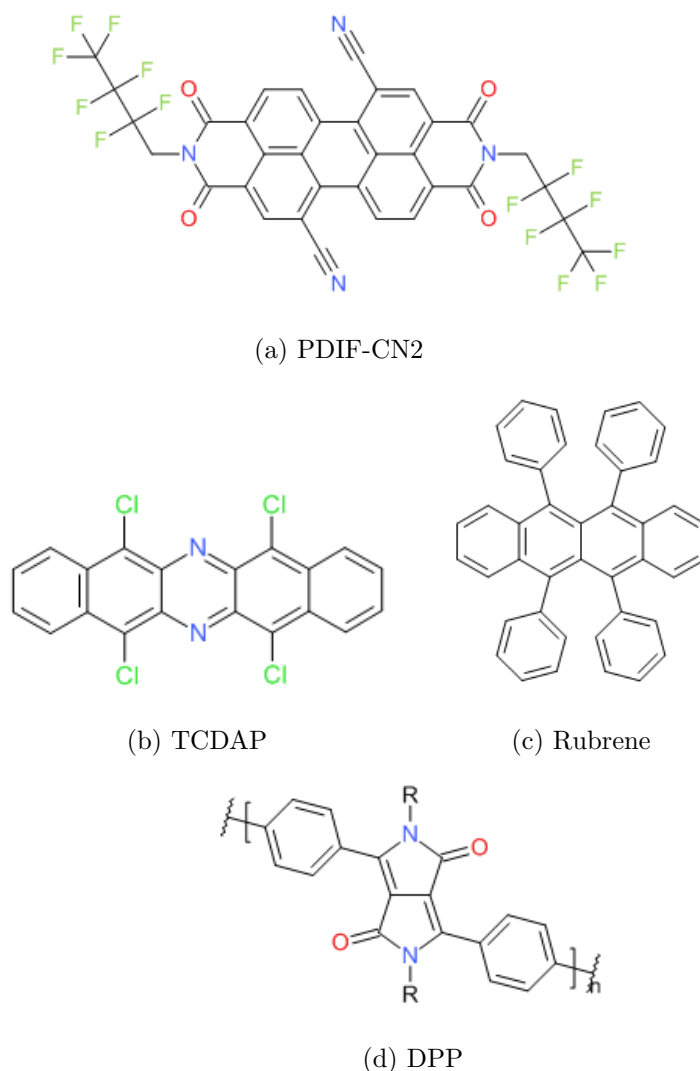


Figure 3.2: Some of the benchmark organic semiconductors

between units (molecules, polymer units etc). Band transport can occur in organic semiconductors^{117;121;122;123} at low temperatures ($<150\text{K}$) and with ultrapure samples. It is usually applied when mobilities reach $10\text{ cm}^2/\text{Vs}$ ¹²⁴. If the crystal was perfect and static, band transport would be achievable at ambient temperatures. In a real crystal however, there are always lattice vibrations (phonons) that disrupt the crystal symmetry. Lowering the temperature reduces the magnitude of the phonons and therefore increases the mobility.

Hopping transport is described in terms of a localised carrier (electron or hole) making its way between molecules. When a carrier lands on a molecule there is an energy cost as the molecule assumes the optimal geometry for the charged state. This energy cost is known as the reorganisation energy (Fig 3.3, λ , and is made up of two contributions: a vertical ionisation (neutral to charged), followed by a relaxation to the optimum charged geometry and the reverse process when the carrier leaves. The total λ is the sum of the

two relaxations and for many systems these values are similar enough to be calculated as twice the λ of each molecule. If the geometries of the neutral and charged molecules are similar, then λ is low and the carrier does little waiting around between hops. However, if the geometries are significantly different, the carrier must wait for a phonon to bring the next molecule to a geometry close to the charged state or the introduction of energy into the system to induce the geometry change. When the charge carrier leaves, there is an intramolecular relaxation from the original molecule back to the neutral state and an intermolecular relaxation as the formerly charged molecule can now move closer to other neutral molecules. This is the main difference between the hopping and band regimes: in the hopping regime the coupling of carriers to vibrations is the main factor, while in the band regime electronic coupling between units (molecules, polymer units etc) is the dominant factor. The electronic coupling is still important in the hopping regime, but is not the dominant contribution. Table 3.1 presents an overview of the two regimes.

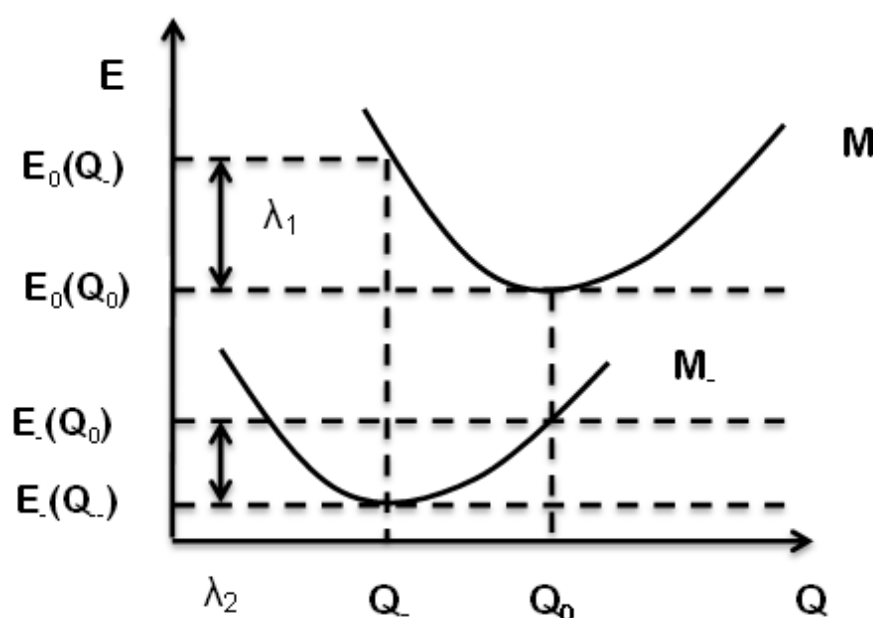


Figure 3.3: Definition of the internal reorganisation energy for an electron-transfer reaction.

Table 3.1: The main differences between band and hopping transport

Band regime	Hopping regime
Delocalization of the charge carrier wave functions	Localisation of the charge carrier wave functions
Electronic coupling	Incoherent hops
	Electron-phonon coupling
	Geometry relaxations
	Both inter and intramolecular modes
Low temperature ($< 150\text{K}$)	Activated process

In energetic terms, if the energy saving from electronic coupling and charge carrier delocalisation is greater than that of electron-phonon coupling, then band transport is

favoured, as several isolated orbitals form a low energy valence band and a high energy conduction band. If the opposite is true, then hopping is favoured and localisation occurs. The calculation of which regime is more likely in a crystal is not straightforward. Due to weak intermolecular forces many properties of the crystal are anisotropic and both modes of transport can exist within the same crystal and there is no comprehensive theory encompassing band and hopping modes of charge carrier transport. In the next section we will discuss a theory that has had success in describing hopping transport within molecular crystals: Marcus theory.

3.2.2 Marcus theory, reorganisation energy and transfer integrals

Marcus theory was initially developed to model the electron transfer between ions in solution which do not undergo large geometrical changes[?]. Since the hopping regime is a single carrier transfer reaction between two molecules of the same composition it can be modelled using classical Marcus theory, where the hopping rate constant is given by equation (3.3),

$$k_{et} = (t^2/\hbar)(\pi/\lambda_{\pm}k_B T)^{\frac{1}{2}} \exp(-\lambda_{\pm}/4k_B T) \quad (3.3)$$

where k_{et} is the hopping rate, t is the transfer integral, λ_{\pm} is the reorganisation energy (discussed above), k_B is the Boltzmann constant and T is the temperature.

The transfer integral is equivalent to electronic coupling discussed in the previous section, given by,

$$t = \langle \phi_m | H_{el} | \phi_n \rangle \quad (3.4)$$

where t is the transfer integral, ϕ_m is a molecular orbital on molecule m , H_{el} is the electronic Hamiltonian and ϕ_n the molecular orbital on molecule n . For hole transport these molecular orbitals are the HOMO, for electron transport the LUMO. If the transfer integral was the dominant term band transport would be achieved, but even in the hopping regime it should still be maximised. As expected of a parameter that depends on the overlap of the molecular wavefunctions the transfer integral is largely dependent on the relative orientation and position of the molecules. As the molecules come together their wavefunction overlap increases exponentially; this results in a splitting of HOMO/LUMO levels into a valence (holes) and conduction (electron) bands. With a perfect cofacial arrangement of the molecules a bandgap of 1 eV is seen, which is consistent with the band regime. Of course the molecules cannot align directly, as the areas of maximum electron density will align and repel. Instead, a displacement along the short or long molecular axis will be seen to adopt a more favourable packing motif. Depending on the specific orbital features of the molecules, large oscillations in bandgap will be seen with a lateral displacement. If the new overlap is unfavourable (due to antibonding/bonding patterns) this could result in the reduction of either of the two

bands. The total bandgap will always be decreasing as fewer molecular orbitals overlap. It must be remembered that it is the wavefunction overlap that matters not just the spatial overlap. The crystal structure therefore has a large impact on the transfer integral. How the molecules pack dictates the most likely way for the charge to move. Later in this chapter a more detailed look at the packing of PAHs (specifically pentacene) will be undertaken.

As discussed above, the reorganisation energy depends on both intra and intermolecular relaxations to move the geometry of the molecule close enough to accept or pass on a charge carrier. Some vibrational modes will have no effect, while ones that lead to a molecular geometry close to that of the ionised state will be strong. This explains the increase in hopping mobility with temperature. The increased vibrations help prepare the molecules to achieve the correct geometry for the hopping to take place. The decomposition of the electron-phonon coupling is beyond Marcus theory, but they can be split into local (affecting the reorganisation energy) and non-local contributions (affecting the transfer integral) in more detailed investigations.

The good electronic properties of pentacene have been attributed to its small λ_+ (λ_+ = hole transport, λ_- = electron). This is due to the delocalisation of its orbitals over the whole molecule, resulting in very little distortion of molecular structure upon the capture of a hole. The reorganisation energy has two (inter and intra) components described as relaxations above, but it has been shown that gas phase reorganisation energies for single molecules can approximate the value in a crystalline environment.¹²⁵ The transfer integral is related to the splitting of the levels transporting the charge and can be approximated as half the HOMO-LUMO splitting.¹²⁶ However work by Valeev *et al*? showed that polarization effects can have a large effect on the energy splitting for non-symmetric dimers. Later on in this chapter there is a description of how charge mobility calculations have been performed on the systems studied.

Marcus theory is not without drawbacks¹²⁷. The idea of a discrete carrier hopping between sites is intuitively attractive and retains popularity in organic semiconductor research. However, Marcus theory is only valid in the non-adiabatic limit (when the transfer integral is very small) and in the high temperature limit. If the transfer integral is too large the charge transfer reaction will proceed along the lowest energy electronic state adiabatically and Marcus theory is no longer valid. Another issue is the existence of a discrete charge carrier. If the transfer integral is greater than half the reorganisation energy the charge carrier does not exist, as the coupling is so strong any potential energy minima collapse into one minimum and the wavefunction is delocalised. For pentacene and rubrene for example, the largest electron couplings are 118 and 142 meV respectively, while half the reorganisation energy is 159 and 90 meV. From these values it can be that pentacene especially can not form a discrete charge carrier and no hopping rate can be defined. Marcus theory is valid in the high temperature limit where motions can be treated classically, the movements of water molecules in solution are much slower than

the rate determining step in intermolecular charge transfer, which are usually bond stretches. By treating these motions classically (rather than quantum mechanically) important effects such as zero point energy are excluded, which can often be larger than the energy barrier between two minima and leave the hopping rate incalculable.

However, it must be stressed that the focus of this work is on seeing how we can explore chemical space and crystal packing space; we use Marcus theory as a commonly used approach, but with the knowledge that it will not be a perfect description of charge transport. What we hope is that it is 'good enough' to reveal trends. Using equation (3.3) it can be seen that for good charge transport the transfer integral needs to be maximised and the reorganisation energy minimised. Much work has gone into molecular and crystal engineering of semiconductors to achieve these goals and will be touched upon in the next section.

3.2.3 PAHs and Pentacene

Like most PAHs, pentacene crystallises in a herringbone arrangement¹²⁸. This structure represents a compromise between edge to face ($C\cdots H$ interactions) and face to face π -stacking ($C\cdots C$ interactions). Desiraju and Gavezzotti¹²⁸ identify 3 other common packing motifs for PAHs, *a*) sandwich herringbone, with two molecules making up the herringbone motif; *b*) γ , a flattened herringbone where the main $C\cdots C$ interaction is between parallel translated molecules; and *c*) β where strong $C\cdots C$ interactions leads to the formation of graphite-like cofacial arrangements. Examples of these motifs can be seen in Fig 3.4

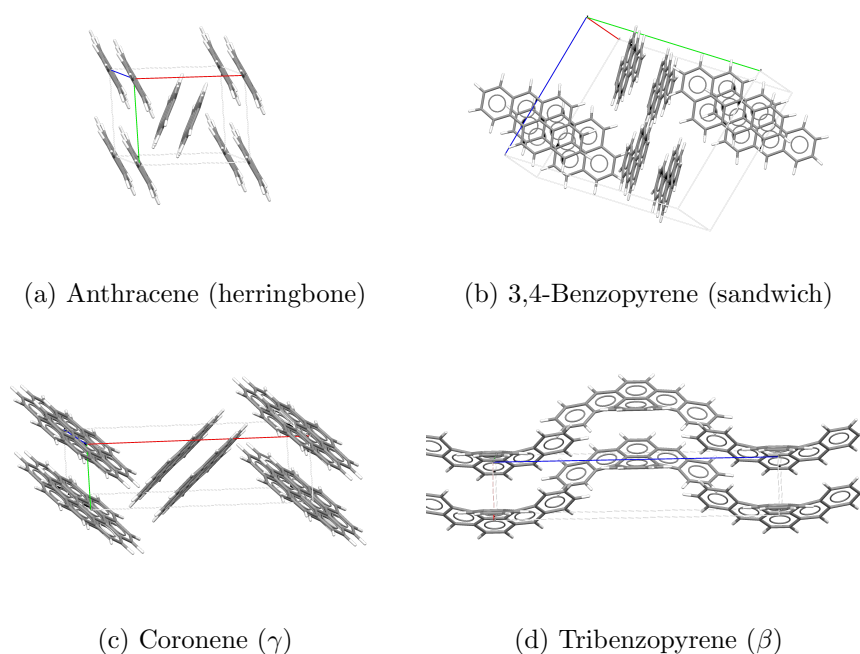
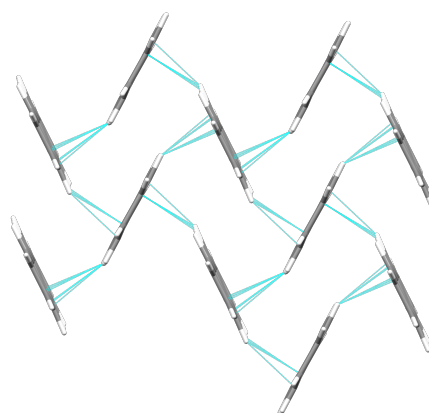
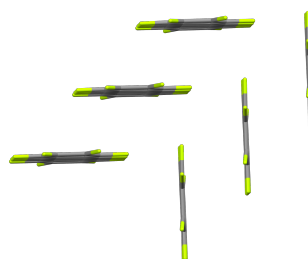


Figure 3.4: The four most common packing motifs in PAHs

Herringbone packing typically results in 2D charge transport between the stacked layers (along the edge-to-face interactions highlighted in Fig 3.5a). This is not the most favourable packing for charge transport, as calculations by Bredas and coworkers¹²⁶ showed that HOMO/LUMO splitting was highest in cofacial arrangements. Thus cofacial arrangements are the most effective for charge transport and edge-to-face interactions must be disrupted in some way to accomplish this. In addition, pentacene is a p-type semiconductor (a hole transporter): p-type materials have been easier to obtain and thus extensively researched¹²⁹. In a 2012 review¹³⁰ of π -conjugated systems, p-type semiconductors outnumber n-type 2 to 1. However n-type materials (electron transporters) are equally important for organic electronics. They enable complementary circuit design which reduces power requirements and the production of p-n junctions.^{131;132} N-type devices are usually made by introducing electronegative atoms into the ring system of a PAH or as substituents on the outside of the molecule.



(a) Pentacene



(b) Perfluoropentacene

Figure 3.5: (a) Herringbone packing of pentacene with edge to face interactions highlighted, (b) altered herringbone packing seen upon perfluorination of pentacene.

Functionalisation offers a way to change the electronic properties and the packing of

pentacene. Perfluorination of pentacene leads to a n-type material (CSD ref BEZLUO), this also changes the packing but the basic herringbone structure remains with an edge-to-face angle of nearly 90 degrees (Fig 3.5b).¹³³ Electron mobility is a respectable 0.22 cm²/Vs, even though theoretical calculations showed that λ is doubled with respect to pentacene.¹³⁴ The same authors also investigated a perfluorinated tetracene¹³⁵ though performance was lower than the pentacene derivative.

The addition of bulky substituents at the central positions on the pentacene ring system is one way to disrupt edge to face interactions. Typically, large functional groups are added to alkyne "spacers".^{136;137} The alkyne spacer is sterically undemanding to preserve good π overlap, while controlling the size and shape of the functional group allows fine control over the crystal packing.¹³⁸ This can be seen below (Fig 3.6) as the packing changes from 1D π -stack with trimethyl to 2D π -stacked "brickwork"/lamellar packing with triisopropyl.

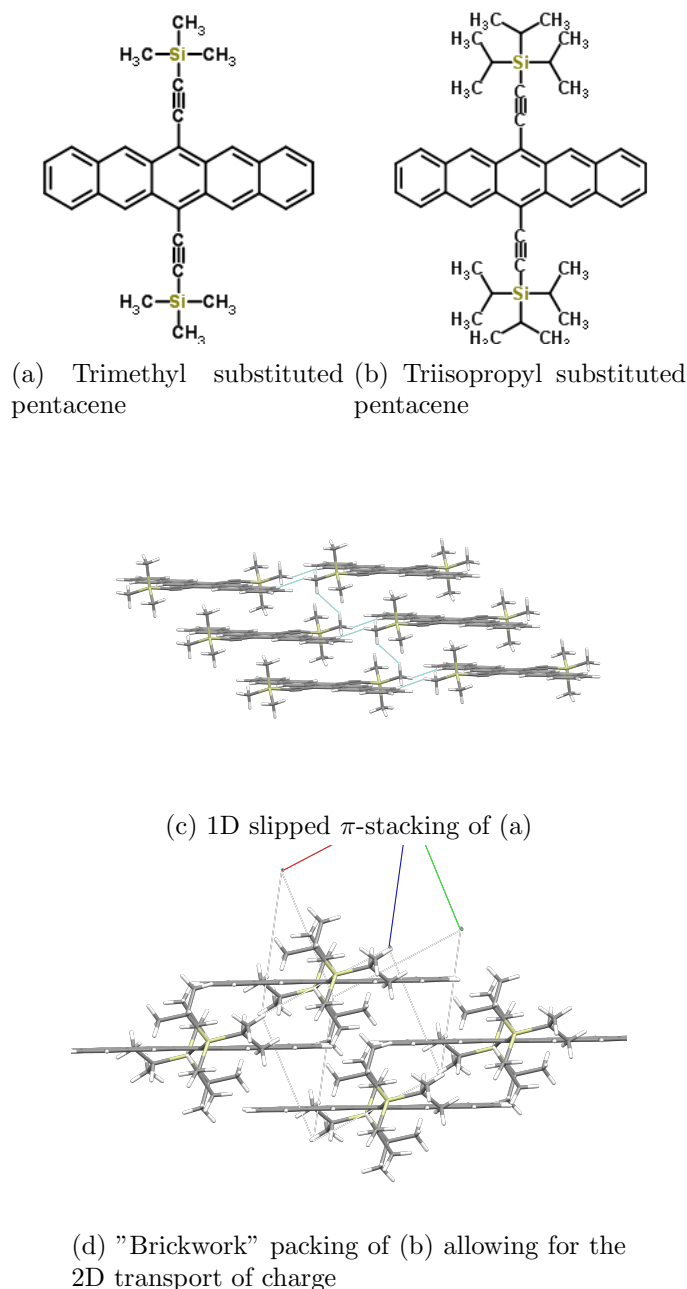


Figure 3.6: Changes in packing as the silyl substituent becomes larger

Many other functionalisation schemes exist¹³⁷ and one of interest is hetero-atom substitution into the pentacene (or other acene) molecule. While the addition of electronegative substituents (as in perfluorination) increases n -character, azaacenes (and azapentacenes in particular) offer a way to favourably modify electronic properties and crystal packing. Interest in azaacenes has increased over the last few years^{139;140} due to this potential control and intriguing theoretical results. Chen and Chao¹⁴¹ investigated a series of azaacenes as they tried to lower the value of the internal λ . The authors showed that too much nitrogen substitution (10N) increased λ_- due to the effects of the electronegative perturbation on the LUMO, increasing its non-bonding character. This

leads to stronger orbital interactions between neighbouring atoms, resulting in a larger geometry change when an electron hops on. However, 10N substitution did result in a large increase in electron affinity (a property needed for a good n-type material¹³¹) so 5N was also investigated and showed good (0.149-0.167 eV) and tunable (depending on N position) λ values.

Winkler and Houk¹⁴² also looked at a series of azapentacenes, calculating reorganisation energies for varied amounts of nitrogen substitution and locations including 5N, where their λ values agree quite well with those of Chen and Chao. N-substitution will not just change the electronic properties but also the molecular packing. The replacement of C-H moieties should reduce edge-to-face interactions and lead to face-to-face dominating. The ability to form $N \cdots H-C$ hydrogen bonded networks will also increase the stability of the material. Some work has been done showing promising results with N substitution into already 6,13-substituted pentacene derivatives, although no thorough examination of the crystal packing was performed.¹⁴³ Winkler and Houk state that "A most interesting question is how substitution of CH by N modifies the solid-state structures (and hence transfer integrals) of azaoligoacenes.". This will be explored thoroughly in Chapter 6.

3.2.4 Charge transport calculations in this thesis

For our charge mobility calculations, the two parameters needed are the reorganisation energy of the molecule, and the transfer integral in the crystal structure. As mentioned above, the reorganisation energy can be calculated with reasonable accuracy from the gas phase structure of the molecule. The λ_- is calculated as the sum of the energy required for the reorganisation of the vertically ionised neutral geometry to the anion geometry, plus the energy required to reorganise the anion geometry back to the neutral equilibrium. The process of calculating the reorganisation energy requires four calculations; *a*) a geometry optimisation of the neutral molecule; *b*) a geometry optimisation of the molecule with a negative charge; *c*) a single point energy of the *a* geometry with a negative charge and *d*) a single point energy of the geometry of *b* with a neutral charge. The energies from these calculations are then summed as,

$$(c - b) + (d - a) \quad (3.5)$$

or more formally,

$$\lambda_- = E_{0opt}(Q_-) - E_-(Q_-) + E_{-opt}(Q_0) - E_0(Q_0) \quad (3.6)$$

where the subscript on E refers to the geometry of the molecule and Q the charge. For the hole reorganisation a similar procedure would be followed, replacing the negative charges with positive ones.

Calculating the transfer integral from predicted crystal structures is a rather more complicated task. The calculation must be performed not on the entire crystal structure but between dimer pairs present in the crystal. The more unique dimer pairs, the more calculations need to be performed. For hopping transfer only two molecules in the crystal need to be treated at one time, allowing the splitting of the crystal structure into fragments. In the tight binding approximation it is assumed that only adjacent molecules can couple. As mentioned above the transfer integral can be approximated as the splitting of the two HOMO levels (for hole transport) or LUMO levels (for electron transport) in a dimer. However this method is only an estimate and a full treatment requires the spatial overlap of the molecular orbitals to also be taken into account.

Frozen density embedding (FDE)¹⁴⁴ as implemented in the Amsterdam Density Functional⁷ allows the use of molecular orbitals on individual molecules as a basis set in calculations on a system composed of two or more molecules. FDE, first developed for use in the study of solvated molecules¹⁴⁵, describes the total electron density as the sum of subsystem densities as below,

$$p(\rho) = p_i(\rho) + p_{ii}(\rho) + \cdots, \quad (3.7)$$

where i and ii are the labels of the subsystems. The partitioning of the system allows FDE to return subsystem-localized electronic structures which resemble the acceptor and donor states in a hopping charge transfer. Running the calculation first requires a single point energy on each fragment of the system which gives the initial density and energy of the fragments with no interaction between them. A FDE calculation is then performed taking into account the dimer structure, giving the embedded density of each subsystem. This gives both the charged and uncharged structures and densities of each half of the system. From there the final electron transfer rate constant is calculated through the computation of the overlap matrix elements which are needed for the coupling matrix elements, which are equivalent to the transfer integral.

The reorganisation energy and transfer integral allows the rate constant of electron transfer to be calculated from (3.3), this value can then be used to calculate the electron diffusivity (\mathcal{D}),

$$\mathcal{D} = \frac{1}{3 \sum_{i,j} N_{ij}} \sum_{i=1}^N \sum_{j=1}^{N_i} r_{ij}^2 k_{ij}^2 \mathcal{P}_{ij}, \quad (3.8)$$

where N is the number of symmetrically independent molecules in the crystal, N_i is the number of nearest-neighbouring dimers for the i -th molecule and r_{ij} is inter-centroid distance. \mathcal{P}_{ij} is the probability for the charge carrier to hop between molecule i and j , and is calculated as

$$\mathcal{P}_{ij} = \frac{t_{ij}^2}{\sum_{j=1}^{N_i} t_{ij}^2}. \quad (3.9)$$

The intermolecular electron transfer rate is calculated according to equation (8.3). The value of the diffusivity can then be used in the Einstein equation to calculate the electrical mobility of the electron as below,

$$\mu = \frac{e}{k_B T} \mathcal{D}. \quad (3.10)$$

All calculations will be performed at $T = 300$ K.

3.2.5 Conclusions

This chapter has presented the types of charge transport, factors needed for good charge transport in organic semiconductors, how to modify them and the most common crystalline molecules used to create organic electronics. Two possible regimes are possible in organic semiconducting crystals, band and hopping. Band transport involves the delocalisation of the charge carrier wavefunction resulting in coherent transport. The electronic coupling (or transfer integral) is the most important factor in the band regime and due to this requirement it is usually seen at low temperatures. As temperature increases, a drop in mobility is seen before hopping transport can begin.

Hopping transport has localised charge carrier wavefunctions, which move from molecule to molecule. It is strongly coupled to the phonons within the crystal and molecule, as the geometry of the molecule must be ready to accept or eject the charge carrier. The smaller the reorganisation of the molecular geometry for this process the lower the molecule's reorganisation energy and the hops are made easier. Outside of considerations in device processing and synthesis, the transfer integral and the reorganisation energy control the charge mobility. The transfer integral depends on the overlap between molecular wavefunctions in the crystal and can be modified by changing the crystal packing. The reorganisation energy depends on how delocalised the molecular orbitals are, and this is often modified by substitutions or additions to the molecule (not always for the better). Using Marcus Theory for the calculation of charge mobility shows that the transfer integral should be maximised while the reorganisation energy should be minimised.

Modifying the crystal packing of PAHs allows the tuning of the transfer integral. Many different schemes have been developed to lead to cofacial packing from the more common herringbone motif. One that shows promise is the creation of heteroacenes (specifically azaacenes) which allow hydrogen bonding networks. Electronegative substituents also change the PAHs from hole transporters to electron transporters. While some azaacenes molecular properties have been investigated, little work has been performed on examining the packing motifs of these molecules. Chapter 5 in this thesis presents a CSP study on a six azapentacene molecules, and calculates the charge mobility from the predicted structures. The next chapter will discuss machine learning, in particular genetic algorithms, and how they can be applied to the search for novel organic semiconductors.

Chapter 4

Machine Learning and Genetic Algorithms

4.1 Machine Learning and Genetic Algorithms

The primary goal of machine learning is to construct computer programs that automatically improve with experience. Since the dawn of the field, machine learning algorithms have succeeded in tasks such as the optimisation of search engines, speech recognition, autonomous vehicles and economics. Machine learning draws from varied fields such as biology, information theory, statistics, philosophy and more thus many different methods have been developed to meet the goal of "learning" programs. While we do not yet know how to make a program learn as well as a human, many algorithms have been developed that are effective for certain types of learning tasks. Mitchell¹⁴⁶ provides a formal definition for machine learning, "A computer program is said to **learn** from experience E with respect to some class of tasks T and performance measure P , if its performance at tasks in T , as measured by P , improves with experience E ."

This framework can be applied to many possible learning tasks. Checkers, being a relatively simple game but with a depth of strategy has been a testbed for many advances in machine learning. The world's first self learning program was designed to play checkers and presented in 1951¹⁴⁷, it was written by Christopher Strachey for a discussion on "logical or non-mathematical programmes". Computing pioneer Arthur Samuel followed by writing another checkers program at IBM in the 50's, and it was demonstrated on television in 1956¹⁴⁸. Using Mitchells definition we can formulate a checkers learning problem. The program's performance would improve as measured by its ability to win, its task would be to play checkers games and the experience would be gained through playing games against itself. In general, to have a defined learning problem we must identify the tasks, measure of performance and experience the program will learn from.

A checkers learning problem:

- Task T : playing checkers
- Performance measure P : percent of games won against opponents
- Training experience E : playing practice games against itself

Many learning problems can be defined this way.

A robot driving problem:

- Task T : driving on public roads
- Performance measure P : distance covered before error
- Training experience E : images and inputs from watching a human driver

A speech recognition problem:

- Task T : recognising and classifying spoken words
- Performance measure P : percentage of words classified correctly
- Training experience E : a database of spoken words with given classifications

Or more relevant to this thesis:

A molecular design and property optimisation problem

- Task T : design and CSP of a molecule
- Performance measure P : favourable molecular property or charge mobility of predicted crystal structures
- Training experience E : molecules with known charge mobilities

Machine learning tasks are typically separated into three broad categories: supervised learning, unsupervised learning and reinforcement learning depending on what feedback is available to the program (Table 4.1). These classes are not mutually exclusive however. It is possible to write a learning program using multiple methods of learning. The first choice is usually the learning method the algorithm will follow. However there are many decisions to be made after that. In supervised learning the program is presented with inputs and the desired outputs with the goal being to learn a function to map inputs

to outputs. This function can then be used to map new inputs to a correct output. To set up a supervised learning problem some initial steps need to be followed. The first decision is determining the type of training examples. In the case of the speech recognition problem above this could be single words, syllables or common phrases. The next step is gathering training data. This set should be representative of the real-world use of the function. Inputs and outputs must be gathered to allow training of the learner. If our speech recognition program is to be used on television speech to text successful examples of this task are needed. Next comes determining the representation of the inputs. This is a vital step as the performance of the learner depends strongly on the representation. Typically a vector is used, with enough features to describe the object. The number of features should not be too large to avoid dimensionality problems¹⁴⁹, but still contain enough information to accurately predict the output. Now the structure of the algorithm and the learning function needs to be considered. Different algorithms perform better with different problems, but many can do well if the input representation is strong enough¹⁵⁰. Once the algorithm is functionally complete it needs to be run on the training set. Depending on the choice of algorithm there may be some parameters of the function left to be decided. These can be chosen by first running on a smaller sub set of the training data. After evaluating the performance of the algorithm on the training set, it can then be given a separate test set of unseen data.

Table 4.1: The three common types of learning used to train machine learning algorithms.

Type of learning	Learning feedback available
Supervised	Inputs with their desired outputs
Unsupervised	No output given, learner must infer patterns present in the input
Reinforcement	Learner interacts with an environment aiming to maximise a reward

While following the simple steps above there are a number of issues to consider. First is the balance between bias and variance. Both bias and variance introduce error into the learning algorithm, and problems can arise when both of these are minimised simultaneously. Bias is the error from false assumptions made during the design of the algorithm. High bias can lead to underfitting as the algorithm misses relations between features of the input to the output. Variance is the error created by the sensitivity of the algorithm to small fluctuations in the training set. High variance leads to overfitting, modelling the noise present in the data, rather than the intended outputs. Ideally a method would be chosen that both accurately captures the regularities in the training data and generalizes well to new data. However in practice this is typically impossible to do both simultaneously. Simplifying the model through reducing the dimensionality of the problem can decrease variance, as can using a larger training set. But adding features decreases bias at the cost of increasing variance. Most learning algorithms have tunable parameters to control the tradeoff, such as the use of hidden layers in neural

networks to increase the variance (which decreases bias) or controlling tree depth in decision tree models to reduce variance.

The second issue is the relationship between the amount of training data present and the complexity of the "true" function being investigated. If the true function is simple then an inflexible algorithm (high bias, low variance) will be able to learn it from a small amount of data. However if the true function is complex (it may depend on a number of combinations of input features or behave differently across the feature space) then it will only be able to learn from a large amount of training data and a flexible (low bias, high variance) algorithm will perform better. A good algorithm will therefore automatically adjust the bias/variance trade off depending on the amount of data available. This occurs in the pruning of neural networks¹⁵¹, where the hidden layer is initialised with a large number of nodes, after training those nodes which have small weights are removed as these are either contributing nothing or adding noise to the model.

The third issue is the dimensionality of the input space. If the feature vectors have a high dimensionality, then learning the true function can be difficult even if it only depends on a small number of those dimensions. The many superfluous dimensions can cause high variance in the algorithm. Thus a high input dimensionality requires tuning of an algorithm to lower variance and increase bias. Removing the useless dimensions can dramatically improve performance, whether this is done by hand or using another method to reduce dimensionality before training the algorithm.

The fourth issue to consider is the presence of noise in the output values. If the desired output values are often incorrect (due to human or sensor error) then the algorithm should not try to learn a function that matches the training examples exactly, as this can lead to overfitting. As discussed above, overfitting can also occur in the absence of noise if the true function is too complex for the learning model. In this case the part of the function that cannot be learned damages the performance of the algorithm as whole. In either case it is better to go with a lower variance higher bias model learner for noisy data. However there are strategies to deal with noise in the output, such as stopping the training early to prevent overfitting or the removal of particularly noisy examples from the training set.

Other considerations include the homogeneity of the data set; if the data is homogeneous, numerical and scaled (leading to simple feature vectors) some methods will perform better than others. Algorithms that use a distance vector are particularly sensitive to the form of the data. If the data is heterogeneous and the feature vector includes features of many different types, then a decision tree method may be more useful. If the data contains many redundant features this can also cause problems for distance based methods, as the redundancy can lead to numerical instabilities. The interactions between features present in the vector must also be taken into account. If the features do not interact and each makes an independent contribution to the output then algorithms based on linear

combinations can perform well. However, if there are complex interactions between the features then methods such as neural networks or decision trees designed to highlight hidden interactions may be a better choice.

Some of the factors discussed above also apply to the other forms of learning experience chosen, specifically those to do with input representation. In unsupervised learning the program must learn the patterns present in the input itself from unlabelled data; there is no output for the algorithm to learn from. The most common unsupervised learning task is the clustering of inputs into specific subsets with the aim of correctly clustering new inputs. Reinforcement learning allows the program to learn from interacting with its environment. Examples of the above include the robot driving program and the checkers program learning through playing itself. Unlike supervised learning, no correct input/output pairs are ever presented to the algorithm, and mistakes made by the algorithm are not corrected. Each action (or set of actions) has an associated reward, and it is up to the algorithm to maximise the reward gained. In addition, there are sub-classes such as semi-supervised learning (in which output values are provided for only a subset of the training data) and active learning (in which the algorithm gathers new training examples as it progresses). As mentioned above, the choice of learning method is largely problem-dependent and many families of algorithms can be used with any of the three methods.

Table 4.2: Desired outputs of machine learning and the common algorithms used for these outputs.

Desired output	Common machine learning algorithms
Classification	Support vector machines, k-nearest neighbours, decision trees
Regression	Linear regression, ordinary least squares, Bayesian
Clustering	Hierarchical, k-means, density based
Density Estimation	Kernel density estimation, spectral density estimation
Dimensionality reduction	Principal component analysis, linear discriminant analysis

Another categorisation scheme used for machine learning programs is the expected output of the system (Table 4.2). Classification is the problem of identifying to which set of sub-groups a new input belongs. Inputs are divided into two or more sub-groups, then the learning program must produce a model which assigns new inputs to one of these groups. Classification is typically considered supervised learning, as there is a set of correctly sorted inputs available to the learner. The unsupervised version of this task is commonly referred to as clustering, where the sub-groups of the inputs are not known initially and the learner must create its own. Clustering and classification can both be thought of as pattern recognition tasks and regression is also included in this group. Like classification, regression is usually a supervised task, but the outputs are continuous variables rather than discrete groups. Regression typically overlaps with machine learning when used to make predictions. Density estimation uses statistical models

to find an underlying probability distribution that gives rise to the observed variables, which is typically an unsupervised task. Dimensionality reduction attempts to reduce the number of variables in a problem to a set that captures the essential behaviour of the system as a whole. This set can be a subset of the observed variables or an entirely new set that better describes the system. Examples include principal component analysis and the use of autoencoders in neural networks.

The next section will discuss four of the common algorithms used in machine learning for the above desired outputs.

4.1.0.1 Support Vector Machines

Support vector machines (SVM) are a common tool for classification and regression tasks. At its most basic a SVM is a representation of input data in space mapped so that different categories of data are separated by as wide a gap as possible. SVMs are linear classifiers, as they are based on a linear combination of the features of the vector. However, not all input data is separable linearly in 2D space, but can be mapped to a linear distribution in feature space. Kernel functions are used to map the data into feature space; the coordinates of the data in this space do not need to be computed, just the inner product between each pair. Many different kernels can be used as long as Mercers condition (that the kernel is positive definite) is satisfied. Each data point is viewed as a \mathbf{n} -dimensional vector and we need to separate these points with a $(\mathbf{n}-1)$ -dimensional hyperplane which allows the surface between our data points to be represented geometrically. There are many hyperplanes that might classify our data and the use of a SVM ensures the best one will be chosen. The best in this case is defined as having the largest distance between data points on each side of the vector (Figure 4.1). Those data points closest to the margin are the support vectors and if they were removed would change the location of the optimal hyperplane.

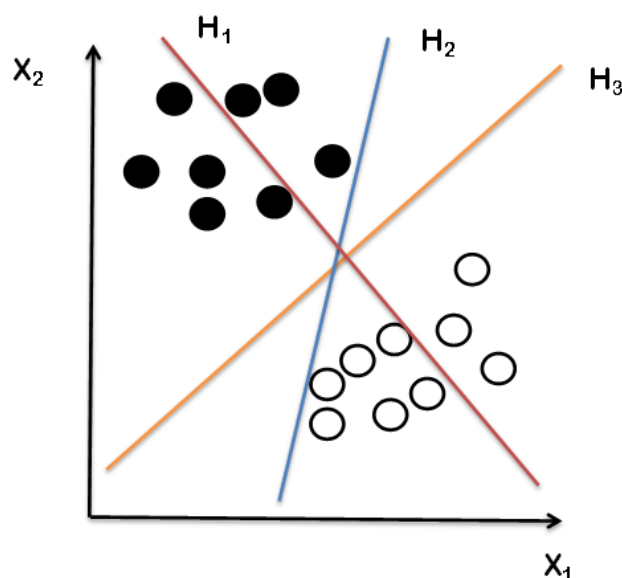


Figure 4.1: Schematic representation of a support vector machine in two dimensions. H_1 does not separate the classes correctly, H_2 does but with a very small margin and H_3 separates them with the maximum margin.

Using the hyperplane with the largest margin is the best choice as it allows for the easy classification of new data and the larger the margin the better the generalisation of the function creating the hyperplane. Optimising the size of the margin is a quadratic programming problem and is usually accomplished using Lagrangian multipliers. SVMs are strong tools for classification as the kernel allows the use of non-linear data. No assumptions are made about the functional form needed for the transformation to feature space and the use of the best margin allows for good generalisation to unseen data even if the data contains some bias and due to the nature of the optimisation problem the best margin will be the global minimum. However, SVMs may not be the best choice for a certain problem, they require both positive and negative training data, selecting the best kernel for the transformation is not always easy, there may be numerical issues in optimising the margin and the parameters of the final model may be unintuitive.

SVMs were first developed in 1963 for use in pattern recognition by Vapnik¹⁵², before being expanded into the most common implementation today in 1995¹⁵³. SVMs have a rich history in chemistry and have been applied to a wide range of problems^{154;155}. For example Burbridge *et al*¹⁵⁶ used an SVM for structure-activity relationship analysis in drug design, proving it outperformed other machine learning techniques in the prediction of inhibition of dihydrofolate reductase by pyrimidines. Borin *et al*¹⁵⁷ used a least-squares SVM as a calibration method for a near-infrared spectroscopy analysis of common adulterants in powdered milk. The three analysed adulterants exhibited non-linear spectral behaviour rendering the use of linear methods difficult. Their SVM could predict both the presence and absence of the adulterants outperforming a partial

least-squares regression method. Fatemi *et al*¹⁵⁸ used their SVM to successfully predict the selectivity coefficients of ion-selective electrodes based on the molecular structure alone.

4.1.0.2 Linear Regression

While linear regression comes from the field of statistics and may not seem a machine learning algorithm at first, it can be thought of as a supervised learner as it essentially attempts to understand the relationship between input variables and their output. As in the simple SVMs, linear regression is a linear model as it assumes a linear relationship between input and output or more specifically, that the output can be predicted from a linear combination of the input variables. If the relationship is one-to-one it is called simple linear regression and the plot of our function describing this will form a straight line (often called a regression line). If there is more than one input per output it is referred to as multivariate linear regression (which may plot a plane, a generalisation of a line). The representation of a simple linear regression problem is as below,

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \epsilon_i \quad (4.1)$$

where y is the output. β_0 is the intercept coefficient (allowing the line to move up and down); this is usually a constant and is included in many statistical model. β_1 is the coefficient of x_1 the input and ϵ_i is the error term. As more input variables are added they have their own β_n coefficients. The learning present in linear regression focuses on the estimation of these coefficient values from training data to make the best prediction of our output value on unseen data. The complexity of the model refers to the amount of coefficients present in it. Often complexity is reduced by regularisation methods by driving the value of some coefficients towards zero. The error term captures all other factors which influence y apart from the input variable x (for example, noise in measurements of the output variables). Discovering the relationship between the errors and the input variables can be important, especially if they are correlated.

The estimation of the coefficients can be performed by many means, the most common and simple is ordinary least squares (OLS). OLS works by minimising the errors/distance between the regression line and the observed values of y . Coefficients are determined that minimise the sum of the squared distances between the prediction and the observed for every point in the training data. The smaller the distances between the regression line and the points, the better the model fits the data and the better it can generalise to new data. To use OLS for a linear regression model, some assumptions need to be made about the data (all of these assumptions can be relaxed, usually with an increase in the complexity of the model).

In simpler models the input variables (x) are assumed to be measurement error-free and can be treated as fixed values rather than random variables. While not realistic in all problems dropping the error can result in underestimation of the coefficients and the OLS becoming invalid. The second assumption is that the mean of the output (y), is the result of a linear combination of the coefficients and the input variables, as the input variables are fixed values the linearity only really applies to the coefficients and the input variables can be transformed or even copied and transformed to allow for non-linear models to be used. The third assumption is that different output variables have the same variance in their errors, and the error does not depend on the value of the input variable. If the variance is too different for each output variable the model will look better than it actually performs. In addition to the variance of the errors of y they are also assumed to be uncorrelated as this can also introduce bias into the predictions. Apart from these assumptions, other properties of the data being studied need to be considered such as the relationship between ϵ_i and the input variables and the sampling of the input variables to ensure an accurate estimate of the coefficients. Many more complex methods of linear regression such as Bayesian, generalised least squares and weighted least squares have been developed to relax the assumptions made above.

Having been around for over 200 years OLS and linear regression has been a popular general prediction method almost since it was derived. Its use in chemistry may have peaked before more advanced machine learning methods were invented and now it is often used as the comparison for the "hot" method of the moment (as above with SVMs). The more complex linear regression models mentioned above have found further use however. For example, Riu¹⁵⁹ developed a bivariate linear regression method to estimate errors in analytical methods; this allowed the relaxation of the third and fourth assumptions mentioned above and can be used when errors may be present in the input and output variables. Koklay¹⁶⁰ developed a stepwise multivariate linear regression model to predict the nitrogen, lignin and cellulose content of leaves from reflectance spectra data across a variety of plant species.

4.1.0.3 Artificial Neural Networks

Neural networks (NNs) are inspired by the observation that biological learning systems are built of very complex webs of interconnected neurons. While a brain is a massively complex object the building blocks are surprisingly simple. A brain contains on the order of 10^{11} neurons each connected to 10^4 other neurons. Despite their relatively slow switching speed (10^{-3} seconds compared to 10^{-10} computer speeds) the brain can perform very complex tasks surprisingly fast. Recognising the face of someone you know can occur within 150 milliseconds of the image hitting the retina. Performing the task this quickly implies the sequence of neurons firing can only be a few hundred steps given the neuron switching speed. This led to the speculation that the information-processing abilities of biological systems is due to highly parallel processes operating on

representations distributed over many neurons. This is the main motivation behind the development of NNs, to try and take advantage of highly parallel computation based on distributed representations.

Typically NNs are made up of connected layers of simple processing units, there is no connection between units in the same layer, only to layers above and below. The signal typically travels front to back, from the input layer through the processing units (neurons) and to an output layer. Fig 4.2 shows this basic structure, but additional layers can be used and the connectivity of the network can be modified.

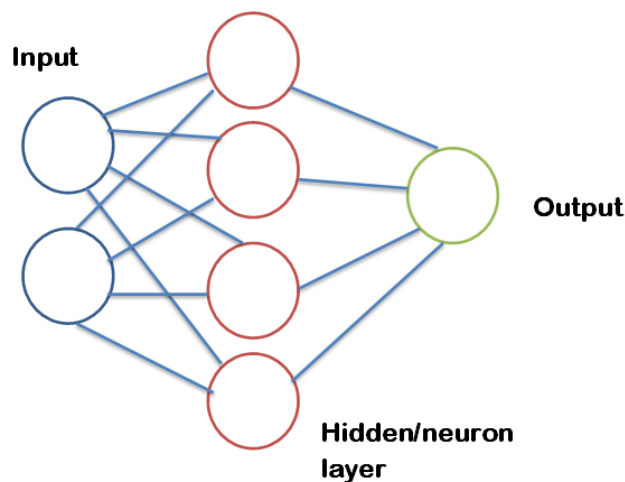


Figure 4.2: Structure of a basic NN.

The input layer does not carry out any arithmetic operations and its main function is to pass input values to the layer of neurons below with each part of the input layer connected to every neuron. The neuron layer is also commonly called a hidden layer, as it sends or receives signals to or from an outside source. The values that arrive to each neuron from the input layer are different as each connection has a different weight. These weights are usually determined in a learning process using the NN. The weights can be thought of as the synaptic strength linking real neurons together and allow the NN to choose which neurons are most useful for solving the problem. The neurons are where the computation of the NN occurs. Each neuron receives one or more weighted input values from the input layer, these weighted values are then summed and passed to the transfer function. The purpose of the transfer function is to introduce nonlinearity into the model. Without this mapping, neural networks can only identify linear relationships between the input and output layers, but even with just a single hidden layer that maps input values to a transfer function, they have been shown to be able to approximate any (well behaved) function to arbitrary accuracy, provided the network has enough hidden layer nodes. The function can take any form but typically a bounded differentiable function is used, such as a sigmoid function (commonly the arctangent function):

$$f(x) = \frac{1}{1 + e^{-x}} \quad (4.2)$$

$$f(x) = \frac{1}{\pi} \tan^{-1}(x) + \frac{1}{2} \quad (4.3)$$

Other functions can be used depending on the construction of the network. Once the transfer function has acted, the value can then be passed to the output layer, or another layer of neurons (Figure 4.3 shows the structure of a neuron/hidden node).

The training process of a NN is through the adaptation of the weights linking the layers until the correct output is achieved and choosing the structure of the network. Many different approaches can be used for this step including the use of genetic algorithms, simulated annealing or interval analysis. Regardless of what method is used, training a supervised NN involves the same basic steps. Weights are initialised randomly (often taking small values), the inputs of the training set are fed into the network and the resulting output calculated. The error between the NN output and the known output are calculated and weights are optimised to reduce this error. The process stops when the error begins increasing again, usually there are a few local minima in the weight space and the best set may be ones found in a reasonable time-frame. The weights are generally optimised (normally by gradient descent with some modifications like stochastic gradient descent) to reduce the average error (i.e like the RMSE) of the training set. Early stopping to prevent overfitting is then accounted for by monitoring the RMSE of the training set and the test set. As the weights are being optimised to the training set the training set error should always decrease (outside of numerical error). The test set error is likely to fluctuate, but if the test set error is increasing while the training set error continues to decrease overfitting to the training data is occurring.

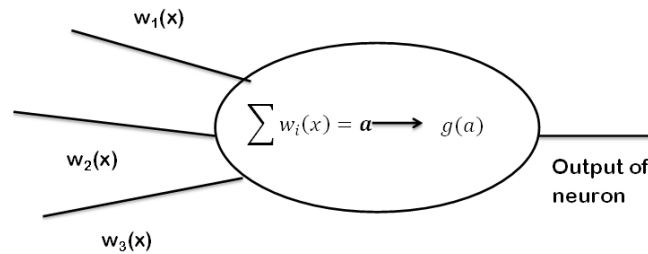


Figure 4.3: Structure of one neuron, where g is the transfer function.

Deciding whether to use a NN is (as with many machine learning algorithms) dependent on the problem you are trying to solve. NNs are well suited for input data which is complex and noisy, discovering unseen associations or regularities in data, and problems in which the relationship between variables in the data is not well understood or the relationships are difficult to describe. However there are some drawbacks. Most NNs

used today are extremely black box: once deciding on the random weights and architecture of the network the user has no other role apart from feeding input, watching learning and awaiting output. The NN cannot explain the relationship between your input and output: the network is the relationship. Therefore the form of the network provides no physical insight into the relationship between input and output. In addition training a NN can be a time consuming process if run on a single computer. Usually the evaluation of the network is fast once training has been completed. Another main drawback is they require a large number of training samples to optimise all the weights, as more hidden nodes are needed to reduce bias with the increase in training samples to keep variance low. Despite these small drawbacks, NNs have been used in many real world applications. ALVINN¹⁶¹ is an autonomous driving system which uses a learned NN to steer the vehicle. The NN was trained to mimic the observed steering commands of a human driver vehicle for 5 minutes. ALVINN eventually graduated to successfully driving on public motorways at speeds of 70 miles per hour. NNs have also found use in speech¹⁶² and pattern recognition¹⁶³ and the diagnosis of cancer¹⁶⁴. NNs have been applied to many problems in chemistry¹⁶⁵ including the prediction of molecular electronic properties¹⁶⁶, the lipophilicity of molecules¹⁶⁷ and the mechanical properties of polymers¹⁶⁸.

4.1.0.4 *k*-Nearest Neighbours

The *k*-nearest neighbours algorithm (*k*NN) is a method typically used for classification and regression tasks in which a data point is characterised by its nearest neighbours. In each case the input is *k* (typically a small, positive integer) closest training examples in the feature space of the data set. If used for classification, the output is which class the feature vector corresponds to, and this class is that which is most common amongst its *k* nearest neighbours. If used for regression the value associated with the feature vector is the average over the *k* nearest neighbours. In both cases, weights are often assigned to the neighbours based on their distance from the feature vector being analysed (typically $1/d$ where *d* is the distance to the neighbour). Neighbours are selected from already known examples, so they are the training experience of the algorithm. Figure 4.4 shows this in practice.

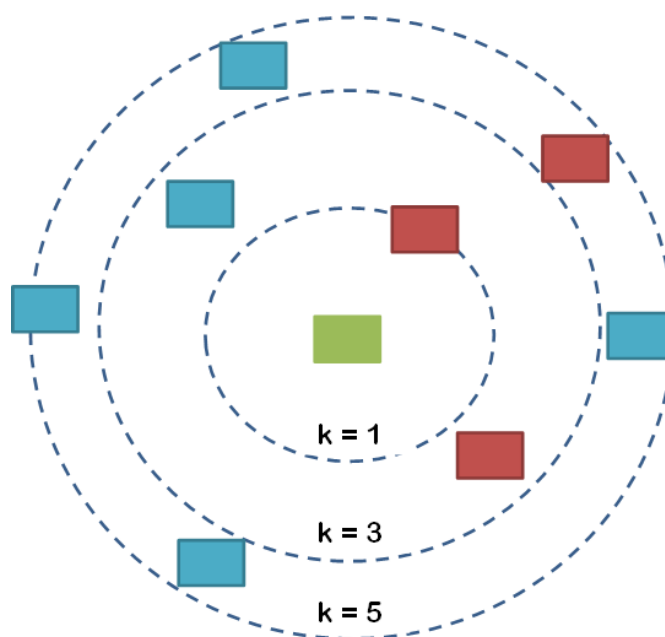


Figure 4.4: How the k NN works for classifying (or predicting) the colour of a rectangle. When $k=1$ the green query will be classified as belonging to the red class, when $k=3$ it is again red by a 2-1 vote, when $k=5$ it will be classified as blue by 3-2.

The k NN is sensitive to the local structure of the data which is a strength if the application requires the calculation of properties with a strong locality. Despite this locality k NN, can still give good global performance as long as the examples are distributed throughout the feature space. Distances are usually assumed to be Euclidean if continuous variables are used, but other metrics can be used depending on the problem being studied. The performance of the method can be improved dramatically if the distance metric is learned using a specialised algorithm. The best value of k relies on the data being used (and the task, as in binary classification problems an odd number is chosen to eliminate ties); noisy data may require a large value of k to reduce interference though this may make boundaries between classes less sharp. Often k is optimised using an internal validation method. It is often useful to scale the features present in the vector so that measurements in different directions are comparable. In addition, the presence of irrelevant and noisy features can severely degrade the performance of the algorithm so these may need to be removed.

k NNs has been applied to many classification problems in chemistry and they excel in the creation of quantitative structure activity relationships (QSAR). Kovatcheva *et al*¹⁶⁹ used a k NN to generate a QSAR for fragrant compounds in ambergris, outperforming a SVM method in the same study. Kühne *et al*¹⁷⁰ used their k NN to select the best method for predicting solubility of a molecule based on similar previously predicted molecules, similarity classified by the k NN. Chavan *et al*¹⁷¹ used a k NN in combination

with molecular fingerprinting to predict and classify the toxicity of a range of molecules. In an early application, Kowalski¹⁷² used a k NN as a comparison against a simple machine learning algorithm for the classification of NMR spectra with the k NN method performing better on the majority of spectra seen.

4.2 Genetic Algorithms

Genetic algorithms (GAs) use concepts from biological evolution for optimisation problems. Evolution presents an attractive inspiration for researchers looking to discover good computational optimisation methods. After all, evolution has been optimising biology for billions of years through the process of natural selection. Many different "solutions" are "tested" through genetic variance at the same time. These solutions are then evaluated based on their ability to survive their environment. Evolution can also design innovative solutions to complex problems, such as the mammalian immune system or the classic example of a giraffe's neck. The fitness landscape that biological evolution operates on depends on a large number of factors and is constantly changing. What may be good for the goose today may make the goose hopelessly outdated tomorrow. While the underlying landscape is complex, the rules of evolution are remarkably simple. The diversity we see on our planet today arises from random variation (through genetic recombination operators) and natural selection in which the fittest tend to survive and reproduce.

All of the above points have a direct analogy to many of the complex problems in biology, chemistry and finance. Many computational problems require searching huge number of possibilities for the best solution. One example is CSP in which millions of trial structures may be generated and lattice energy minimised in the search for the global minimum. Additional examples could be the search for a set of rules or equations to predict financial market behaviour, or the search for a sequence of amino acids leading to a protein with the desired properties. All these problems benefit from parallelism—the ability to evaluate many solutions simultaneously in an efficient way. In addition, many of these problems require the next set of solutions to be optimised to be chosen intelligently.

Many problems require the program to be adaptive, such as those in robot control performing a task in a variable environment. Other problems require the program to be innovative by creating new algorithms or new solutions. Many requirements or rules for the program are too complex to be encoded completely in the program (such as AI or the simulation of neurons) and must rely on emergent behaviour from a simpler set of rules. It can be seen that biological evolution as discussed above, can provide an adaptive, parallel search for the solutions to these complex problems.

In GAs our rules are those biological operators (loosely) borrowed from biological evolution. Exploration of the search space through variation (crossover or mutation) and emergent behaviour from the design of high quality solutions. While there is no rigorous definition of what exactly a genetic algorithm constitutes, most share the same basic building blocks. A population of candidate solutions (often called genomes), selection according to fitness, crossover to produce new solutions and the mutation of a small number of these new solutions. Solutions are usually represented in a way that is easy to crossover and mutate; most common are bit strings but many other representations have found use based on what problem is being investigated. Each solution is a point in search space, and the GA iteratively updates the population. To do this, the GA needs a fitness function to score each member of the population.

Fitness is how well the solution scores the problem at hand. The fitness function varies widely depending on what is being studied. For example, the protein design problem above may use the molecular energy of the proteins generated. However the fitness is calculated it is used to select which members of the population are used for crossover. A number of selection methods can be used, which can have a large impact on how well the space is sampled. In most examples of crossover two parent solutions produce two offspring solutions. These are collected into a new generation and fitness is calculated again. This process continues until some stopping condition is met, either a convergence of the fitness, a maximum number of generations or a lack of new unique solutions. This basic structure can be seen in Fig 4.5. Overall, a GA is a simple procedure analogous to the simple rules of biological evolution. The next section will discuss many of the concepts needed for a successful GA in more detail.

1. Generate a random population.
2. Evaluate fitness of each member.
3. Order by fitness and select parents of next generation.
4. Crossover selected parents to create next generation.
5. Pick new members to mutate.
6. Repeat steps 2-5 until stopping conditions are met.

Figure 4.5: Structure of a basic GA.

4.2.1 GA operators and concepts

The easiest way to explain the operators in a GA is to explain them as they appear in the running of a standard GA. The first choice to be made is how the problem will be

encoded. This is integral to the performance of any learning or search method. The most common (and earliest) method is the use of fixed length binary encodings, strings of 0s and 1s also called bitstrings. Much of the earliest work investigating suitable parameter values in GAs used bitstrings. However they are unwieldy for many problems (such as evolving neural network weights) and can be prone to arbitrary orderings. Other encodings including multiple characters or real numbers have also been developed. These encodings may be more useful for problems such as exploring chemical space, where elemental symbols can be used. Tree encoding systems used in genetic programming allow the search to be open ended as they have no limit on the size of the tree that can be formed. However large trees can sometimes become difficult to classify and uncontrolled growth can lead to unstructured solutions. Choosing the right encoding system can be difficult and Davis¹⁷³ proposes possibly the most sensible solution: choose the most suitable encoding scheme for your problem and develop the GA around that.

The initial population is often generated randomly, with the size of the population depending largely on the problem being studied. A smaller population can lead to the GA converging early on a local minimum, whereas a too large population can rapidly increase the cost of the GA. Sadly there is no formal way of choosing a population size, requiring it to be chosen empirically as the GA is being developed. While random generation is popular to try and ensure a good mix of starting population members, certain members can be seeded in the population to encourage evolution in a certain direction¹⁷⁴. Due to the inherent randomness present in many steps of the GA, multiple runs are often performed and average statistics reported.

The first generation of a GA is the simplest: after creation, each member of the population needs to be evaluated for its fitness. The fitness function is an important part of ensuring that the GA locates the best solution. As more and more solutions cross over, the fitness landscape becomes more complicated and it is possible that many solutions are very close to one another. Any accuracy lacking in the fitness function makes it possible that the GA will not find the correct minimum. This is less a problem of the GA and more the choice of fitness function. However fitness is calculated, the next step is the selection of solutions for crossover.

4.2.2 Selection methods

Selection is not just a process of pairing off the best solutions for crossover. Selection pressure is of major importance. This is essentially how hard the GA is driving towards the minimum. If the selection pressure is too high, the search will converge too early. If too low, the cost and time spent on the GA will increase. Fitness proportionate selection (FPS) was used in the first major GA¹⁷⁵ (to be discussed later): solutions are chosen probabilistically based on their fitness. The first step in selecting genomes for the next generation involves the normalisation of the fitness. In this method the probability that

a solution can be expected to crossover is that individuals fitness divided by the average fitness of the population. This method is often called roulette wheel selection, as each member of the population can be thought as a space on a roulette wheel. The size of each pocket is related to the fitness of the genome. Then a random selection is based off the spin of the wheel. Equation 4.4 shows this in practice,

$$p_i = \frac{f_i}{\sum_{j=1}^N f_j} \quad (4.4)$$

where f_i is the fitness of the i th member of the population, p_i the probability of that member being selected and N the number of individuals in the population. Figure 4.6 also shows how a member is selected. With this selection method higher fitness individuals are obviously favoured, although weaker individuals still have a chance to make it through. This is good, as a weak member may have a positive feature that can be expressed during crossover. Statistically, this method will always result in the expected number of crossovers for each member. However, the usually small population sizes of most GAs mean that sampling of members for crossover will often differ from their statistical weights assigned in (4.4). This could result in the space being saturated by high fitness members or low fitness members, neither of which is preferable for the diversity of the next generation.

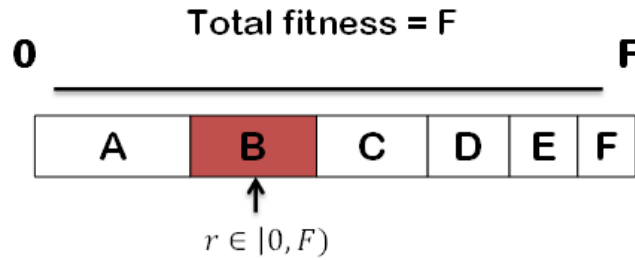


Figure 4.6: Example of the selection of a single individual using FPS.

To combat this Baker developed stochastic universal sampling (SUS) in 1987¹⁷⁶. SUS works in a very similar manner to fitness proportionate selection but aims to reduce the spread of the expected offspring value. Rather than spinning N times to select N members, SUS spins once but with N equally spaced pointers to select the N parents. Figure 4.7 shows the difference between SUS and FPS. SUS helps the spread of expected offspring but does not ameliorate the main issue with using fitness proportionate methods. At the beginning of a GA run, the spread of fitness will be high as some members are much fitter than others. Using a FPS method will result in these individuals and their descendants multiplying quickly in the population leading to premature convergence of the population. FPS as a method puts too much emphasis on exploiting highly fit individuals against exploration of other regions of the search space. Towards the

end of a GA run, when many population members are alike, there are no real fitness differences for the selection method to exploit and evolution can grind to a halt. One of the important factors for keeping a GA viable is the fitness variance across a population.

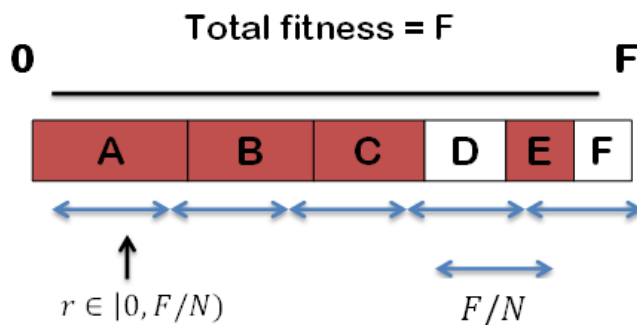


Figure 4.7: Example of selection of multiple members of the population using SUS. F/N is the width of the pointers. $r \in [0, F/N)$ the initial "spin" to choose where to begin selecting.

Another attempt to fix the fitness variance problem in FPS is the sigma scaling method¹⁷⁷. The fitnesses of members of the population are mapped to the expected value of offspring of each member. This keeps the selection pressure relatively constant across generations rather than relying on the fitness variance of the population. Using sigma scaling, the expected value of a genome is a function of its fitness, the population mean and the population standard deviation. At the beginning of the GA, the standard deviation of the fitnesses will typically be high, but the fitnesses should not be too many standard deviations above the mean so they will not dominate the crossover. When the population is more converged at the end of the run and the standard deviation is lower, the fitter individuals will stand out more, allowing evolution to continue. An example of sigma scaling is as below¹⁷⁸,

$$\text{ExpVal}(i, t) = \begin{cases} 1 + \frac{f(i) - \bar{f}(t)}{2\sigma(t)} & \text{if } \sigma(t) \neq 0 \\ 1.0 & \text{if } \sigma(t) = 0 \end{cases} \quad (4.5)$$

where $\text{ExpVal}(i, t)$ is the expected value of individual i at time t , $f(i)$ the fitness of i , $\bar{f}(t)$ is the mean fitness of the population at time t . This function would give an individual with fitness one standard deviation above the mean 1.5 expected offspring. If $\text{ExpVal}(i, t)$ is less than 0 it is set to 0.1 to give weaker population members a chance to produce offspring.

Different methods of selection have been developed that use no fitness scaling. Rank selection¹⁷⁹ was developed to prevent too quick convergence, and as the name suggests uses the rank of the individual in the population rather than a fitness value. This approach hides large absolute differences in fitness between population members (Fig

4.8), which avoids giving a large proportion of the offspring to a small number of highly fit individuals and reduces selection pressure when the fitness variance is high. It also helps keep up selection pressure when the variance is low, as the ratio of offspring between i and $i + 1$ will be the same irrespective of their absolute fitness difference. However, method is disadvantageous in situations where it is helpful to know one individual is much fitter than the rest of the population.

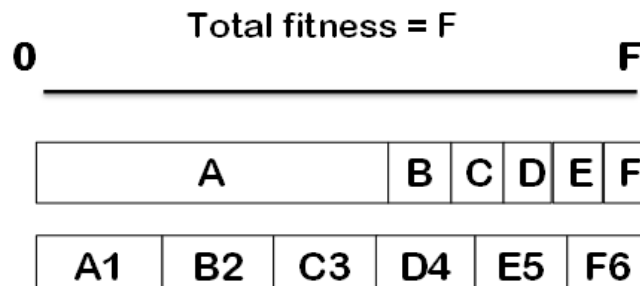


Figure 4.8: In this example, population member A would dominate the members chosen for crossover and negatively impact the diversity of the next generation. With rank selection A still has a higher chance to be selected but the fitness variance has been masked.

Tournament selection involves running tournaments among a number of (usually 2) randomly selected members of the population. A random number between 0 and 1 is selected and if it is less than a user defined parameter (usually around 0.75) the fittest individual in the tournament goes through to crossover, otherwise the less fit member will go through. Selection pressure can be easily adjusted by changing the size of the tournament as larger tournaments disadvantage weaker members. Tournament selection has many advantages compared to other selection methods: like rank selection it can mask large fitness variances, but avoids the potentially time consuming step of ranking all the solutions. Additionally, it is simple to code and can easily be ported to parallel architectures.

Elitism, first introduced by de Jong¹⁸⁰ is an addition to many selection methods. A small percentage of the fittest individuals are taken through to the next generation unchanged as these individuals may be lost through crossover or mutation. Elite individuals are still allowed to be chosen as parents for crossover. This can have a good effect on the performance of the GA as long as the number of elite individuals remains small in relation to the overall population.

Producing a new generation entirely made up of new offspring may not be the best strategy for certain kinds of problems. For the development of rules-based systems¹⁸¹, classifiers¹⁸² or genetic programming¹⁸³, it is advantageous to keep what already has been learned about the problem. To ensure this steady state selection is used, of each

generation only a small percentage of the population is replaced. These are usually the weakest genomes and are replaced by mutation or crossover of the fittest members.

4.2.3 Crossover and mutation

Crossover is where the magic of a GA happens. Here genomes picked for crossover are paired together and produce two offspring by copying selected parts of both parents. The chromosome at position i in the child is the same as position i from one of the parents. These two offspring hopefully explore a new part of the search space different from their parents. There are a number of different crossover operators which splice and combine different ratios of the parents.

The simplest is single point crossover: a random number n is chosen, and the parents cut at this point in the bitstring. The first offspring will contain the first n characters of parent one and the rest of the string made up of parent two. The second offspring will be the reverse of this: first n characters of parent two with the rest made up of parent one. Single point crossover does have drawbacks. Short parts of the bitstring are preferentially swapped with this method, as are the ends of each bitstring. It is possible that longer slices are needed for a fitter individual and that the ends are propagated through the crossover when they contribute little to the overall fitness of an individual.

To reduce this positional bias, many GAs also incorporate two-point crossover. In this operator two random numbers are selected, n_0 and n_1 , defining a segment in the parents to be exchanged. This results in child one containing upto n_0 of parent one, between n_0 and n_1 of parent two, with the rest being made up of parent one. Again, the process is reversed to produce the inverted offspring. Two-point crossover is less likely to disrupt large parts of the chromosome and can combine more parts of the genome together. In addition, the segments that are exchanged do not necessarily contain the endpoint of the strings. However there may be fit parts of the chromosome still not able to propagate this way.

Some researchers¹⁸⁴ believe that uniform crossover is the most suitable for all applications. In uniform crossover, exchange can happen at every point of the parents: for each point in the children a random number is chosen between 0 and 1. If it is above 0.5 take the character from parent one, if below from parent two. The second offspring is then the inverse of the first one. Uniform crossover has no positional bias and can propagate any fit section of the chromosome. However this is also a weakness of the scheme; any parts of the chromosome that have co-adapted can be easily destroyed. The three types of crossover mentioned here are demonstrated in Fig 4.9 using bitstrings.

Like many parameters of the GA, it is difficult to assume which crossover operator is the best *a priori*. The success or failure of a crossover operator depends on the fitness function, encoding and the problem that is being investigated. Uniform crossover may

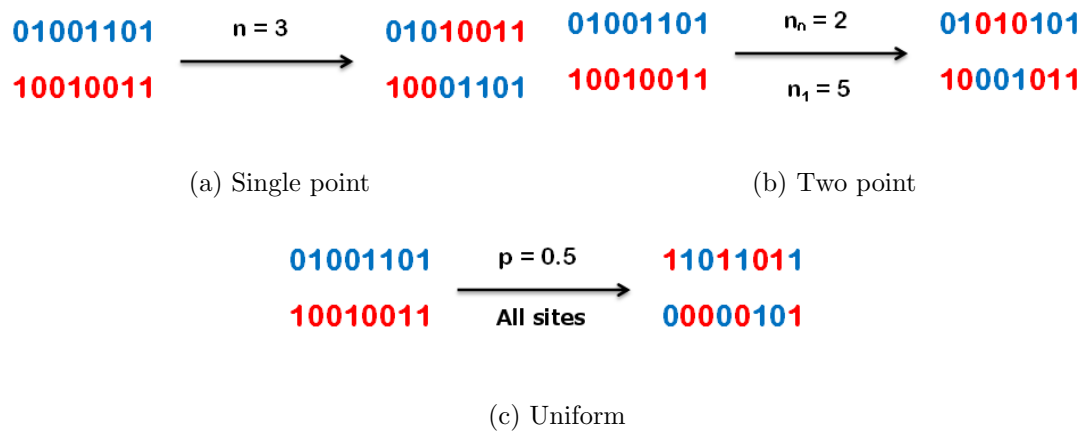


Figure 4.9: The three common crossover schemes seen in GAs, parent solutions on the left, new children on the right.

ensure each child explores a new part of the space, but if the best solution has features that interact with one another it may be unlikely to be made. Due to the variety of problems GAs are applied to, even empirical test sets shine little light on which one may be best. Many of today's GAs use either two-point, uniform or a combination of the three different operators.

After your new population has been created, there is a final operator applied before their fitness can be evaluated. Mutation changes one point in the bitstring of a chosen member of the population as seen in Fig 4.10. This usually occurs with a very small chance (0.5-1%) per generation. Mutation can be a helpful way to keep diversity present in a population and stop the GA focusing too far in one direction. There has been some debate^{185;186} about which is the more useful operator and which one should be favoured. However the success of a GA rarely relies on one parameter, it is the balance between selection, crossover and mutation that allows the GA to effectively explore the search space.

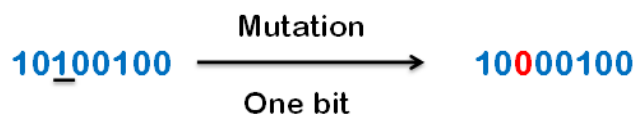


Figure 4.10: Mutation operator acting on a bitstring

4.2.4 Other operators and parameters

Crossover and mutation are the basic operators present in most GAs though the way they are applied is varied. Several operators have been developed to preserve diversity in

the population. De Jong¹⁸⁰ developed an operator where a new offspring would replace the existing individual most similar to itself. Fitness sharing decreases each individuals fitness proportional to how many other similar individuals are present in the population, punishing individuals similar to ones already present and favouring different individuals. Goldberg and Richardson¹⁸⁷ showed this could induce speciation in their population. Their GA converged to multiple peaks in the fitness landscape rather than one. Through a bidding process, Smith¹⁸⁸ showed this fitness sharing could be accomplished without the entire population calculations needed for a fitness sharing function.

Another way to focus on diversity is to put restrictions on which individuals can crossover with one another. If only sufficiently similar individuals are allowed to crossover speciation will occur¹⁸⁹. If the desire is to keep the overall population as diverse as possible, the opposite strategy can be applied^{190;191}.

Important parameters to consider in the construction of a GA are: population size, crossover rate, values inside whichever crossover method is chosen, elitism rate, mutation rate and the various parameters inside the chosen selection method. These parameters do not have a linear relationship in most problems, so can not be optimised one at a time. Aside from experimentation with the GA, most researchers stick to parameters that have worked well in the past. De Jong¹⁸⁰ tested various parameters on a small test suite of problems. Parameters were judged based on the average fitness of a population at a time and the overall best fitness seen. Conclusions from his study suggest a population size of 50-100, a single crossover rate of 0.6 per pair of parents and a mutation rate of 0.001 per bit. These settings became widely used despite doubts about their application to problems outside of the test suite.

In a matryoshka doll type insight Grefenstette¹⁹² proposed that a GA be allowed to evolve the parameters for another GA. In this study the overarching GA evolved a population of 50 parameter sets for the problems in De Jong's test suite. Each individual encoded six parameters and fitness was measured in the same way as De Jong's original study. The fittest individual called for a population size of 30, the crossover rate to 0.95 and the mutation rate 0.01. These parameters gave a small improvement over De Jong's best parameters. Again these values were quickly taken up, ignoring the applicability of these values to problems outside of a small, specialised test suite.

A similar study¹⁹³ on a slightly different test suite found parameter values close to those by Grefenstette. A population size of 20-30, crossover rate 0.75-0.95 and a mutation rate of 0.005-0.01. The population size is very small, but the authors may have biased towards smaller populations in their measurement of the average fitness. Due to the parameter values relying intrinsically on the problem studied and the fact the optimum parameter values may change during the course of a run, many have called for parameters that adapt over the life of the GA. Davis¹⁹⁴ created a steady state GA in which each operator is also assigned a fitness. This value is a measure of how many highly fit individuals

has been created using that operator. Operators gain fitness both for the production of good individuals and creating fit parts of the genome that may go into fitter individuals later. Each operator started out with the same initial fitness and at each time step an operator is chosen to create a new individual, which replaces a low fitness member of the population. Each individual keeps a record of how it was created, and if it is fitter than the current best member, the operator gets some credit as do the individual's ancestors. In theory, the fitness of the operators should match up with which one is the most useful at that stage in the search. Davis did show that this method improved the performance of his GA evolving the weights for a neural network.

4.2.5 When to use a GA

GAs have succeeded in many problems reported in the literature, but there are also examples of them performing poorly. Deciding whether a GA is suitable for the problem in question is not always straightforward. However there are some rough guidelines. If the search space is rough and hilly, not well understood, the fitness function is noisy and the task does not require the global minimum to be found (rather a family of solutions) then a GA can outperform many other search methods. If the search space is small, can be exhaustively searched to locate the global minimum then a simpler search method is probably best, as a known problem in GAs is premature convergence on a local minimum. If the space is smooth to one minimum then a cheaper gradient based method will be more efficient. If the space is well understood, then a method incorporating specific knowledge of that space can perform better. If the fitness function is noisy, however, a GAs collecting of fitness of many generations may outperform a single candidate at a time method which may be led astray by noise.

These are only guidelines for a prospective GA and the performance of the GA will depend on the construction of the GA and the choices made for the operators, fitness function and encoding.

4.2.6 Brief GA history

With a discussion of the basic building blocks of a GA complete a quick jaunt through the development of the field up to today will elucidate how to design a GA for a range of problems.

Like the overall field of machine learning, GAs got their start in the 1950 and 60s. The first computational study of evolution was probably the work of Barracelli in 1957¹⁹⁵. His trials were essentially simulations of artificial life, in which numbers were placed in a grid and moved according to local interaction rules. These numbers propagated throughout the grid and depending on how they interacted, could create new patterns

using rules of both of the parents. In this game the origin of Conway's famous game of life can be seen¹⁹⁶. Alex Fraser¹⁹⁷ published a series of papers on the selection of artificial organisms with multiple chromosomes controlling a measurable trait. Fraser's simulations contained most of the framework seen today in GAs as seen in Fig 4.11. At a similar time, Rechenberg¹⁹⁸ developed "evolution strategies" to optimise parameters for the design of aerofoils, which was further developed on by Schwefel¹⁹⁹. Bremermann²⁰⁰ also published a series of papers which adapted a population of solutions to an optimisation problem, with solutions able to undergo crossover, mutation and selection. This work led to the Bremermann limit, a theoretical limit to any computation.

Structure of Fraser's "Epistasis" program

1. Extract without replacement two random parents from the set.
2. Form a set of progeny from these parents.
3. Determine phenotypes of the progeny.
4. Select potential parents from the progeny
5. Repeat (1)-(4) until all parents have produced the specified number of progeny
6. Repeat (1)-(6) using the selected progeny as parents.

Figure 4.11: The structure of Fraser's evolution program, containing the foundation of today's GAs

All of this was focused on developing evolution inspired machine learning techniques. The first true GA was published by John Holland in the 1960s¹⁷⁵, and further developed by him and his students at the University of Michigan. In contrast to earlier work, Hollands focus was not to design specific algorithms for specific problems, but to study the process of adaptation as it occurs in nature and import this into computer systems. Holland's work was the first to attempt to put computational evolution on a theoretical footing. The introduction and use of the operators discussed in previous sections represented a large leap forward for the field. Rechenberg's work began with only two population member's, one parent and a mutated offspring, while other early work relied on only mutation to move around the search space.

GA research continued to grow and Holland's original framework was built upon by scientists from a range of different fields. The rest of this section will focus on the applications of GAs in chemistry, which has been growing almost exponentially since the early 90s, when Forrest²⁰¹ published a paper which gave a general overview of GAs and how best to use them. In 1995 Deaven²⁰² used a GA to optimise the energy of carbon clusters upto C_{60} . Using a starting population of random atomic coordinates (with some

rules) and relaxed molecular energies as a measure of fitness, this was the first successful calculation of the buckyball structure. Blommers²⁰³ *et al* used their GA to perform a conformational analysis of a dinucleotide photodimer. Judson²⁰⁴ *et al* also used a GA for conformational analysis of small organic molecules (1-12 rotatable bonds). This was combined with a local gradient optimiser and compared to a commercially available cheminformatics package. For molecules with more than eight rotatable bonds the GA was more efficient, and performed better as the number of rotatable bonds grew. There have been many other GAs used in conformational searches, many summarised in the 2001 Leardi review²⁰⁵.

More relevant to this thesis is the use of GAs in the design of novel chemical materials with desired properties. Venkatasubramanian and coauthors applied their GA to the design of novel polymers^{206;207}. An alphabet of symbols that represented chemical building blocks was used to encode molecules, while operators included single point crossover and some designed specifically for their study. Their GA performed well in locating target structures, though its effectiveness decreased as the search space increased in size. Unsurprisingly, they also discovered that the most effective parameter set for their GA varied with the search it was performing. Burden²⁰⁸ *et al* used a GA and a neural network to predict the bioactivity of a number of dihydrofolate reductase inhibitors. Sundaram²⁰⁹ *et al* followed a similar strategy in the design of fuel additives. Williams²¹⁰ (of W99 fame) developed a GA to minimise the energy of three dimensional intermolecular interactions for use in docking simulations, using chromosomes consisted of three genes representing rotation and three for translations. GAs are also an active avenue of research in protein folding^{211;212;213}, typically using a free energy based fitness function and some local optimisation when the GA has revealed promising solutions, these GAs have shown some success.

GAs have been applied to most of the problems in computer-aided molecule design and the book *Evolutionary Algorithms in Molecular Design*²¹⁴ provides a thorough overview as does *Genetic Algorithms in Molecular Modelling*²¹⁵, Mitchell provides an overview of the theoretical underpinning of GAs in *An Introduction to Genetic Algorithms*²¹⁶.

4.3 Conclusions

The first half of this chapter introduced machine learning, points to consider when choosing a machine learning algorithm, categorisation schemes for said algorithms and four popular methods used today. Machine learning is concerned with developing algorithms that can learn from training experiences and that perform better with time. The training of an algorithm can be supervised (in which known data outputs are supplied the learner), unsupervised (in which the learner must discover patterns present in the data itself) or through reinforcement (in which positive learning is rewarded). Machine

learning algorithms can be used for tasks in classification, clustering, density estimation and dimensionality reduction. Choosing an algorithm principally relies on the problem you want to solve and the data available to the learner. Getting the representation of the data correct is one of the most important steps in formulating a successful machine learning algorithm. Other issues to take into account include the balance of bias and variance, the complexity of the true function being learned, the dimensionality of the input space and the presence of noise in the input data.

The second half of this chapter covered genetic algorithms and the steps needed to create a successful GA. GAs provide an adaptive, easily parallelised search for complex problems using concepts from biological evolution. Usually a random population is generated for which each member's fitness is calculated. Population members are then selected for crossover. Selection methods vary in effectiveness but are all used to apply selection pressure to the GA. Once selected a new generation is created using one of more crossover operators. The choice of operators is largely problem dependent, as are many of the parameters needed through the GA. These are best chosen through testing before being used for a production run. Chapter 6 will describe the design and testing of a GA for the discovery of novel organic semiconductors.

Chapter 5

Structure Generation and Blind Test

5.1 Introduction

This chapter introduces some of the additional work done during the course of my PhD outside of the core aims of my project. Chapter 2 introduced both different methods for the generation of trial crystal structures and the blind tests of CSP approaches. In this chapter the testing and design of a novel trial structure generator is described. Initial testing was performed on nitrogen containing organic semiconductors (with the full CSP presented in the next chapter) with full testing performed on three separate molecules each representing a research interest of the group. The second part of this chapter will focus on the 6th blind test of CSP methods, one molecule from the test set was investigated by me in detail.

5.2 Structure generator

One of the first steps in crystal structure prediction is the generation of trial structures. Many different algorithms exist for this task, covering a wide range of techniques. The simplest methods involve a systematic grid search over the structural degrees of freedom. While such methods can be successful for a low dimensionality problem (and can slightly outperform random searches²¹⁷), they quickly become expensive as the number of degrees of freedom increases. Random searches, as the name suggests, use random values for the parameters involved in generating the crystal, such as unit cell lengths, angles, molecular positions and orientations. However, random sampling can be inefficient (a "clumpy" distribution is common) and fail to uniformly sample the potential energy surface. Random searches can be improved, by deciding on how "random" the numbers

really are. True random numbers have three characteristics: they are unpredictable; uncorrelated and unbiased. Random number generators provide pseudo-random numbers, they are not unpredictable but are generated by a definite arithmetic method. Quasi-random numbers are predictable and correlated but unbiased, as they are spread as evenly as possible. The discrepancy of the sequence is a measure of how well this uniform distribution is maintained with a low discrepancy being favoured. Quasi-random sequences are designed to be low discrepancy and have found use in crystal structure prediction.^{45;218} Both previous implementations of quasi-random searching in CSP use the Sobol’⁴⁶ sequence, which has been shown to perform better than other sequences²¹⁸.

An in-house structure generator method was developed which uses Sobol’ numbers to sample unit cell lengths, angles and molecular positions. Rotations also make use of the Sobol’ sequence with the Shoemake method²¹⁹ defining how these are converted to rotations.

Space group symmetry is used in generating structures. Once the unit cell and asymmetric unit are defined, the full crystal structure is built using the space group symmetry operators. Therefore, independent searches are performed in each chosen space group. In detail, each independent molecule in the asymmetric unit requires three parameters for its position and three parameters for its orientation. Up to six further parameters are needed to describe the internal angles and lengths of the unit cell; six for the case of a triclinic cell, though all other lattices have free parameters due to restrictions on lengths or angles. The Sobol’ numbers are generated in the range [0,1] and a n-dimensional search requires a length n vector. Each of these parameters is associated with a quasi-random number from the Sobol sequence, but not all are independent.

Sampling the molecular positions involves the mapping of three random numbers to the three positions of a molecule’s centroid. Each element of the Sobol’ vector is taken as a position in fractional coordinates along an axis of the unit cell. As mentioned above, orientations are converted using the Shoemake method using quaternions. Quaternions have the advantage of capturing orientations and rotations in a relatively simple (four numbers versus a nine number orthogonal matrix) way. They have found use in computer graphics applications and have previously been used in the generation of molecular dimers²²⁰ and to calculate the root-mean-square-deviation (RMSD) between molecules²²¹. After the asymmetric unit has been created other molecules in the unit cell are generated by applying space group symmetry operators to it.

Next the unit cell angles are sampled. Each angle that is not constrained by space group symmetry is sampled as:

$$\theta_i = \left(\frac{1}{n} \arccos(1 - 2x_i)\right) + \theta^{min} \quad (5.1)$$

where, θ_i is the angle being sampled, $n = 2$, $\theta^{min} = \frac{\pi}{4}$ and x_i is the relevant element of the Sobol' vector. This equation gives an even distribution in $\cos(\theta_i)$ through the range $\theta^{min} = \frac{\pi}{4}$ to $\theta^{max} = \frac{3\pi}{4}$ with a probability density highest at the centre. This equation was chosen to balance sampling a range of angles and avoiding problematic non-physical structures. Particularly in in triclinic cells, many options for the unit cell have very acute (or obtuse) angles, which are computationally inefficient and awkward to lattice energy minimise. These "flat" cells were a particular nuisance during the initial design stage of the structure generator and the process of biasing away from them will be discussed later on.

So far, no information about the molecule has been used in generating the crystal structure. Choosing the bounds on the cell lengths, however, requires some system-dependant information be used. Pidcock²²² developed a box model to rationalise packing motifs seen in a range of experimental crystal structures across a number of space groups. This model established relationships between molecular dimensions and unit cell lengths which is useful for our purpose. When given an input structure, space group and a target volume, a box is generated around the molecule bounded by the (quasi)random lengths and Z. A target volume for each unit cell is calculated as the sum of all the molecule volumes in the cell multiplied by a constant, which we call the target volume parameter (TVP). The default value for is TVP 1.0 but may need to be varied depending on the system studied. Each molecular volume is calculated as that of a box enclosing all the atoms of the molecule. The box is defined by the axes of inertia of each molecule and finding the maximum and minimum value of projection of each of its atomic coordinates onto the axes using standard van der Waals radii. The difference between maximum and minimum value of the projection onto the axis is called the "shadow" onto that axis. Using a box to measure molecular volumes may overestimate compared to a more normal measure of volume, but in CSP we expect the cell to contract when lattice energy minimised and the extra volume in the generated structure is expected to allow easier sampling, while avoiding molecular clashes.

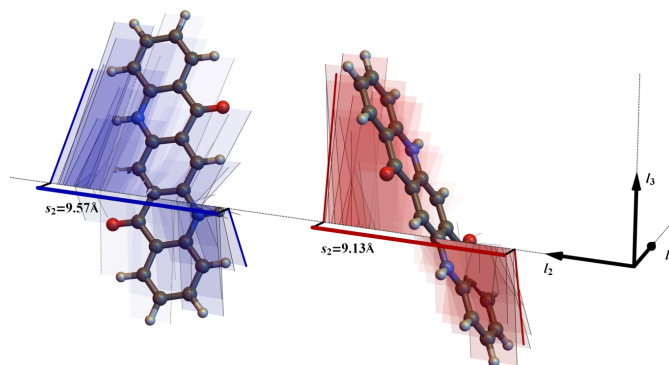


Figure 5.1: Molecular projections onto the lattice vectors, used to define the sampling range for unit cell lengths. The directions of the three lattice vectors, $l_{1,2,3}$ are shown, and the molecular projections of two quinacridone molecules are shown onto lattice vector l_2 . Thin lines show the projection of the edges of the van der Waals radii of each atom onto the lattice vector. Bold red and blue lines show the molecular shadows onto l_2 . This figure is adapted from reference [223](#)

The bounds of the three cell lengths are calculated considering this target volume and the projections of the atomic positions onto the lattice vectors (Fig 5.1). As the unit cell angles have been already determined, the direction of each unit cell vector can be fixed in a global axis frame and the "shadow" of each lattice vector on the global axes can be calculated. Each molecule in the unit cell that differs by rotation must be considered separately and the minimal (s_j^{min}) and maximal (s_j^{max}) molecular projections for each vector (j) must be found. To ensure a realistic sampling, the first unit cell length is chosen in the range from cs_j^{min} to $cN^{mols}s_j^{min}$, where N^{mols} is the number of molecules in the unit cell and c is a constant (value of 0.75) used to scale the sampling:

$$l_j = c(s_j^{min} + x_i(N^{mols}s_j^{max} - s_j^{min})) \quad (5.2)$$

where x_i is the relevant element of the Sobol' vector. The second unit cell vector is sampled the same way, with the third chosen to give a normal distribution of the cell volumes centred on the target volume described above. This sampling only changes when lattice types impose restrictions on cell lengths and fewer independent lengths must be sampled. The development of sampling cell lengths will also be discussed later.

Once the unit cell has been created, the structure is now ready for energy minimisation. However, not all structures will be physically sensible. If molecules overlap in a generated crystal structure they should be either rejected or adjusted to remove the overlap. The separating axis theorem (SAT) [224](#) is used to judge overlap. The SAT determines whether two convex shapes are intersecting. Initially, a convex hull [225](#) for the molecule is generated which contains the shape of the molecule, it may not contain a vertex per

atom but all atomic positions lie within its volume. The use of convex hulls is efficient for larger systems (because the number of vertices grows slower than the number of atoms) and rigid molecules when it only needs to be calculated once. SAT allows us to determine whether two convex hulls are overlapping and the shortest displacement vector to remove the overlap can be calculated. A set of vectors is chosen (either normal to the faces of a hull or to an edge of each) upon which to project the "shadow" of each convex hull. The overlap of the two shadows is measured and the minimal length of the overlap along any vector gives us the smallest vector required to separate the hulls (Figure 5.2). SAT can be used to simply reject any structure that contains overlapping molecules. The proportion of rejected structures typically falls as TVP is increased (increasing the target unit cell volume) but this may lead to an increase in the time spent on lattice energy minimisation, as structures may be further from their minima than they would be with smaller target volumes. An alternative use of SAT is to adjust the trial structure to remove the overlap of the molecules. This is done by expanding the unit cell lengths in proportion to the overlap vector. To stop cells growing too large a maximum volume parameter is also implemented, which is usual set as 2.5 times the sum of the molecular volumes in the unit cell.

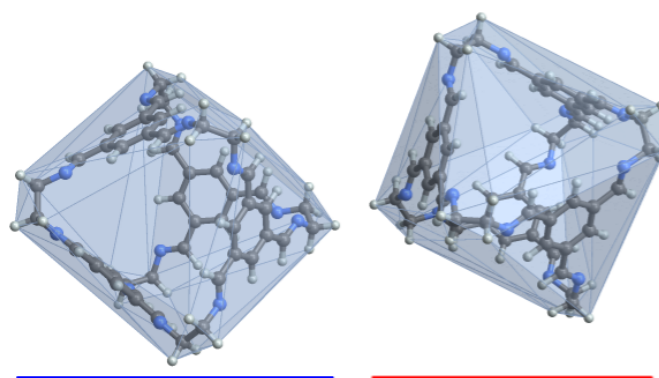


Figure 5.2: SAT test for molecular overlap. SAT prescribes the vectors upon which to project the vertices of the convex hulls when testing polytopes for their overlap in space. An example for the cage molecule CC1 is shown with convex hulls overlaid on the molecular geometry. In the geometry shown there is a vector upon which the "shadows" of their hulls, the blue and red vectors, do not overlap. If they did overlap, the set of overlapping blue and red vectors would determine the minimum displacement necessary to separate them in the direction of that vector. This figure is adapted from reference [223](#)

The structure generator has been used in different ways since it was first developed. Initially a set number of structures were generated in a space group and then lattice energy minimised. The total number of valid (ones that successfully reach an energy minimum) structures was whatever came out of that, and if this needed to be increased the structure generation was restarted from the highest Sobol' seed used so far. Recent

developments in our overall code now continues until a desired number of valid, energy minimised, structures is reached.

The next section will discuss early work on the structure generator code that gave rise to the cell angles and cell length sampling seen above. The section after that will discuss the full testing of the structure generator on three different interesting molecules.

5.2.1 Initial testing of the structure generator

The first real hiccups in the development of the structure generator were when it was tested on a number of azapentacene molecules previously been studied using Materials Studio Polymorph Predictor⁴⁸. When testing molecule 7A (Fig 6.3d) problems began to appear. Very "flat" unit cells were being generated and accepted. With a molecule pointing out of the cell and through perhaps 4 neighbouring cells, lattice energy minimisations failed or produced interesting though chemically implausible structures (see Fig 5.4). Initially the cell angle sampling proceeded as in equation 5.3, where Θ_i is the final angle, x_i a Sobol' number between 0 and 1, $\text{maxval} = 135^\circ$ and $\text{minval} = 45^\circ$.

$$\Theta_i = x_i * (\text{maxval} - \text{minval}) + \text{minval} \quad (5.3)$$

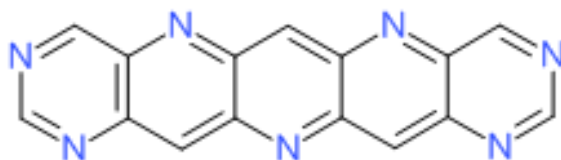


Figure 5.3: 7A one of the azapentacenes studied in other parts of the thesis.

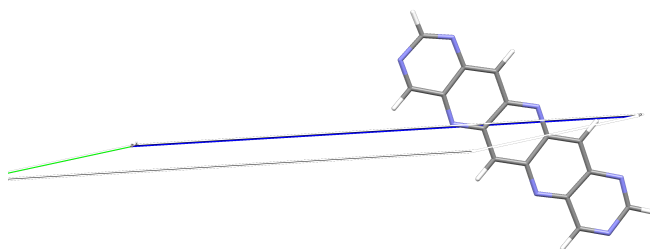


Figure 5.4: Typical flat cell produced.

Flat cells are particularly prevalent among the structures generated in spacegroups $P1$ and $P\bar{1}$. This was problematic as ca. 40% of $Z'=1$ and ca. 20% of $Z'=2$ structures that made it into DMACRYS from Polymorph Predictor were from those space groups. To prevent these squashed cells from being generated, the lattice deformation parameter (ldp)⁴⁵(eq 5.4 where ω = the unit cell angles) was placed in the code as a test on generated cells. Unit cell volume is equal to $a \times b \times c \times ldp$ so, ldp quantifies the unit cell volume relative to an orthorhombic unit cell with the same cell lengths. A value of 0.1 was chosen as the minimum allowed value for the ldp, initially based on observed values from squashed structures.

$$ldp = \sqrt{1 - \sum_{a=1}^3 \cos^2 \omega_a + \prod_{a=1}^3 \cos \omega_a} \quad (5.4)$$

Instead of settling on this arbitrary value, an examination of experimentally observed triclinic cells in the CSD (release 5.34) was performed. 67896 structures in total were examined and the distribution of the ldp plotted as a histogram (Fig 5.5). A comparable histogram (Fig 5.6) was made with 2000 generated structures in $P1$ for molecule 7A.

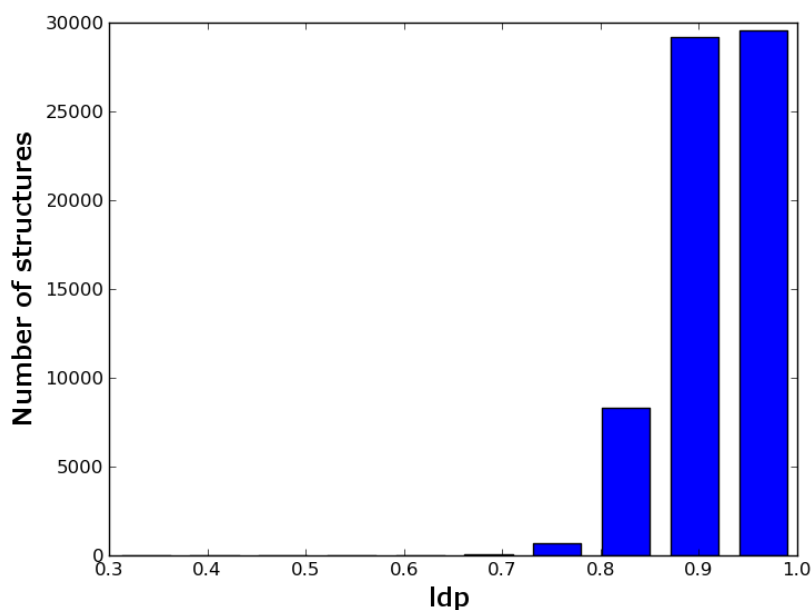


Figure 5.5: Distribution of ldp from the CSD (5.34) (search restricted to organic molecules in a triclinic space group)

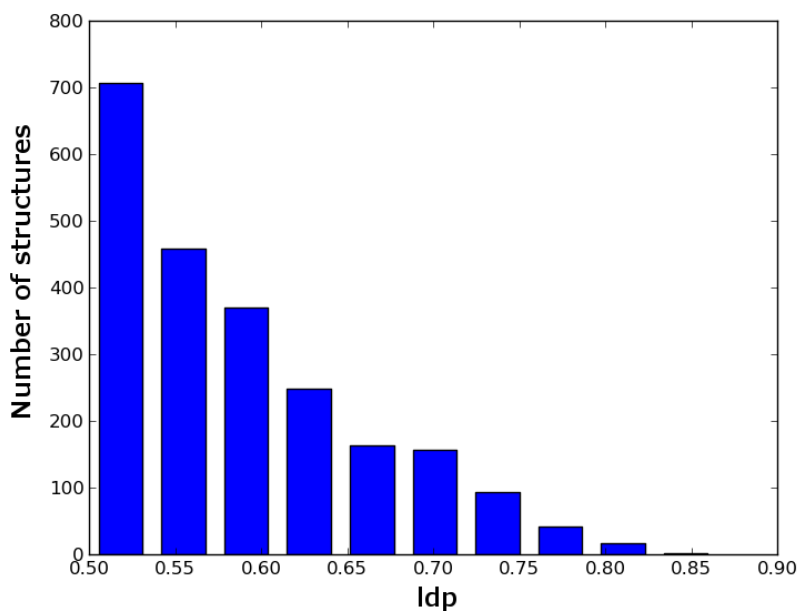


Figure 5.6: Initial ldp distribution of structure generator for 2000 structures in *P1* for molecule 7A.

A minimum value of 0.5 was chosen. While values from the CSD are largely confined to 0.8-1.0, picking a value near these would limit the search space too much, possibly excluding interesting structures. However, this did not entirely solve the problem. Structure generation in both triclinic spacegroups was glacial due to the sheer number of structures being rejected for having a too low an ldp value. The spread of ldp is almost opposite to cells in the CSD (Fig 5.6) demonstrating the sampling of unit cell angles needed adjustment. Since many structures were still being generated that failed the ldp test, it was decided to modify the sampling of the unit cell angles. Flattening the curve of 5.3 as it passes between the bounds would have the effect of improving sampling of plausible angles. Equation 5.5 accomplishes this (taking the same variables as eq 5.3).

$$\Theta_i = \frac{\frac{1}{2} \arcsin(2x_i - 1)}{\pi} \times (maxval - minval) + minval \quad (5.5)$$

Plotting both functions clearly shows the difference in sampling (see Fig 5.7).

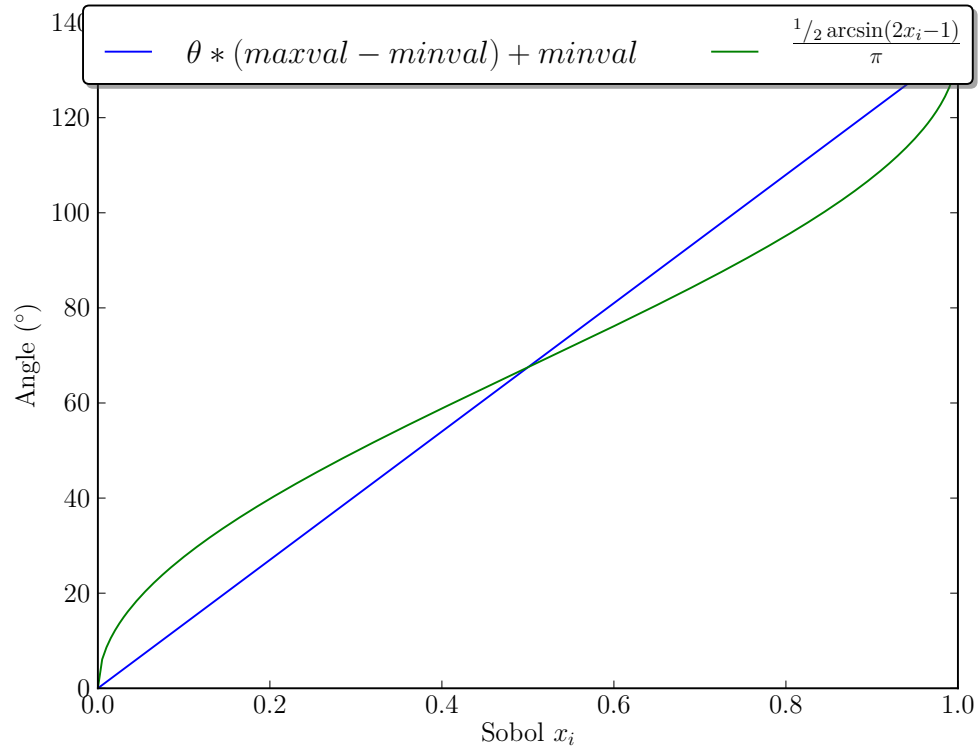
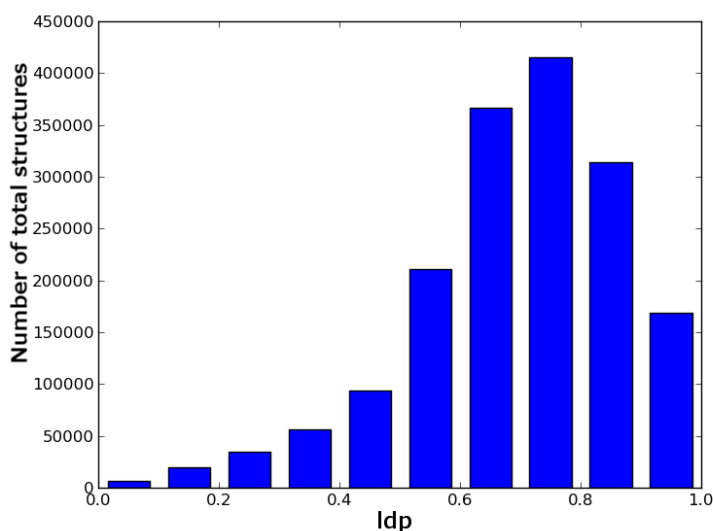
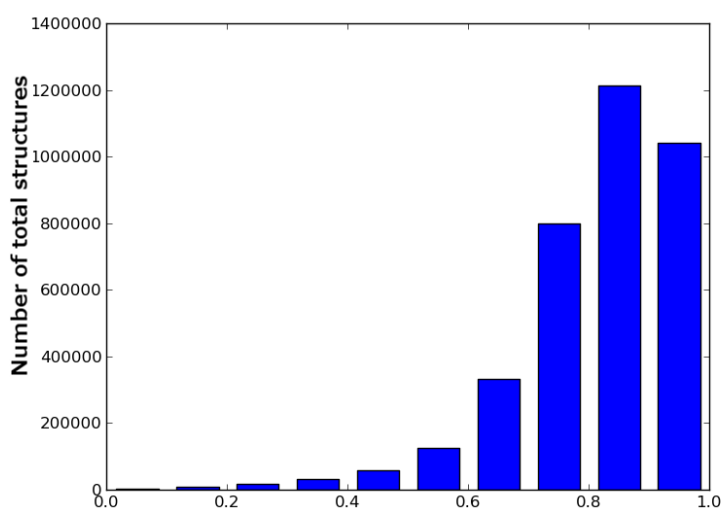


Figure 5.7: Plot showing the differences between sampling functions

1000 structures were generated with both methods and ldp from both accepted and rejected structures plotted (Figs 5.8, 5.9). Even though the new method used more trial structures (3631733 vs 1686645, these numbers being the total number of Sobol' vectors sampled to get 1000 accepted structures), it was much quicker.

Figure 5.8: Linear sampling ldp distribution (7A *P1*)Figure 5.9: Arcsine sampling ldp distribution (7A *P1*)

However, there still remained the problem of the long axis of the molecule possibly pointing along the shortest axis of the unit cell, so that a molecule could span several cells. The original cell length sampling used the three dimensions of the molecule (long, middle and short) and *Z* (number of molecules in the unit cell) to define the allowed ranges on the unit cell lengths. A new implementation uses the unique space group operations to project each molecule in the unit cell onto the vector defining the direction of the unit cell axis and setting the bounds on the length of the axis from the magnitudes of these projections. The van der Waals radii of the atoms within the molecule define

the shape to be projected and for each symmetry operation a value is obtained. The maximum and minimum projections define the bounds of the cell length. For the first cell length, the procedure is given in 5.6 where l is the length, x_i is the value from the Sobol' sequence and Z is the number of molecules in the unit cell.

$$l_1 = \min(l) + x_i(l) \times (Z \times \max(l) - \min(l)) \quad (5.6)$$

The second length is then bounded by equations (5.7) and (5.8), then sampled using (5.9). S, M, L refer to the short, medium and long lengths of a box created around the molecule and V^* has the same value as the lattice deformation parameter, being the volume of a cell with unit lengths.

$$\text{lower bound} = \max(\max(l) \text{ and } \frac{S \times M}{V^* \times L}) \quad (5.7)$$

$$\text{upper bound} = \min(\frac{Z \times M \times L}{V^* \times l_1}) \text{ and } Z * L \quad (5.8)$$

$$l_2 = \text{lower bound} + \zeta(l) \times (\text{upper bound} - \text{lower bound}) \quad (5.9)$$

The third length is then chosen to create a gaussian distribution around the target volume. However if the random length is less than the minimum value from the axis shadowing operation, the random length is set to the minimum shadow value. To remove any bias from picking the longest first, the order of projection is shuffled using the Sobol' sequence. The cells created with this method have significantly larger volumes than before and this necessitated a change in picking the target and maximum volume of the cell. Automatic target volumes were calculated in a similar way to the original cell lengths method, involving Z multiplied by the box lengths.

It can be seen from the introduction to our method both these changes were implemented into the final method in (slightly) modified ways. Once these problems had been ironed out a test set of new molecules was created for in-depth testing.

5.2.2 Quinacridone and full testing of rigid structure generation

Three molecules (Fig 8) were chosen for our new test set representing three research themes in the group, as well as diversity in terms of shape and intermolecular interactions. Quinacridone is a industrially produced pigment used in high performance paints and four polymorphic forms are currently known²²⁶. Due to its insolubility, all initial characterisation was done using powder X-ray diffraction, leading to the erroneous identification of additional polymorphs. Quinacridone displays photoconductive and photovoltaic properties, shows remarkable air stability and has been fabricated into devices^{227?}. Theoretical work has been done on the charge transport properties of the individual polymorphs²²⁸. Three of the polymorphs have the expected n-character

(electron transporter) and good overall mobility, while one may be used as an ambipolar (both n and p-type) charge transporter. Artemisinin is a pharmaceutical used in the treatment of malaria²²⁹ being the most used anti-malarial treatment today (in a combination with other therapies to reduce the risk of parasite immunity) and shows interesting anti-cancer properties²³⁰. Isolated from sweet wormwood the discovery of its anti-malarial properties won Tu Youyou half of the 2015 Nobel prize in medicine. The third molecule (CC1) is one of series of porous organic cages that had previously been studied using simulated annealing²³¹. These cages are solution processable porous materials and CC1 can switch between porous and non-porous polymorphs.

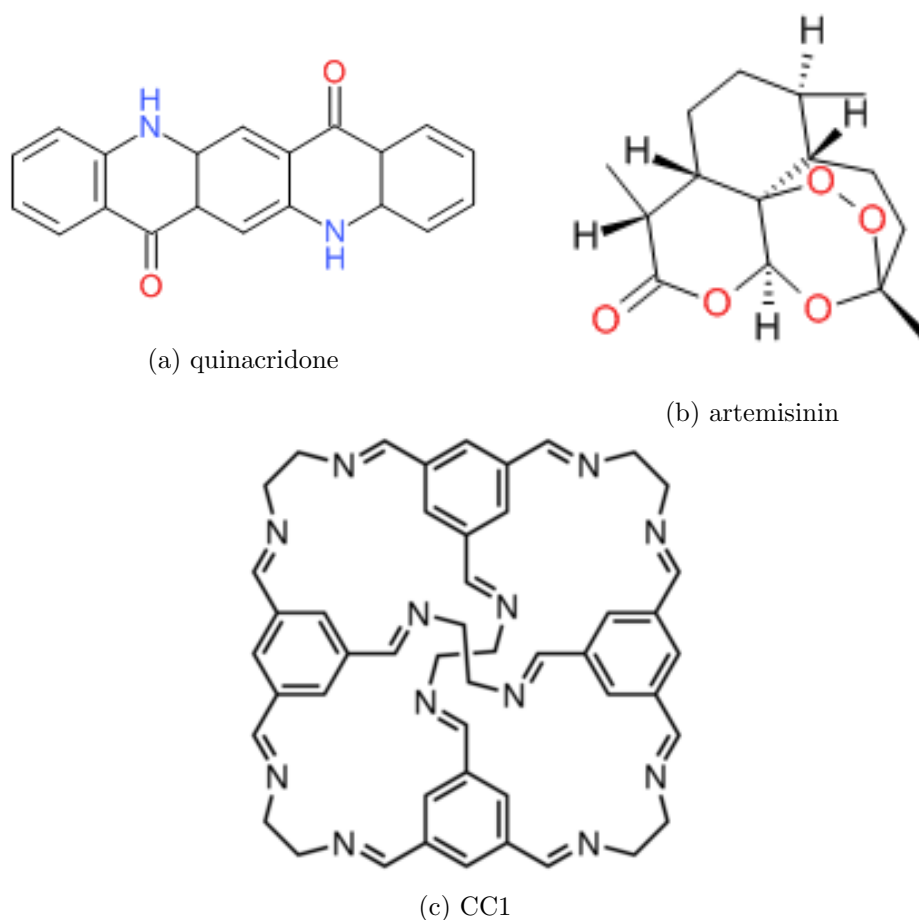


Figure 5.10: The three molecules chosen to test the structure generator.

I performed all the quinacridone calculations and half of the CC1 work that will be presented here. Discussion of the artemisinin results that were produced by Dr David Case will be included where appropriate.

For each molecule, every space group with a known polymorph was sampled. For quinacridone $Z' = 1$ this was $P2_1/c$ and $P\bar{1}$: the γ and β polymorphs appear in $P2_1/c$ while α_I appears in $P\bar{1}$. There is also a $Z' = 2$ polymorph that appears in $P\bar{1}$ so a search was performed with two molecules in the asymmetric unit. For CC1, searches

were performed in $P1$ and $P2_1/c$ with $Z' = 1$. Artemisinin was sampled in $P2_12_12_1$ for $Z' = 1$ and $P1$ was sampled for $Z' = 4$ polymorph. For each search, five sets of trial structures were generated using different target cell volumes, 1.0, 1.5, 2.0 and 2.5 where trial structures with molecular clashes were rejected. An additional set was generated using SAT to expand structures with overlapping molecules rather than rejecting them, and a TVP of 1.0. SAT was allowed to expand structures to a maximum of 2.5 times the initial target volume. If molecular overlap could not be removed with this expansion, the trial structure was rejected. For each $Z' = 1$ search 10,000 structures were generated. For the higher Z' searches 50,000 accepted structures were generated.

5.2.2.1 Lattice Energy Minimisation and Clustering

All lattice energy minimisations were performed using the DMACRYS⁶⁶ crystal structure modeling software, which employs a quasi-Newton Raphson, rigid-molecule optimization of molecular positions, orientations and unit cell parameters with space group symmetry constrained. The intermolecular interaction energy between molecules M and N was modelled with an anisotropic model potential of the form:

$$E_{MN}^{\text{intermolecular}} = \sum_{i,k} A^{\iota\kappa} \exp(-B^{\iota\kappa} r_{ik}) - C^{\iota\kappa} r_{ik}^{-6} + E_{ik}^{\text{elec}}(\text{DMA}) \quad (5.10)$$

where i, k are atoms of type ι and κ belonging to molecules M and N , respectively, separated by the distance r_{ik} . The first two terms model the repulsive and attractive non-electrostatic intermolecular interactions, whose parameters are taken from a revised version^{232;233} of the Williams99 force field²³⁴. The final term, describing electrostatic interactions, is calculated from atom-centered multipoles up to rank 4 (hexadecapole) on all atoms, obtained from a distributed multipole analysis[?] (DMA) of the B3LYP^{235;236}/6-311G(d,p) charge density. Charge-charge, charge-dipole and dipole-dipole interactions were calculated using Ewald summation, while repulsion-dispersion interactions and all higher multipole-multipole interactions were truncated after a cutoff distance. The summation cutoff (for exp-6 interactions and higher-order multipole-multipole interactions) was set to 30 Å for CC1 and quinacridone.

After lattice energy minimisation many trial structures may have converged to the same minimum, so duplicates can be removed and the lowest energy example of a cluster kept. As part of the Global Lattice Energy Explorer (GLEE, of which the structure generator is a part) a clustering method was developed based on inter-atomic distances and performing an overlay of molecules. A cluster of symmetry related molecules is created around the asymmetric unit (usually 25 molecules) and a list of atom separation distances to the asymmetric unit is calculated. Once these lists have been computed for two crystals they can be compared by placing both origins at the same point, and testing whether for every molecule in one cluster a molecule exists in the second that can

be overlaid onto the first through rotation only. The RMSD can be computed between the two clusters and if they match they are considered the same minimum structure.

5.2.2.2 Results and discussion

The first performance aspect of the structure generator to be analysed is the efficiency of the Sobol sequence with each TVP value. The low discrepancy nature of the sequence is meant to uniformly sample the configurational space. However, high rejection rate could harm the uniform sampling. Table 5.1 shows the number of trial structures required to generate 10,000 accepted structures for each TVP value and search (50,000 for $Z' = 2$).

Search	CC1 ($P1$)	quinacridone ($P\bar{1}$)	quinacridone ($P2_1/c$)	quinacridone ($P\bar{1}$, $Z' = 2$)
TVP = 1.0	17863 (9581)	251805 (9804)	501181 (9767)	96693852 (32021)
TVP = 1.5	17225 (8999)	55400 (9827)	78400 (9533)	8325359 (46213)
TVP = 2.0	17215 (8274)	30696 (9761)	36348 (9315)	1057042 (46443)
TVP = 2.5	17215 (7681)	23262 (9651)	24135 (9075)	504452 (45714)
SAT-expand	10090 (8514)	25131 (9862)	26617 (9353)	480626 (43541)

Table 5.1: Number of trial structures required to generate 10000 accepted crystal structures (50000 for $Z' = 2$) for each system. $Z' = 1$ unless otherwise stated. The number in parentheses is the number of accepted structures that lead to a successful lattice energy minimization.

Unsurprisingly the number of rejected structures decreases as the target volume increases. For all molecules, the chance of molecular overlap decreases with an increase in volume per molecule. It can be seen that there is a large difference in the rejection rate between molecules. This is due to the differences in molecular shape. Despite the features introduced during initial development of the structure generator long, flat molecules such as quinacridone still lead to a large fraction of rejected structures as long, thin molecules can clash with neighbours in most orientations. The structures created for the almost spherical CC1 contain much fewer rejections (as does artemisinin). Variations are also observed between space groups of the same molecule. Quinacridone $P2_1/c$ has more rejected structures than the simpler $P\bar{1}$. More overlaps can occur if molecules lie close enough to a symmetry element (this also holds true for the other CC1 search in $P2_1/c$). With increases in target volume, the differences in rejection between space groups begin to disappear. The increase in TVP from 1.0 to 1.5 brings the acceptance ratio of both quinacridone $Z' = 1$ searches from ca. 4% ($P\bar{1}$) and 2% ($P2_1/c$) to ca. 17% and 12%. When the maximum volume parameter of TVP = 2.5 is used acceptance ratios for searches are over 40% for CC1 and quinacridone. The $Z' = 2$ search is particularly difficult at low TVP values with the generation of 50,000 structures requiring almost 10^8 trial structures. As for $Z' = 1$, the trend follows that increasing the TVP will result in higher acceptance ratios: a TVP of 1.0 results in an acceptance ratio of only 0.033%, while at TVP = 2.5 it is 9.06%.

While increasing the TVP makes more efficient use of the Sobol sequence, it can be seen that the number of valid minimisations decreases at the same time. Trial structures with smaller volumes are closer in cell lengths and angles to the lattice energy minimised structures. When $TVP = 1.0$, over 95% of structures result in a successful lattice energy minimisation. Larger volume unit cells pose more problems during lattice energy minimisation, with the largest molecule (CC1) particularly suffering from this (9581 successful minimisations at $TVP = 1.0$, 8514 at $TVP = 2.5$). This trend is slightly different for $Z' = 2$ as the $TVP = 1.0$ search struggles to make valid trial structures and those structures are difficult to minimise, for $Z' = 2$ in particular $TVP = 1.0$ is too small and the number of successful minimisations for 1.5 and 2.0 are practically the same. It may be of benefit to generate structures close to the final lattice energy minimised cells as this increases the chances of sampling smaller, narrow wells on the lattice energy surface, which may be missed with high TVP values.

The SAT-expand method offers the best balance between efficient use of the Sobol sequence and number of successful minimisations. SAT-expand structures are initially generated with a TVP of 1.0, ensuring they are close to a minimum, in addition only those structures which require a very large unit cell expansion are rejected. For both quinacridone $Z' = 1$ searches, the number of minimisations compares favourably with $TVP = 1.0$ and the number of trial structures required is close to $TVP = 2.5$. For CC1 only 10,090 trials were required to create 10,000 structures. For $Z' = 2$, despite resulting in the least amount of successful minimisations it was by far the most efficient search.

The convergence of the search (as in how long it takes for the search to produce no new structures and be functionally complete) can be measured in a number of ways. One approach is to monitor the unique, low energy minima that have been located as the search progresses. Figure ?? shows the results for all quinacridone searches and CC1 P1.

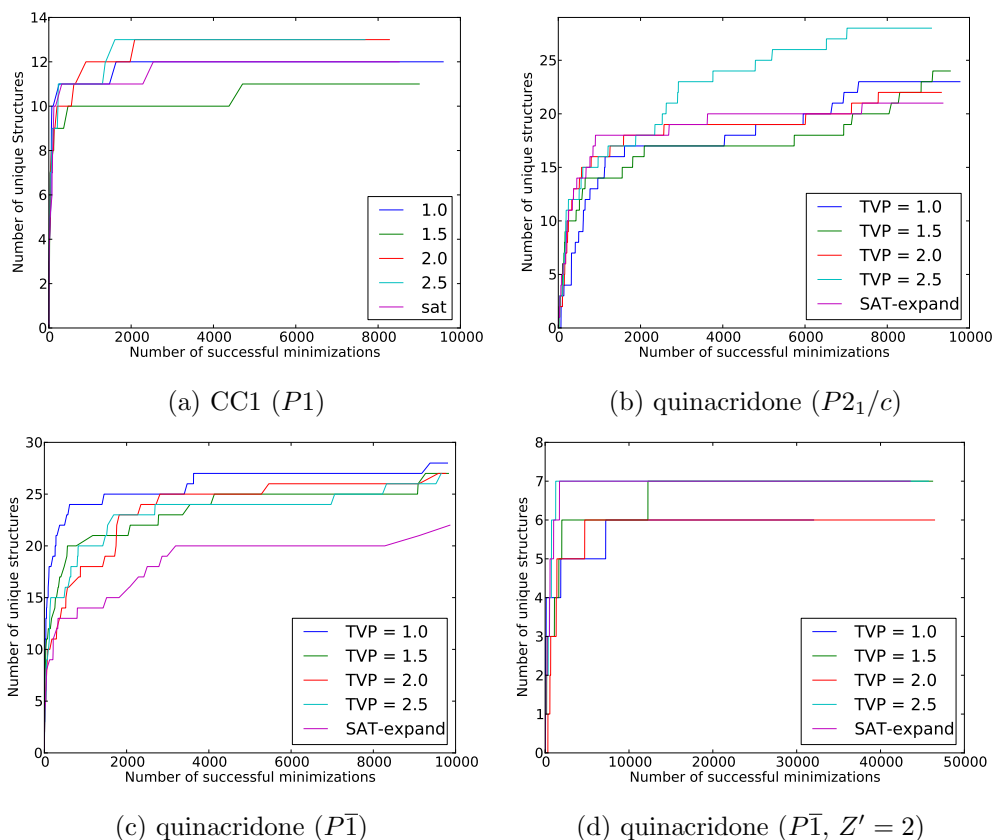


Figure 5.11: The number of unique crystal structures within 15 kJ/mol of the global minimum (60 kJ/mol for CC1) for all searches and all space groups, displayed as a function of the total number of successfully energy minimised structures.

It can be seen that for all searches the rate of discovery of new unique structures is high at the beginning of the search but tails off. Depending on the TVP value the number of unique structures can differ, this is likely due to clustering issues (especially prevalent in the flat quinacridone) but it is expected that the numbers of unique structures should be close. No general trend in the rate of convergence of each method is apparent, though the SAT-expand method compares favourably with all rejection methods apart from quinacridone ($P\bar{1}$).

As well as measuring the number of unique low energy structures it is also helpful to look at the energy evolution of the global minimum as the search progresses. Figure ?? displays the lowest energy structure encountered so far in the search and the average energy of the 10 lowest for the three $Z' = 1$ searches.

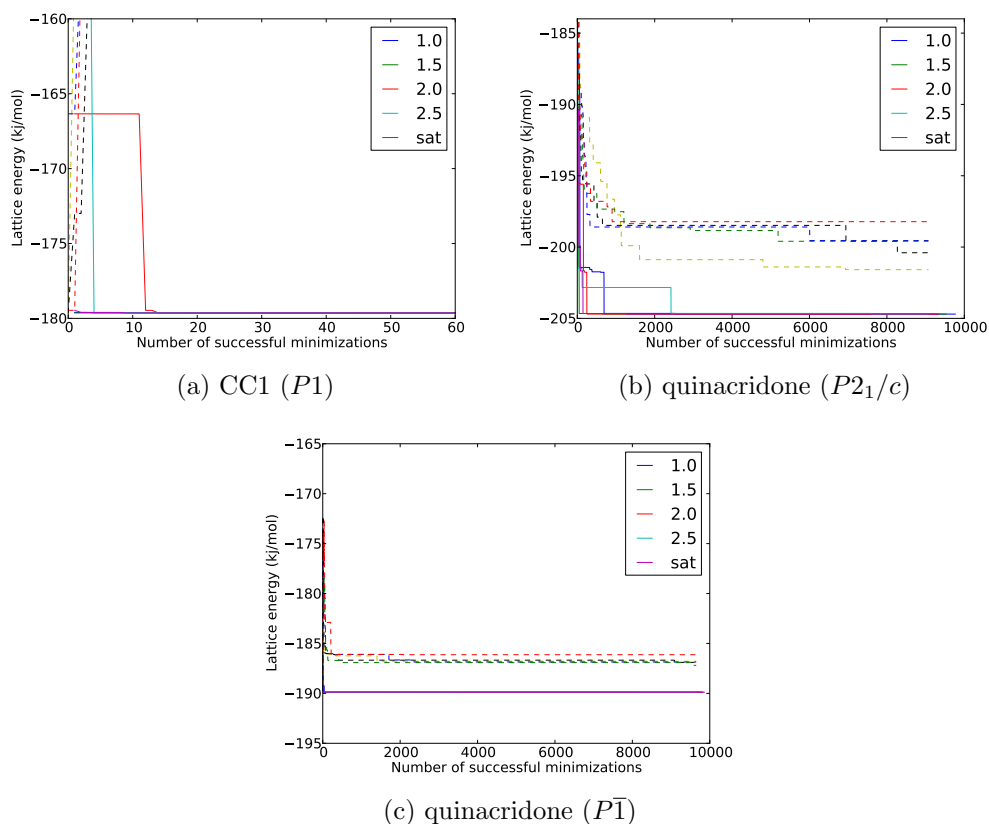


Figure 5.12: The average lattice energy of the ten lowest structures is shown as function of the number of minimised structures. The solid line indicates the energy of the single lowest energy structure, where the colour matches the legend. For CC1 the data had converged at 60 minimisations so is truncated.

It can be seen that all methods find the lowest energy structure rapidly, with the slowest being quinacridone $P2_1/c$ TVP = 2.5, which found the lowest energy of the structure a 1/5th of the way into the search. It is assumed that once the energy of the lowest structure remains stable throughout the search that this structure corresponds to the global minimum in that space group. For CC1 this global minimum is found almost immediately in the search and sampled many times as the average energy of the 10 lowest is almost the same as the single lowest energy structure. It seems that for larger TVPs the search converges slower, so the best results are obtained with a low TVP and rejection or using the SAT-expand method, which locates the minimum early for each search. While the global minimum is found quickly in each search allowing for a cheap (in terms of numbers of structures generated) search across many space groups, structure generation is often much cheaper than the corresponding lattice energy minimisation. Therefore, many 1000s of structures are generated to ensure the minimum is found.

Locating low energy structures is not the whole story when generating trial structures. The number of times each low energy structure is hit is also important as this is a measure of the uniformity of the search and how this translates to the sampling of the low energy

minima. Figure ?? shows the number of times each of the 10 lowest energy structures was hit during the course of the search for each method. For CC1 the frequency of locating each distinct minimum decreases as the lattice energy increases and for all methods well over half the trial structures minimise to the same two wells with 4000-5000 belonging to the global minimum. CC1 also exhibits the largest energy differences between structures as there are few ways to achieve low energy structures with only translational symmetry ($P1$ being the simplest space group). Between the five search methods, the results are broadly the same with the number of hits dropping slightly as TVP increases, due to fewer successful lattice energy minimisations overall, before rising again when SAT-expand is used. For quinacridone this trend is repeated with the global minimum being the most sampled structure (apart from the 3rd structure in $P2_1/c$). The lattice energy landscapes are more detailed than that of CC1 with structures closer in energy to the global minimum, that have also been sampled frequently and no clear relationship between the relative energy of the minimum and its frequency of being sampled.

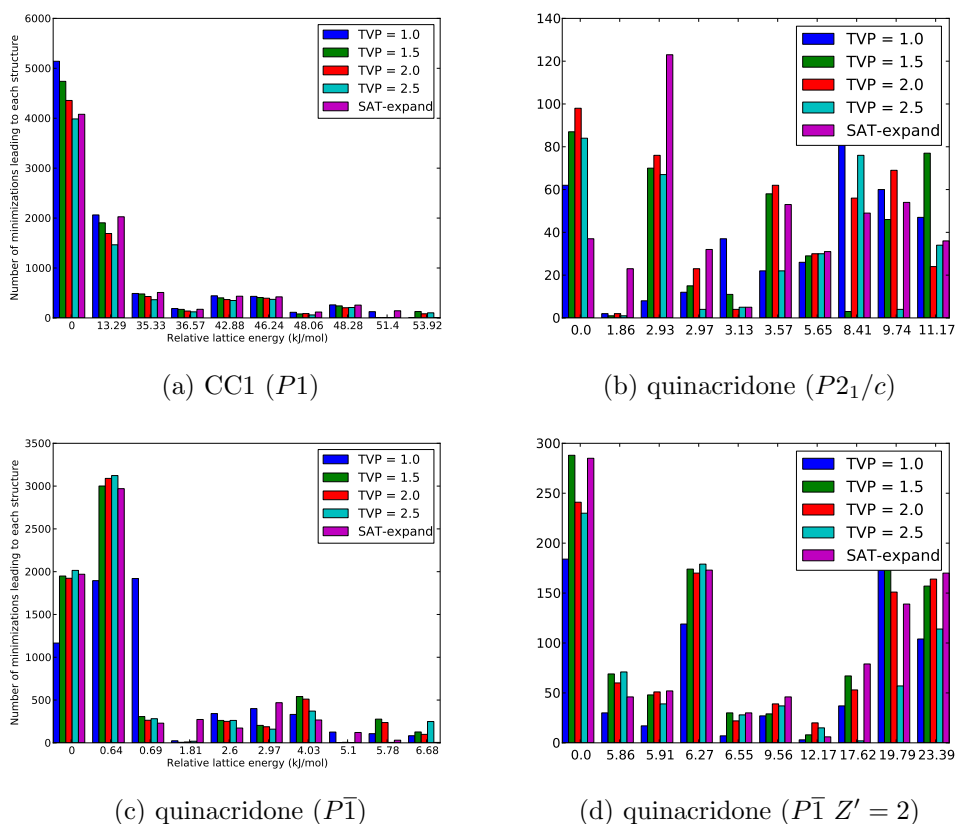


Figure 5.13: Bar charts showing the frequency with which each low energy structure is located. The lowest 10 are shown, with number of hits on the vertical axis and relative energy difference from the global minimum on the horizontal. Each of the five searches appear alongside each other.

Quinacridone $P\bar{1}$ shows the least lattice energy spread amongst low energy structures.

Upon visual inspection the packing motifs are similar and while the first two structures are easy to find, the rest are found far less frequently. It is structures that are found least frequently that are worrying. If the search was stopped early it is possible that these will be missed. For examples of this in our searches (structures 4 (1.81 kJ/mol) and 8-10 for quinacridone in $P\bar{1}$). There seems to be no solid relationship between TVP and frequency though SAT-expand samples the least frequently located structures better, as a range of initial volumes is generated during the search.

In addition to analysing the frequency of low energy structures, knowing where they were generated in the Sobol sequence can be useful. Figure 5.14 shows this for quinacridone $P\bar{1}$ (further examples can be in Figs 5.21, 5.22 and 5.23 at the end of this chapter). Each point on the plot represents a trial crystal structure and shows where in the Sobol sequence it was generated. Plotting the data this way shows the Sobol sequence is sampling the trial space uniformly as the points leading to each structure are evenly distributed through the sequence. The difference between TVP values is also apparent as a TVP of 2.5 hinders the sampling of some low energy structures (the 8th and 9th structures in particular), while SAT-expand samples structures more evenly.

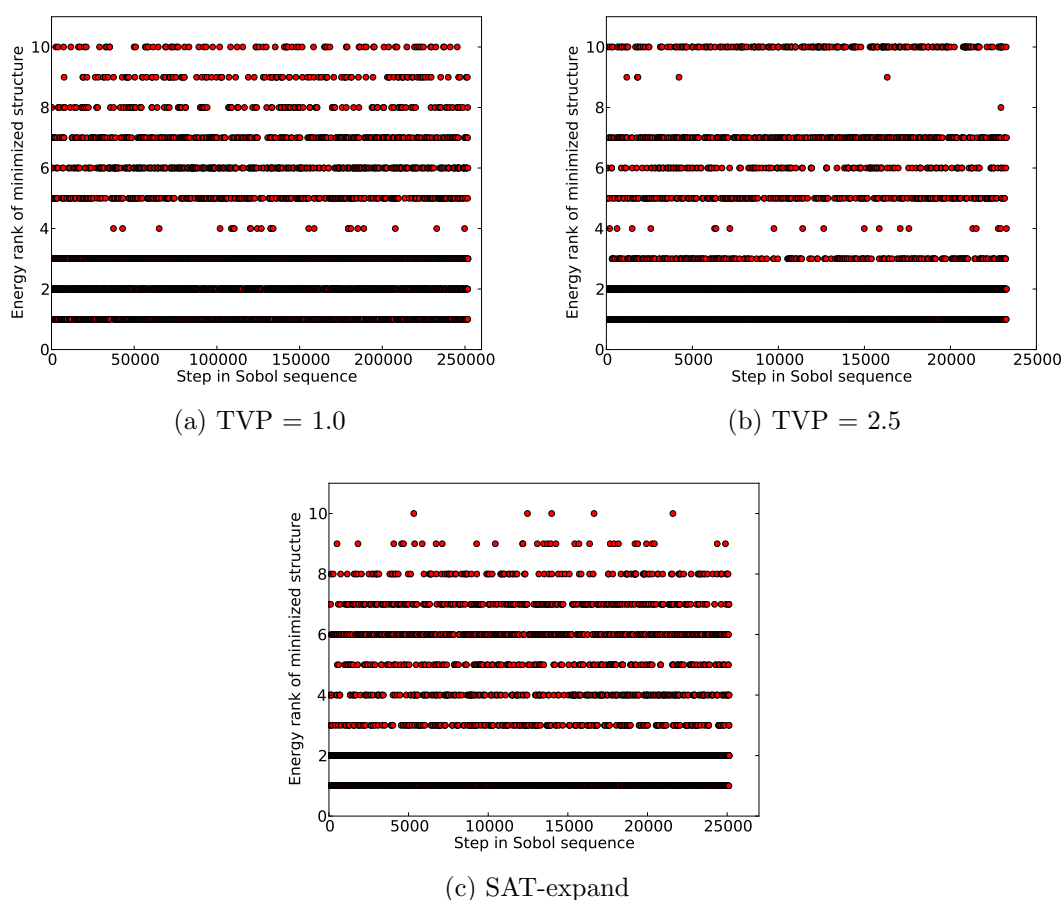


Figure 5.14: Hits to the 10 lowest ranked crystal structures of quinacridone $P\bar{1}$. Each point represents a minimisation from a trial structure showing where the trial structure was generated in the Sobol sequence.

From the analysis above it can be seen large target volumes led to less reliable sampling of structures. The use of SAT allows us to quickly rule out unphysical trial structures, and by expanding the cell to relieve clashes, we can keep a larger proportion of trial structures. This is important, as the Sobol sequence is designed to cover the configurational space in an efficient way, and helps us to rapidly consider a wide range of potential structures in the search. This SAT-expand approach has the best characteristics, overall of the differing search methods.

5.2.2.3 Full searches

After the testing described above, the SAT-expand method (showing the best sampling characteristics) with a max volume parameter of 2.5 was chosen to perform a full CSP for each of the molecules. For quinacridone, 16 space groups were searched ($P1$, $P2_1$, $C2$, $P2_12_12$, $P2_12_12_1$, $C222_1$, $P4_12_12$, $R3$, $P\bar{1}$, Cc , $P2_1/c$, $C2/c$, $Pna2_1$, $Pbcn$, $Pbca$ and $Pnma$) representing the most commonly observed for organic molecules, with 5000 structures generated in each with one molecule in the asymmetric unit. The plots of lattice energy versus density can be seen in Fig 5.15.

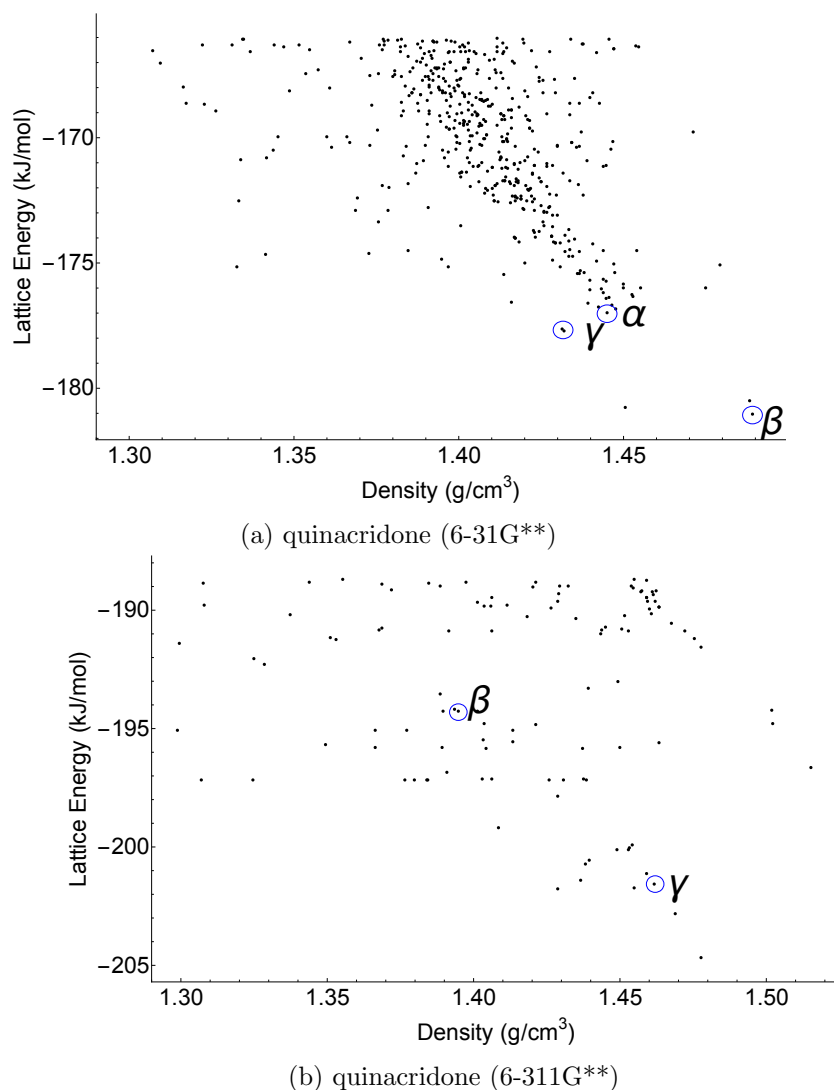


Figure 5.15: Lattice energy versus density plots for quinacridone. Each point is a unique minimum in the lattice energy surface upto 15 kJ/mol from the global minimum. Two searches were performed for quinacridone with the basis set used for the calculation of the multipoles in parentheses. For 6-311G** the α polymorph appeared too high in the set to be plotted with all other matches to experimental structures circled.

The central assumption of CSP is that the global minimum predicted by lattice energy is the most likely structure to appear. Free energy effects from the dynamics of molecules around their equilibrium positions can be important²³⁷, but in this study only lattice energy was considered as this is the largest contribution to the total energy of the crystal. While polymorphism indicates the physical process of crystallisation is not a simple march to the lowest energy structures we ignore possible solvent and temperature effects. The results for quinacridone were surprisingly sensitive to the basis set used to generate the electrostatic model. For B3LYP/6-311G** multipoles, the γ polymorph was the lowest ranked experimental structure located (5th in the search, 3.4 kJ/mol above the global minimum). The β and α_I polymorphs are 9.9 kJ/mol and 16.9 kJ/mol

away from the global minimum respectively. No full search was performed for $Z' = 2$ but the α_{II} polymorph was found in all the searches used for TVP testing. As in previous quinacridone polymorph studies²²⁶ it was found that the $Z' = 2$ α_{II} structure minimises to the $Z' = 1$ γ structure, suggesting both polymorphs correspond to the same minimum on the lattice energy surface. The energies seen are very high for experimentally known polymorphs. Other CSP studies performed on quinacridone in the past²²⁶ using a simple, isotropic force field, ranks the polymorphs in their correct order. Therefore there is no reason to believe these are truly high energy crystal forms. Table 5.2 shows that the predicted cell parameters are not in great agreement with those determined from X-ray diffraction. To investigate whether this was due to the nature of the electrostatic model, the same search was repeated with atomic multipoles generated with a smaller basis set (6-31G**). The ranking of the observed structures changes dramatically: β is now the global minimum, with γ and α_I 3.4 kJ/mol and 5.68 kJ/mol above the global minimum. As before, the structure of $Z' = 2$ α_{II} was located when the smaller basis set was used with the same relaxation to the γ structure. On comparison to experimental structures and cell parameters, the 6-311G** model performs much worse than 6-31G**, especially in the case of the α polymorphs. The RMSD values for the α polymorphs are both greater than 1.0 suggesting poor geometric matches.

For all the molecules used in the testing of the structure generator, the same methodology was used in generating the force field, while (as Table 5.2 shows) it was successful in the cases of artemisinin and CC1 it was not as successful for quinacridone. The anisotropic electrostatic model we include is highly sensitive to the quality of the underlying density functional theory (DFT) calculations. While the electrostatics are *ab initio*, all other intermolecular interactions are calculated using empirically fitted terms in the force field, which are not fitted to our systems of interest in particular. This by itself should not be enough to explain the changes in lattice energy or polymorph ranking seen in this case. It is known that the electronic structure of a π system will alter as multiple rings are joined together, Grimme²³⁸ suggested that stronger dispersion effects arise in molecules with more than three fused rings. This includes quinacridone, which as an organic semiconductor has strong directional effects in its electronic structure, but few molecules included in the parametrisation of our force field. The use of quinacridone has highlighted that for systems such as quinacridone the transferability of standard force fields may be limited.

For CC1, both experimental polymorphs were found in the low energy region and are good geometrical matches to X-ray data. The results of the CSP match with an earlier study using simulated annealing²³¹. The search for artemisinin was equally successful with the 2nd ranked structure corresponding to the known $Z' = 1$ polymorph. In addition, the second $Z' = 4$ polymorph was found as the global minimum through the generation of 50,000 structures in *P1*. The difficulties of searching higher Z' values make this an excellent result, though it was only hit three times.

Crystal structure		Cell lengths			Cell angles			RMSD ₃₀
		<i>a</i>	<i>b</i>	<i>c</i>	α	β	γ	
Artemisinin (<i>P</i> ₂₁ 2 ₁ 2 ₁)	expt.	24.066	9.439	6.354	90.00	90.00	90.00	-
	pred.	24.456	9.399	6.386	90.00	90.00	90.00	0.131
Artemisinin (<i>P</i> 1), <i>Z'</i> =4	expt.	9.881	9.891	15.343	93.28	90.92	102.99	-
	pred.	9.892	10.020	15.164	90.81	93.64	102.32	0.247
CC1 (<i>R</i> 3, β')	expt.	21.015	21.015	10.491	90.00	90.00	120.00	-
	pred.	21.623	21.602	10.851	90.02	90.02	119.98	0.603
CC1 (<i>P</i> ₂₁ / <i>c</i> , α')	expt.	12.810	10.910	36.810	90.00	97.49	90.00	-
	pred.	13.425	11.156	37.761	90.00	94.45	90.00	0.812
Quinacridone (<i>P</i> ₂₁ / <i>c</i> , γ)	expt.	13.697	3.881	13.402	90.00	100.44	90.00	-
	pred. (6-31G**)	12.847	4.251	13.370	90.00	97.08	90.00	0.288
	pred. (6-311G**)	13.397	4.115	13.002	90.00	98.21	90.00	0.439
Quinacridone (<i>P</i> ₂₁ / <i>c</i> , β)	expt.	5.692	3.975	30.020	90.00	96.76	90.00	-
	pred. (6-31G**)	5.746	4.110	29.565	90.00	93.90	90.00	0.369
	pred. (6-311G**)	8.972	5.296	34.750	90.00	123.27	90.00	0.492
Quinacridone (<i>P</i> $\bar{1}$, α^I)	expt.	3.802	6.612	14.485	100.68	94.40	102.11	-
	pred. (6-31G**)	4.331	6.203	13.632	97.49	97.07	98.00	0.451
	pred. (6-311G**)	4.620	6.372	12.530	95.25	97.82	103.62	1.245
Quinacridone (<i>P</i> $\bar{1}$, α^{II})	expt.	14.934	3.622	12.935	91.39	107.13	92.84	-
	pred. (6-31G**)	13.684	4.369	13.239	90.00	115.39	90.00	0.219
	pred. (6-311G**)	13.397	4.115	13.002	90.00	98.21	90.00	1.019

Table 5.2: Matches from the full CSP to experimentally determined structures of the observed polymorphs. RMSD₃₀ is the deviation in atomic positions of a cluster of 30 molecules taken from predicted and experimental structures, not including hydrogen atoms. Non quinacridone results are also included. CC1 (*R*3) was generated in the *P*1 spacegroup, which reduces to *R*3 on account of intramolecular symmetry, hence the cell angles differ at the second decimal place. The experimental structures of CC1 also contained residual solvent, which was removed for purposes of comparison. All structures were converted to their reduced unit cell for comparison. Å and degrees are used throughout.

5.3 CSP of a blind test molecule

Chapter 2 introduced different approaches to CSP and progress in the field has been measured through six blind tests held by the Cambridge Crystallographic Data Centre. For each test a number of molecules are sent to active researchers in the field to attempt a "blind" prediction. No information is given apart from molecular connectivity and some crystallisation conditions and X-Ray data of the actual structures are held until completion of the test. A review of results from previous tests can be found in Chapter 2. On the 12th of September 2014 the 6th blind test was announced and the molecules sent to researchers. The molecular categories have changed over time with the advancement of CSP methods for the 6th test they were as follows:

- I) a rigid molecule, with functional groups restricted to C,H,N,O, S, P, B and halogens, with one molecule in the asymmetric unit and 30 atoms total

- II) a partially flexible molecule with two to four internal degrees of freedom, one molecule in the asymmetric unit and 30-40 atoms total
- III) a partially flexible molecule with two to four internal degrees of freedom as a salt, with two charged in the asymmetric unit and 30-40 atoms total, in any space group
- IV) multiple partially flexible (one to two internal degrees of freedom) independent molecules as a co-crystal or solvate, in any space group upto 40 atoms
- V) a molecule with four to eight internal degrees of freedom, no more than two molecules in the asymmetric unit, any space group and 50-60 atoms.

The molecules chosen to satisfy these requirements are shown in Fig 5.16. This section describes the prediction of the first molecule²³⁹, XXII, which is in category I.

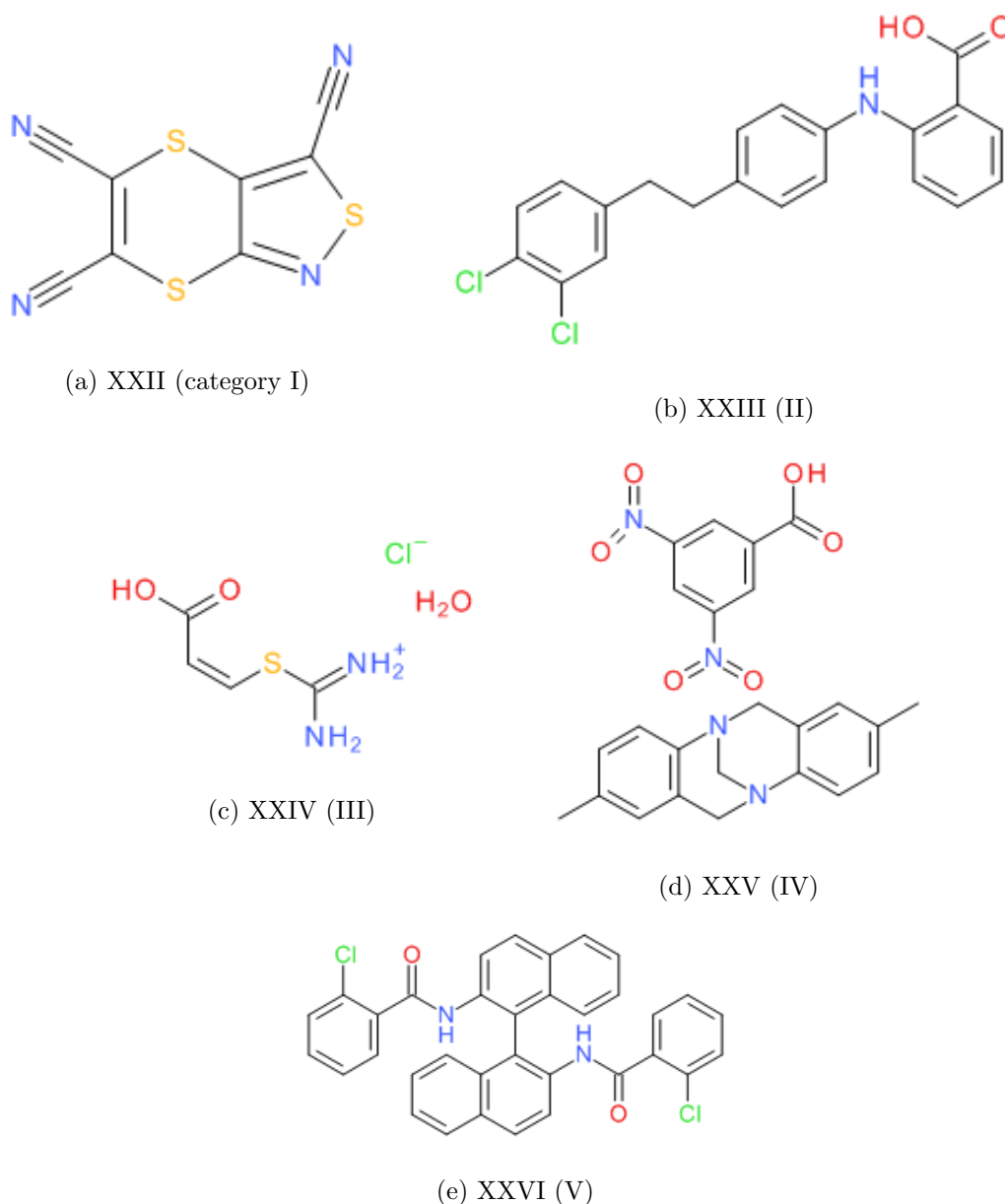


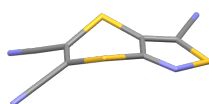
Figure 5.16: The five molecules chosen for the 6th blind test of CSP.

The information given for XXII at the beginning of the blind test for the crystallisation conditions was "Crystallised from an acetone/water mixture; chiral-like character due to the potential flexibility of the six-membered ring, but no chiral precursors used in synthesis". For each candidate molecule groups are allowed to submit a list of 100 candidate structures ranked in some way, in addition to a second list of 100 structures re-ranked in some way.

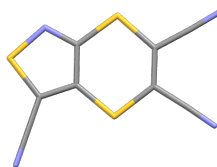
The possible flexibility in the ring system of XXII necessitated a conformational search of possible conformers to perform CSP with. This search was performed using a low-mode conformational search (LMCS)^{240;241} as implemented in MacroModel²⁴². This method ensures conformer sets are as complete and unbiased as possible. The starting

molecular geometry (a 3D geometry drawn in the MacroModel visualiser) is optimised, perturbed along a random combination of its calculated normal modes and re-optimised, with new conformers clustered on the fly (if they have an RMSD within 0.02 Å). Due to large numbers of optimisations required a force field was used for energetic assessment. OPLS2005^{243;244;245} was used due to its strong performance in a conformational study²⁴⁶. After generation of conformers all within 50 kJ/mol of the global minimum were re-optimised using DFT-D with B3LYP/6-311G** and Grimme's D3 dispersion correction scheme and Becke and Johnson damping^{247;248}. After this optimisation the conformers were clustered again to remove duplicates.

For XXII a buckled conformer was found (Fig 15) as the global minimum (C1) with a symmetry-related twin, a planar conformation was also located as a first order saddle point between the two minima (C2). After the DFT-D re-optimisation the saddle point was only 5.7 kJ/mol above the global minimum so it was chosen for CSP in addition to the global minimum.



(a) XXII (C1)



(b) XXII (C2)

Figure 5.17: The two XXII conformations chosen for CSP.

Rigid CSPs were performed for both conformers. Structure generation proceeded using GLEE²²³ and the structure generator described above, with the difference of aiming for a number of valid structures after lattice minimisation rather than just after structure generation. C1 is chiral so 94 space groups were sampled, the 25 most common had 5000 valid structures in each, with 2000 in each of the remaining 69. If any of the additional

space groups gave a structure within 25 kJ/mol the number of structures was increased to 5000 (46 of the 69 required this). No change in the lowest energy structures was seen and only five crystal structures from the additional space groups were within 15 kJ/mol of the global minimum. C2 is non-chiral so 87 space groups were sampled the same way as C1. 34 uncommon space groups required additional sampling and four structures from these made it within the 15 kJ/mol of the C2 global minimum. The global minimum of the combined searches was -120.9 kJ/mol

Lattice energy minimisations were performed in the same way as quinacridone above with a 15 Å cutoff on van der Waals and higher order electrostatics. In addition the revised W99 force field was supplemented with parameters for sulphur²⁴⁹ using standard combining rules. After the rigid body CSP was completed and results clustered (using the in-house method described above in addition to COMPACT⁵⁴) structures within 15 kJ/mol of the global minimum were re-optimised taking into account intramolecular flexibility. Crystal Optimiser²⁵⁰ uses an atom-atom force field for the description of intermolecular interactions with a quantum mechanical description of the intramolecular energy. Specified torsions and angles are optimised in response to intermolecular packing forces within the crystal. After this minimisation and re-clustering 753 unique crystal structures populated the XXII lattice energy landscape. The top 100 structures are plotted with relative lattice energy and density in Fig 5.18.

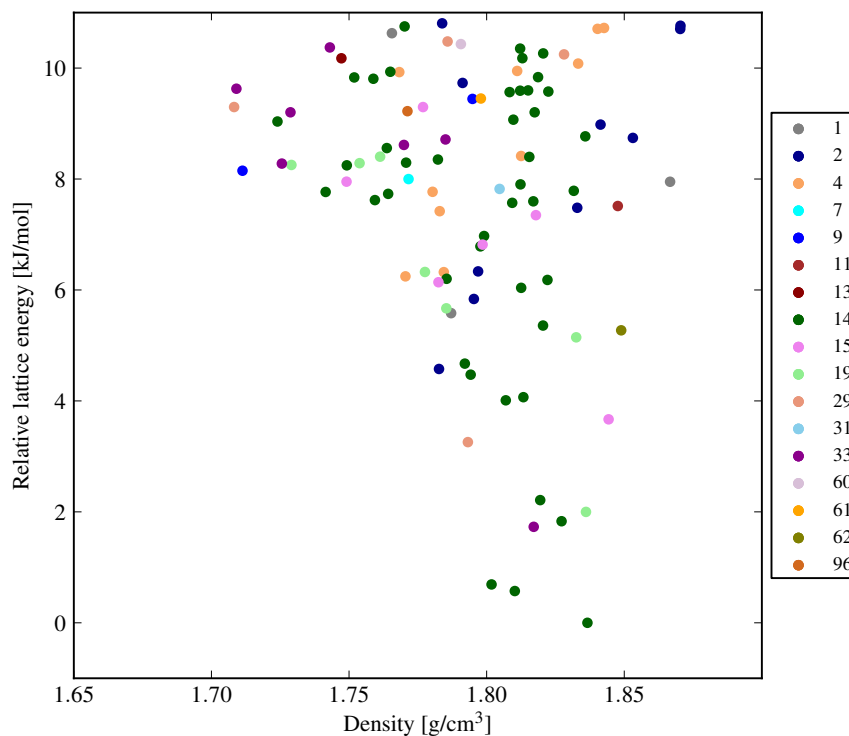


Figure 5.18: Lattice energy landscape of XXII, plotting relative lattice energy versus density. The lowest 100 structures are shown. Labels refer to the space group number.

This created the first submission list but another re-ranked list was also submitted. As briefly mentioned above, the lattice energy is not the total energy of the crystal; free energy effects can also be important²³⁷. The thermodynamic stability of polymorphs is governed by their free energy differences which included contributions from zero-point and thermal motion to the enthalpy and entropy of the lattice. The development and performance of the free energy calculations were attempted by another member of our group Mr Jonas Nyman but the results are presented here (Fig 5.19). The Helmholtz free energy (the sum of the static lattice energy and lattice vibrational contributions) is calculated from phonon frequencies using the rigid molecule, harmonic approximation, with no intramolecular vibrations or phonon-phonon couplings included at $T = 300\text{K}$.

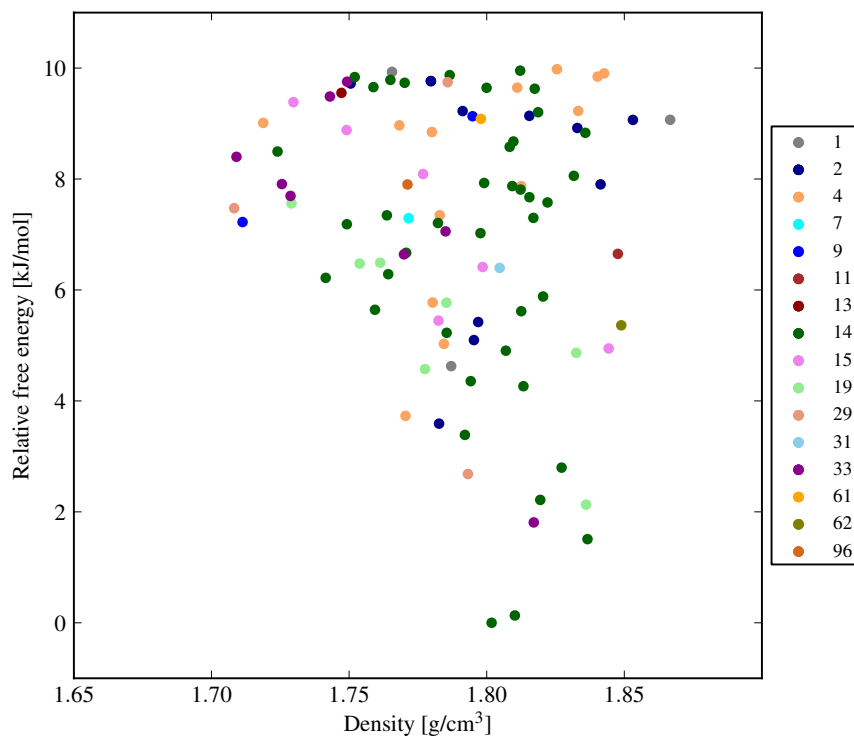


Figure 5.19: Lattice energy landscape of XXII. Plotting relative free energy ($T=300\text{K}$) versus density. The lowest 100 structures are shown and labels refer to the space group number.

The main difference between the two sets of structures was the re-ranking of the 3rd lowest energy structure to be the new global minimum. Once results were submitted and the blind test paper prepared 12 out of 21 total submissions found the experimentally known structure. This was the 3rd ranked structure in our lattice energy list, and the global minimum in the free energy list (Fig 5.20 shows the overlay of our predicted structure with the experimental structure). Table 5.3 shows the cell parameters of the match to known experimental structure. This re-ranking shows the importance of taking into account all energetic contributions in the crystal.

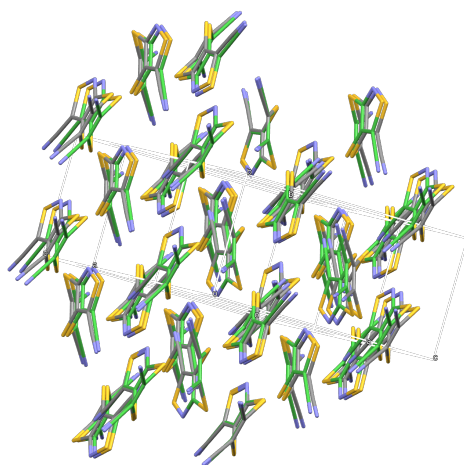


Figure 5.20: Overlay of the global minima (green) from our free energy ranked list of submitted structures with the experimental structure of XXII ($\text{RMSD}_{30} = 0.292 \text{ \AA}$).

	Cell lengths			Cell angles			RMSD ₃₀
Crystal structure	a	b	c	α	β	γ	
XXII predicted	11.700	6.651	12.191	90.00	105.20	90.00	0.292
XXII experimental	11.947	6.696	12.598	90.00	108.60	90.00	

Table 5.3: Matches from the full CSP to experimentally determined structures of the observed polymorphs. RMSD_{30} is the deviation in atomic positions of a cluster of 30 molecules taken from predicted and experimental structures, not including hydrogen atoms. \AA and degrees are used throughout.

5.4 Conclusions

In this chapter two pieces of work outside the main theme of this thesis are presented. The first deals with the development and testing of an in-house structure generator for CSP. A quasi-random sequence is used to sample molecular positions, orientations and unit cell parameters. The SAT method is used to check for molecular overlaps and can be used to expand the cell rather than just rejecting structures out of hand. Initial testing involved the discovery of flat cells which required a new angle sampling method to be used in addition to the choosing of cell lengths. Once this was completed the structure generator was applied to three molecules representing a range of molecular interactions and shapes. The three molecules (quinacridone, artemisinin and CC1) were searched multiple times using different TVP values and the SAT-expand method. Results showed that the acceptance of trial structures increased with a larger target volume but that

more successful lattice energy minimisations occurred with a smaller target volume. The SAT-expand method offers a good middle ground between these two considerations and was chosen to perform full CSPs on the molecules of interest. For artemisinin and CC1 the structure generator was successful, locating known experimental structures (including a $Z' = 4$ polymorph in the case of artemisinin). For quinacridone the ranking of the polymorphs was difficult, showing a surprising sensitivity to the quality of the DFT calculations used in the generation of the electrostatic model. The problems with ranking the polymorphs are due to this fact rather than the structure generator.

The second piece of work is the CSP of a blind test molecule from the 6th blind test molecule XXII (a small, rigid molecule, containing C,H,N,S atoms). A conformer search was performed locating two conformers of interest for CSP. A rigid search was attempted first sampling many space groups for both conformers. After clustering and combining the lists all structures within 15 kJ/mol of the global minimum were re-minimised allowing intramolecular flexibility in response to intermolecular interactions. This created the first submission list of 100 structures, the second being created through the calculation of the free energy of each structure, which was then used to re-rank. The experimental structure was located in both lists, but was re-ranked from 3rd to 1st upon the inclusion of free energy, showing the importance of including all energetic terms when performing CSP.

Our results show that CSP works well for a range of relatively rigid molecules. Artemisinin, CC1 and the blind test molecule are predicted very well, showing that search and energy ranking methods are working. Quinacridone is a warning about sensitivity to the model, but results improve with a smaller basis set. The next chapter will detail the CSP of a number of azapentacene derivatives and their evaluation as potential organic semiconductors.

5.5 Additional figures

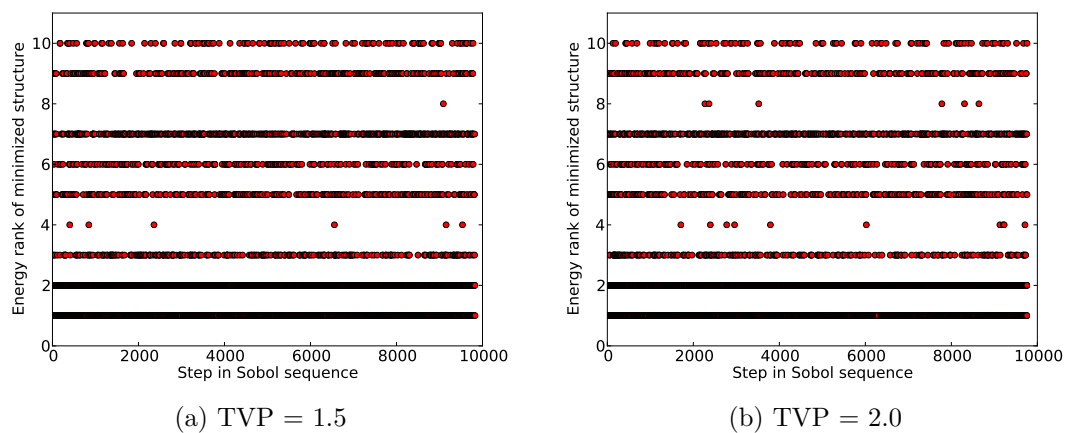


Figure 5.21: Hits to the 10 lowest ranked crystal structures of quinacridone $P\bar{1}$. Each point represents a minimisation from a trial structure showing where the trial structure was generated in the Sobol sequence.

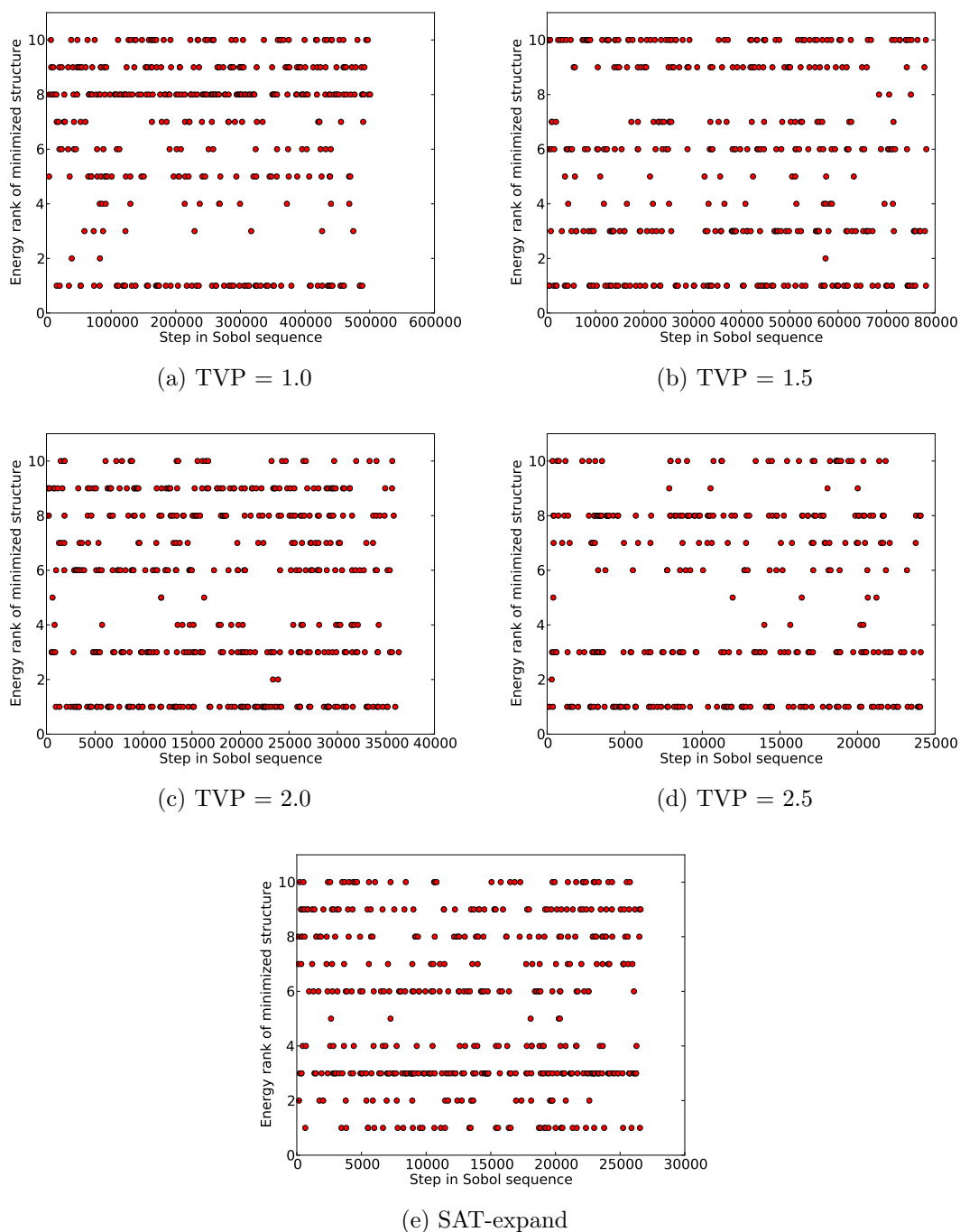


Figure 5.22: Hits to the 10 lowest ranked crystal structures of quinacridone $P2_1/c$. Each point represents a minimisation from a trial structure showing where the trial structure was generated in the Sobol sequence.

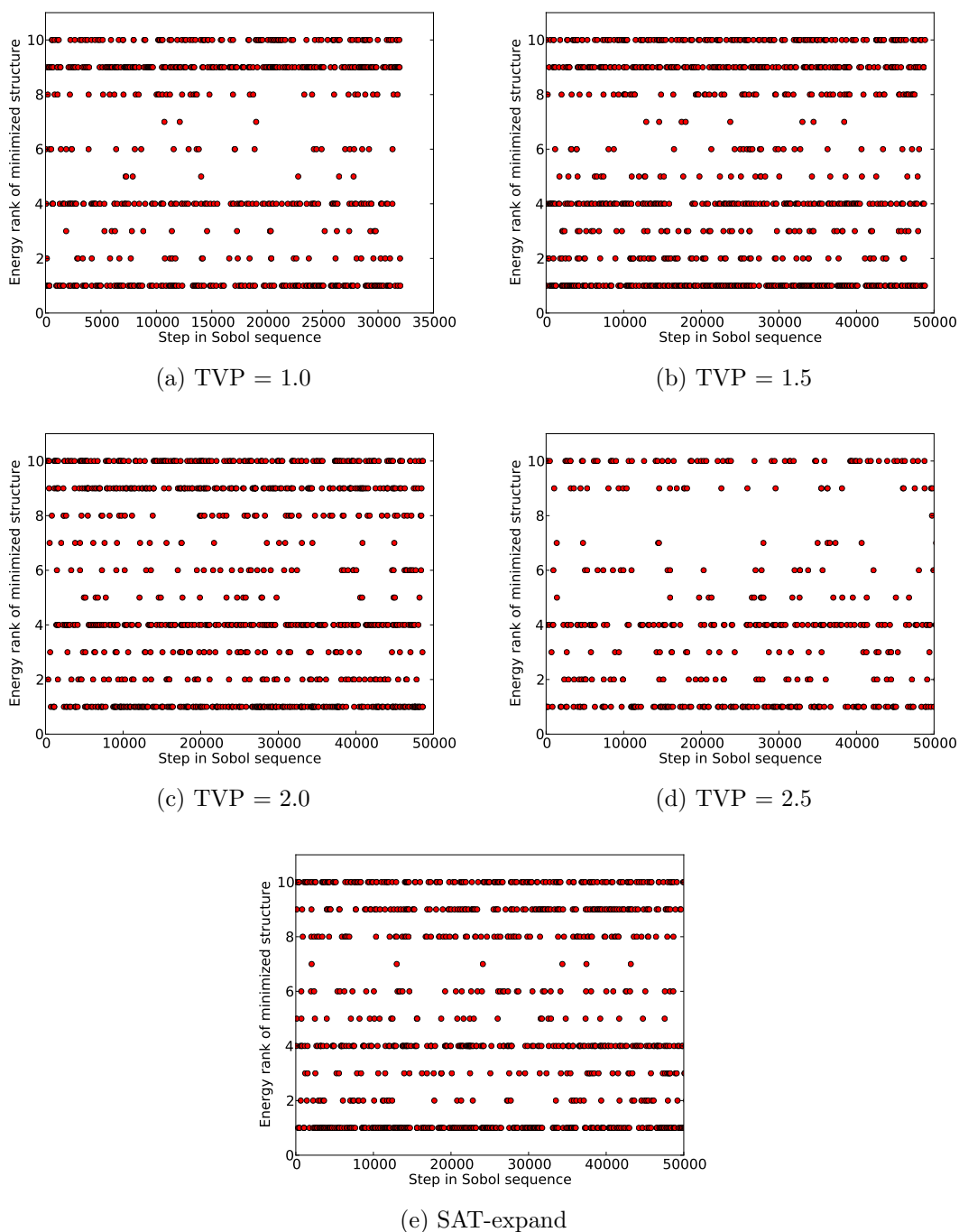


Figure 5.23: Hits to the 10 lowest ranked crystal structures of quinacridone $Z' = 2 P\bar{1}$. Each point represents a minimisation from a trial structure showing where the trial structure was generated in the Sobol sequence.

Chapter 6

Azapentacene Crystal Structure Prediction

6.1 Introduction

Chapters 2 and 3 introduced CSP and the parameters needed for good charge transport in molecular crystals. CSP can be used as a tool for crystal engineering by predicting possible polymorphs of a given molecule. When combined with the prediction of properties CSP can be used to computationally design molecules with desired properties. No experimental data is needed, allowing for the screening of potential molecules before they are synthesised. Organic semiconductors are an attractive field to apply this methodology to; charge mobility depends on both intra and intermolecular parameters. The mobility shows a particularly fine sensitivity to small changes in the crystal structure and the composition of the molecule. This chapter presents the CSP of novel organic semiconductors, the analysis of how substitution patterns affect crystal packing and the calculation of the charge mobility of low energy structures.

Pentacene is an attractive starting point for the design of novel organic semiconductors with promising electronic properties^{251;130;252} attributed to its small λ . Pentacene crystallises in a herringbone arrangement[?], which is characterised by tilted edge-to-face C-H $\cdots \pi$ interactions. However, the electronic coupling between molecules is known to vary strongly with the interplanar angle and is maximised in a co-facial molecular arrangement²⁵³. Thus, the herringbone packing seen in many PAHs is not optimal for charge transport and there have been efforts to modify molecular packing by introducing substituents²⁵⁴. In addition, pentacene is a p-type semiconductor (it transports holes rather than electrons), which can also be modified by substitution, with electronegative atoms offering a way to change the transport character and offer potential hydrogen bonding networks to modify the packing.

Azaacenes (and azapentacenes in particular) offer a way to favourably modify electronic properties and crystal packing, in particular, the possibility to form $N \cdots H-C$ hydrogen bonded networks, which could promote sheet-like packing in the crystal phase. Interest in azaacenes has increased over recent years^{139;140;255;256} due to this potential control of crystal packing and intriguing theoretical results.

Chen and Chao¹⁴¹ investigated a series of theoretical azapentacenes as they tried to lower the value of the internal reorganisation energy. The authors showed that too much nitrogen substitution (10N) increased λ_- due to the effects of the electronegative perturbation on the LUMO, increasing its non-bonding character. This leads to stronger orbital interactions between neighbouring atoms, resulting in a larger geometry change when an electron hops on. However, 10N substitution did result in a large increase in electron affinity (a property needed for a good n-type material¹³¹) so 5N was also investigated and showed good (0.149-0.167 eV) and tunable (depending on N position) λ values. Winkler and Houk¹⁴² also theoretically examined a series of azapentacenes, calculating reorganisation energies for varied amounts of nitrogen substitution and locations, including 5N. Their λ values agree quite well with those of Chen and Chao. Some theoretical work has been done showing promising results with N substitution into already 6,13-substituted pentacene derivatives though no thorough examination of the crystal packing was performed.¹⁴³ Winkler and Houk state that ‘A most interesting question is how substitution of CH by N modifies the solid-state structures (and hence transfer integrals) of azaoligoacenes’. This is the central focus of this chapter.

6.2 Molecules studied

6.2.1 Validation molecules

For the CSP presented in this chapter two molecules with known crystal structures were chosen as validation for our method. Pentacene (Fig 7.1b, CSD reference code PENCEN) is a polycyclic aromatic hydrocarbon (PAH) consisting of five linearly fused benzene rings. Pentacene shows promising electronic properties and has been the subject of intense research.^{251;130;252} A search of Web of Knowledge using the term ”pentacene semiconductor” returns over 7000 hits showing its popularity as a research target.

Pentacene has three experimentally known polymorphs. The first (known as the bulk polymorph, PI) was reported in 1961²⁵⁷ and revised in 1962²⁵⁸. Polymorph II (single crystal polymorph, PII) was discovered in 1999²⁵⁹ and the third (thin film polymorph, PIII) and was characterised in 2007²⁶⁰. All three polymorphs crystallise in spacegroup $P\bar{1}$ and unit cell dimensions differ only slightly between them. The distinguishing characteristic between polymorphs is their d_{001} spacing, that is, the distance between repeating layers²⁶¹ (Table ?? shows the lattice parameters of the three polymorphs).

Pentacene as the archetypal organic semiconductor is a good test for our method on predicting structures of fused ring systems. As we are also interested in how the effect of nitrogen substitution affects the crystal packing, the other validation molecule chosen was tetraazatetracene (Fig 6.1b). Tetraazatetracene (CSD reference code YEBMEZ, hereafter referred to as TT) is a n-channel semiconductor and was characterised in 2012²⁶². TT is an azaderivative of tetracene that crystallises in $P2_1/c$ from CH_3CN . So far, no additional polymorphs have been discovered.

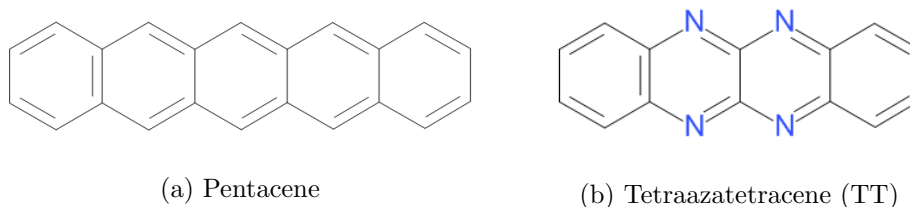
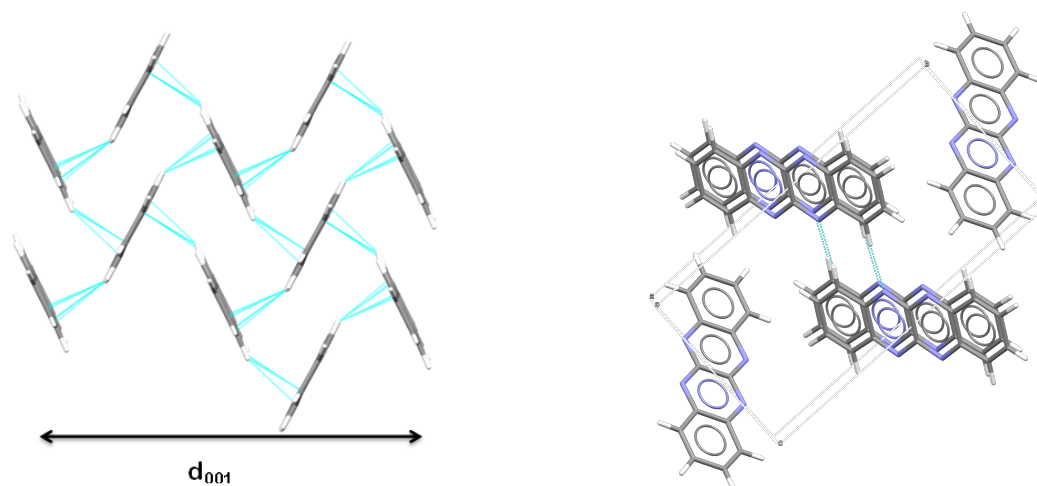


Figure 6.1: The two validation molecules chosen for our study.

Crystal Structure	a	b	c	α	β	γ	$d_{001}(\text{\AA})$
PENCEN (bulk) ²⁵⁸	7.90	6.06	16.01	101.9	112.6	85.8	14.4
PENCEN01 (single crystal) ²⁵⁹	6.28	7.71	14.44	76.752	88.01	84.52	14.1
PENCEN10 (thin film) ²⁶⁰	5.96	7.60	15.61	81.25	86.56	89.80	15.4
YEBMEZ ²⁶² (TT)	4.71	14.91	7.65	90.00	94.70	90.00	-

Table 6.1: Lattice parameters (vectors in \AA , angles in $^\circ$ and d spacing of the three pentacene polymorphs in addition to the experimentally observed TT structure.

Pentacene packs in a herringbone motif across all the experimentally known structures (Fig 6.2a. Directed by edge to face ($\text{C}\cdots\text{H}$ interactions) and face to face π -stacking ($\text{C}\cdots\text{C}$ interactions), this packing is not considered the best for charge transport, but due to pentacene's excellent (low) reorganisation energy mobility is still high. TT packs in a γ motif (Fig 6.2b) directed by $\text{N}\cdots\text{H}$ hydrogen bonds. Observed charge mobilities for TT were lower than that of the unsubstituted parent molecule (tetracene), with $8.9 \times 10^{-5} \text{ cm}^2/\text{Vs}$ (versus $0.15 \text{ cm}^2/\text{Vs}$ ²⁶³) a fact the authors of the original paper attribute to device processing issues.



(a) Herringbone packing in pentacene, directed by the highlighted edge to face interactions
 (b) A γ packing in the observed structure of TT, directed by the highlighted NH hydrogen bonds.

Figure 6.2: Crystal packing of the two validation molecules.

6.2.2 Hypothetical molecules

Four molecules from reference 142 were chosen to investigate the effects of differing amounts and arrangements of nitrogen substitution, with alternating (A) nitrogen substitution and all nitrogens on one side of the molecule (B). Two further molecules were devised with unsymmetrical N-substitution patterns (C) with non-complementary long edges of the molecule, to present a more complete picture of the intermolecular interactions in azaacenes. We will refer to these molecules as nA/B/C where n refers the number of nitrogen atoms in the molecule. The structures of the molecules are shown in Fig 6.3. 5A and 5B were chosen as intermediates between the validation molecules and 7A/B to try and elucidate how many interactions it takes to disrupt herringbone packing in favour of cofacial sheets. 7A/B were explicitly designed to have an electron affinity above 3.0 eV¹⁴² and have a slight increase in reorganisation energy when compared to pentacene (0.131 to 0.18-0.2 eV). However the potential for regular cofacial arrangements in the crystal should still allow for high mobilities. While $\text{CH} \cdots \text{N}$ hydrogen bonds are generally weak compared to other hydrogen bonding motifs²⁶⁴, the sheer number of potential interactions in this series of molecules should lead to cofacial motifs. While larger theoretical acenes should have better charge transport properties (the HOMO LUMO gap decreases with additional rings) they are difficult to synthesise and very unstable. Therefore pentacene based derivatives are more realistic for device applications.

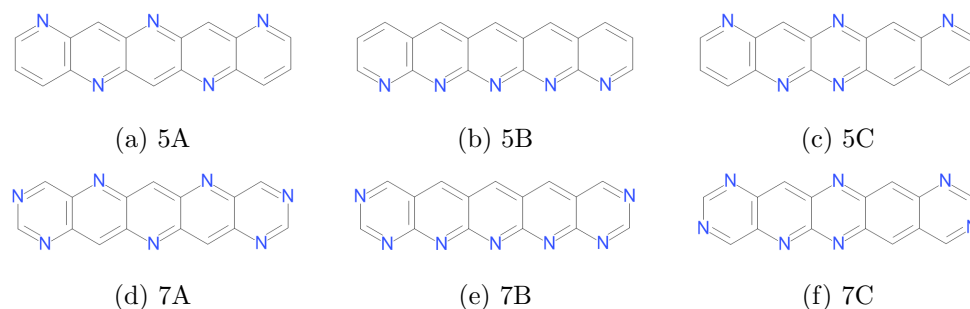


Figure 6.3: The hypothetical molecules chosen for this study (hydrogen omitted).

6.3 Methods

6.3.1 CSP

The CSP methodology used in this study is broadly the same as that in Chapter 5. The geometry of each molecule was kept rigid throughout the crystal structure calculations, at the optimised gas phase structure from a B3LYP⁷²/6-311G** calculation using Gaussian 09⁷¹. Trial crystal structures were then generated in a wide range of space groups, considering 1 and 2 molecules in the asymmetric unit. Searches were performed using the Global Lattice Energy Explorer (GLEE) software, which is described in Chapter 5 and a recent paper²²³. 4000 lattice energy minimised crystal structures were generated in each of the 23 most commonly adopted space groups for organic molecules³⁷ ($P2_1/c$, $P4_12_12$, $P2_12_12_1$, $P2_12_12$, $P\bar{1}$, Pc , $P2_1$, $P3_1$, $Pbca$, $P4_1$, $C2/c$, $Fdd2$, $Pna2_1$, $Pccn$, Cc , $P2/c$, $C2$, $P6_1$, $Pca2_1$, $I4_1/a$, $P1$, $R\bar{3}$, $Pbcn$), all with one molecule in the asymmetric unit ($Z'=1$). For $Z'=2$, 10000 structures were generated in each of 12 space groups ($P2_1/c$, $C2/c$, $P2_12_12_1$, $P1$, $Pca2_1$, $P\bar{1}$, $Pna2_1$, $P2_1$, $C2$, $Pbca$, Pc , Cc), leading to a total of 212,000 lattice energy minimised structures for each molecule.

All lattice energy minimisations were performed using DMACRYS⁶⁶, using the W99⁶³? model potential for all intermolecular atom-atom interactions. Electrostatic interactions were described using atomic multipoles derived from a distributed multipole analysis of the calculated molecular electron density. Ewald summation was used for charge-charge, charge-dipole and dipole-dipole interactions, while all higher order electrostatics and repulsion-dispersion interactions were summed to a 25 Å cutoff. Lattice energy minimisation was initially performed within the space group of the generated structure. In cases where this led to a saddle point, lattice energy minimisation was continued after removing the space group symmetry operators that allowed minimisation from the saddle point. This process led to some structures of higher Z' in the final structure sets.

Clustering was performed to identify and remove duplicate crystal structures. An initial screen was performed using the clustering method described in Chapter 5 (and in more

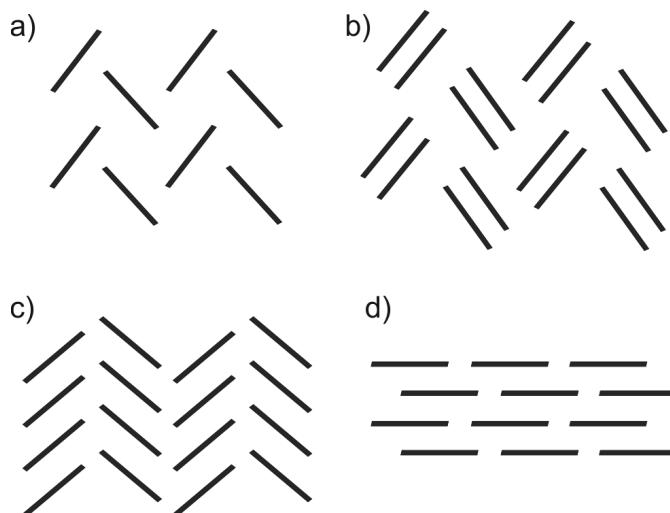


Figure 6.4: The four packing types seen in crystal structures of polycyclic aromatic hydrocarbons: a) herringbone; b) sandwich herringbone; c) γ and d) sheet (β).

detail in reference [223](#)) within individual space groups. An overall clustering across all space groups was then performed using COMPACK⁵⁴.

6.3.2 Classification of Predicted Crystal Structures

As discussed in Chapter 2, it is expected that polyaromatic hydrocarbons will crystallise in one of four packing motifs (Fig 6.4). To analyse the effect of nitrogen substitution on these tendencies, the predicted crystal structures were categorised according to their packing motif, using an nearest-neighbouring angle-based metric. First, we classified all dimers formed between the molecule in the asymmetric unit and all molecules within a 20 Å distance cutoff according to the angle between their principal moments-of-inertia. Crystal structures in which all the angles are between 0–9° are classified as sheet structures (the β packing type). For structures where some dimers are not co-planar, the packing type is assigned using the four nearest neighbouring molecules. Structures in which none of the four nearest neighbours are co-planar are classed as herringbone packing. Where only one of the nearest neighbours is co-planar, the structure is classified as sandwich herringbone. Two or more co-planar neighbours indicates a stack of molecules, so these structure were classed as the γ packing type. This last category contains traditional γ structures and more sheet like structures, where parallel sheets are tilted along the short axis of the molecule (usually 3–10°). The classification of structures was checked manually, and an additional category of "other herringbone" was created, to encompass structures whose packing did not fit into one of our four initial categories.

The set of rules used for classification should be good enough to uncover trends in properties between these broad families of packing types. The classification of borderline

cases would be affected by a change in the angle bounds, which were taken from an initial visual analysis of crystal structures. A faster and more rigorous method of classifying predicted crystal structures into structural families would facilitate the analysis of crystal energy landscapes, and is currently being developed in our group.

6.3.3 Mobility calculations

In addition to classifying the packing motifs seen in our predicted structures, the mobility of charge carriers was calculated to elucidate the relationship between packing motifs and charge mobility. The calculations are performed as described in Chapter 2, with the hopping rates of charge carriers calculated using Marcus Theory,

$$k_{et} = \frac{t^2}{\hbar} \sqrt{\frac{\pi}{\lambda_{\pm} k_B T}} \exp \left[-\frac{\lambda_{\pm}}{4k_B T} \right]. \quad (6.1)$$

where t is the transfer integral (a measure of the overlap of molecular wavefunctions), λ the reorganisation energy (the reaction of the molecular geometry to a charge carrier landing on the molecule), λ_{-} is used in this case as all azapentacenes are expected to be electron transporters. For the calculation of the transfer integrals present between unique dimer pairs in the crystal the nearest-neighbouring molecular dimer electronic coupling matrix elements were calculated using subsystem density functional theory¹⁴⁴ at PW91²⁶⁵/DZ level of theory, as implemented in the Amsterdam Density Functional²⁶⁶ (ADF) package. For each crystal, the nearest-neighbouring dimers were extracted based on the criterion that at least a pair of intermolecular atom–atom distances was less than the sum of van der Waals radii plus 1.5 Å. In the case where the energy minimized crystal structures contain more than one symmetrically independent molecules, the nearest-neighbouring dimer is extracted for each unique molecule in the structure. To reduce the total number of DFT calculations required, coupling matrix elements for duplicated molecular dimers in a given crystal were not calculated explicitly. The duplicated dimers were identified based a root-mean-squared-deviation (RMSD) between two dimers being less than 0.1 Å. In this way, we ensured that all dimers that will play dominant contributions to the electron transport properties in a given crystal have been properly accounted for. Crystal structures within a 7 kJ/mol energy cutoff of the global minimum crystal structure. This is the expected range in which observed polymorphs are expected to be found¹.

The intramolecular reorganisation energies were calculated in Gaussian 09 at B3LYP/6-31G** level of theory. The electron diffusivities (\mathcal{D}) were calculated in a similar approach to Sokolov *et al*²⁶⁷:

$$\mathcal{D} = \frac{1}{2nN} \sum_{i=1}^N \sum_{j=1}^{N_i} r_{ij}^2 k_{ij}^2 \mathcal{P}_{ij}, \quad (6.2)$$

where N is the number of molecules in the crystal, N_i is the number of nearest-neighbours for the i -th molecule, r_{ij} is inter-centroid distance and $n = 3$ is the dimensionality of diffusion. \mathcal{P}_{ij} is the probability for the charge carrier to hop between molecule i and j , and is calculated as

$$\mathcal{P}_{ij} = \frac{k_{ij}}{\sum_{j=1}^{N_i} k_{ij}} = \frac{t_{ij}^2}{\sum_{j=1}^{N_i} t_{ij}^2}. \quad (6.3)$$

The intermolecular electron transfer rate (k_{ij}) is calculated according to Eq. (8.3). The final electron mobility is calculated based on the Einstein relationship as

$$\mu = \frac{e}{k_B T} \mathcal{D}. \quad (6.4)$$

All calculations were performed at $T = 300$ K.

6.4 Results and discussion

6.4.1 CSP

6.4.1.1 Validation molecules

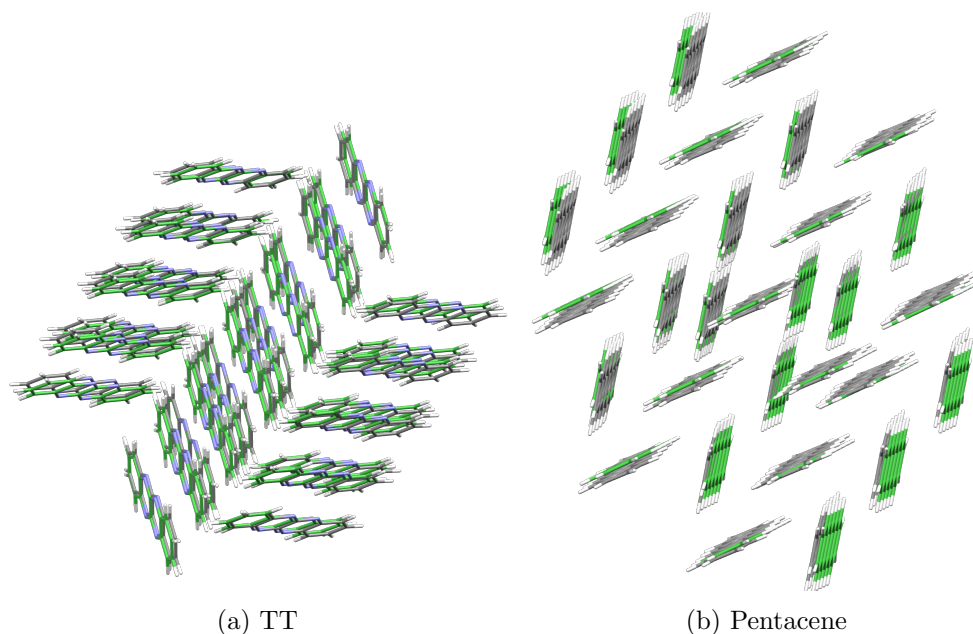


Figure 6.5: Overlays of the global minima (green) from our validation CSPs with the experimental structures for TT (RSMD₃₀ = 0.355 Å) and pentacene (bulk, RSMD₃₀ = 0.393 Å).

All three known polymorphs of pentacene crystallise in space group $P\bar{1}$. The distinguishing character between these polymorphs is their $d_{(001)}$ spacing, which is the distance

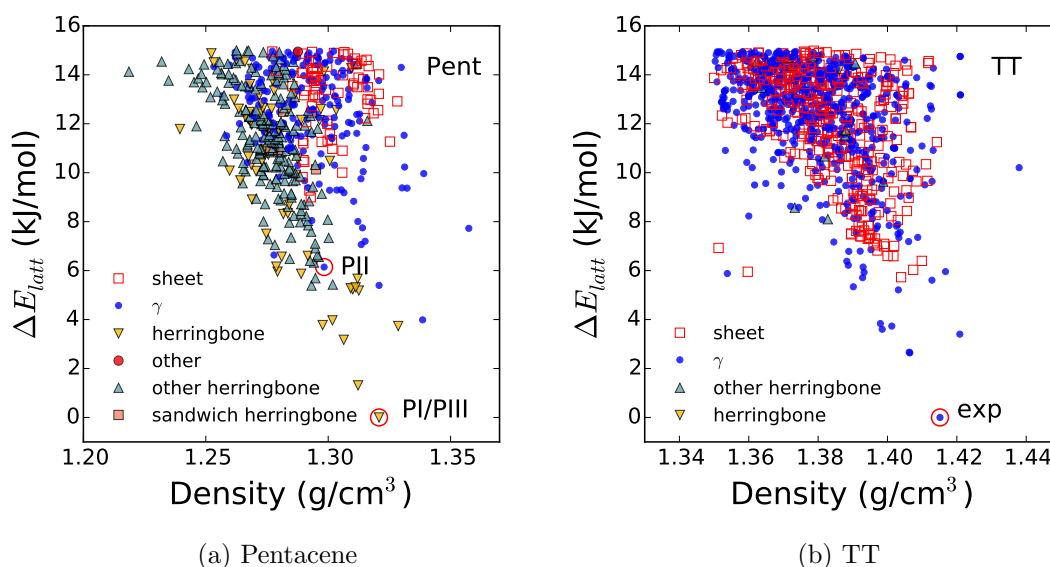


Figure 6.6: Structure–energy landscapes for the predicted crystal structures of pentacene. (a) and (b) are the low–energy (within 15 kJ/mol above the predicted global minimum) lattice energy landscapes for pentacene and TT, respectively, where each point is coded with respect to its crystal packing type.

Table 6.2: Matches from the full CSP to experimentally determined structures of the observed polymorphs. RMSD_{30} is the deviation in atomic positions of a cluster of 30 molecules taken from predicted and unoptimised experimental structures, not including hydrogen atoms. All structures were converted to their reduced unit cell for comparison. Å and degrees are used throughout.

Crystal Structure		Cell lengths			Cell angles			RMSD_{30}
		a	b	c	α	β	γ	
PI (bulk) ($P\bar{1}$)	expt	6.060	7.900	14.884	101.90	112.60	85.80	-
	pred	5.889	8.215	14.847	97.87	99.10	93.64	0.393
PII (single crystal) ($P\bar{1}$)	expt	6.275	7.888	14.709	76.01	87.23	85.00	-
	pred	5.973	8.015	15.219	77.11	83.93	86.22	0.526
PIII (thin film) ($P\bar{1}$)	expt	5.958	7.596	15.610	81.25	86.56	89.80	-
	pred	5.889	8.215	14.847	97.87	99.10	93.64	0.396
TT ($P2_1/c$)	expt	4.710	14.910	7.652	90.00	94.70	90.00	-
	pred	4.881	14.328	8.768	90.00	117.27	90.00	0.355

between repeating herringbone layers²⁶¹. For comparison with the predicted structures, the three experimental polymorphs were lattice energy minimised using the same energy model and molecular geometry used in the CSP calculations. We observed that, upon minimisation, both the bulk (PI) and thin film (PIII) phases converge to the same structure, which corresponds to the global minimum from CSP (Fig 6.6). This points to the difference in $d_{(001)}$ spacing seen between PI and PIII possibly arising from substrate effects that the pentacene thin-film is laid upon. In addition, this agrees with an earlier study on the lattice dynamics of pentacene polymorphs performed by Della Valle *et al*²⁶⁸. The single crystal phase was ~ 6 kJ/mol above the predicted global minimum

and was also located in the set of predicted crystal structures. As expected, herringbone packing is present throughout the clustered results with the main differences arising from interactions between herringbone pairs. The global minimum of the TT search corresponds well to the known experimental structure. The packing landscape is quite different to that of pentacene (Fig 6.6b, $N\cdots H-C$ hydrogen bonds can now form and the preference for herringbone packing is weakened. Sheet-like structures and γ are now the preferred motifs for low energy crystal structures. It must be noted that these sheet-like structures are not the "brickwork" packing that is expected to be optimal for charge transport, but are more similar to flattened gamma/herringbone motifs. For pentacene, the sheet structures seen are a mixture of this and conventional sheet structures. Isolated cases of sandwich herringbone packing can still be found, albeit being higher up in the energy range on the pentacene landscape.

Table 6.2 shows the cell parameters of the matches to known experimental structures. The results from the validation studies are encouraging for applying CSP to the theoretical azapentacenes. Both global minima of the validation searches match a known experimental structure and Table 6.2 shows good agreement with the experimental cell parameters.

6.4.1.2 Hypothetical azapentacenes

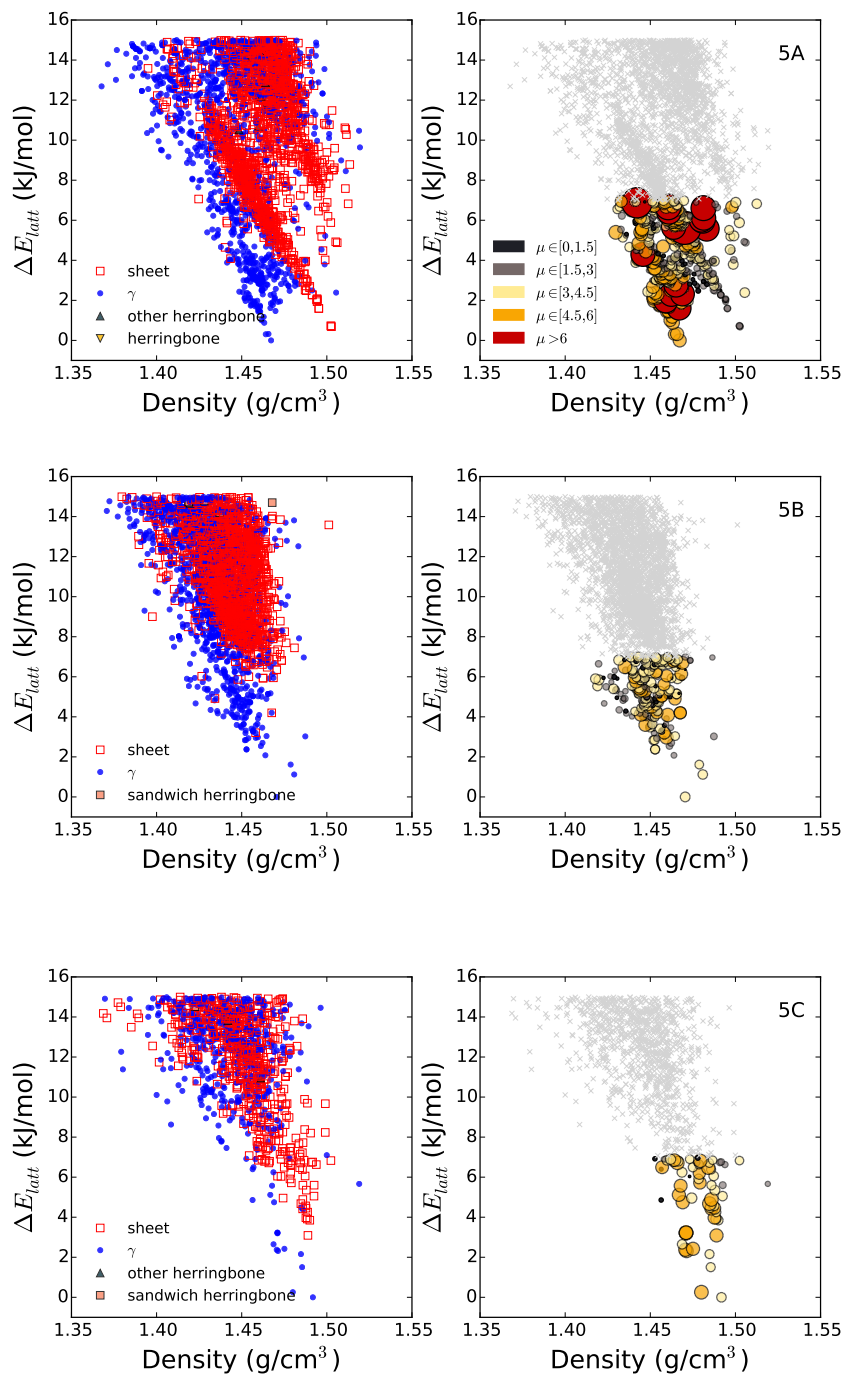


Figure 6.7: Structure–energy–mobility landscapes for the predicted crystal structures of azapentacenes with 5N. Colouring and the size of circles on the right–hand–side correspond to the magnitudes of calculated electron mobilities in cm²/Vs.

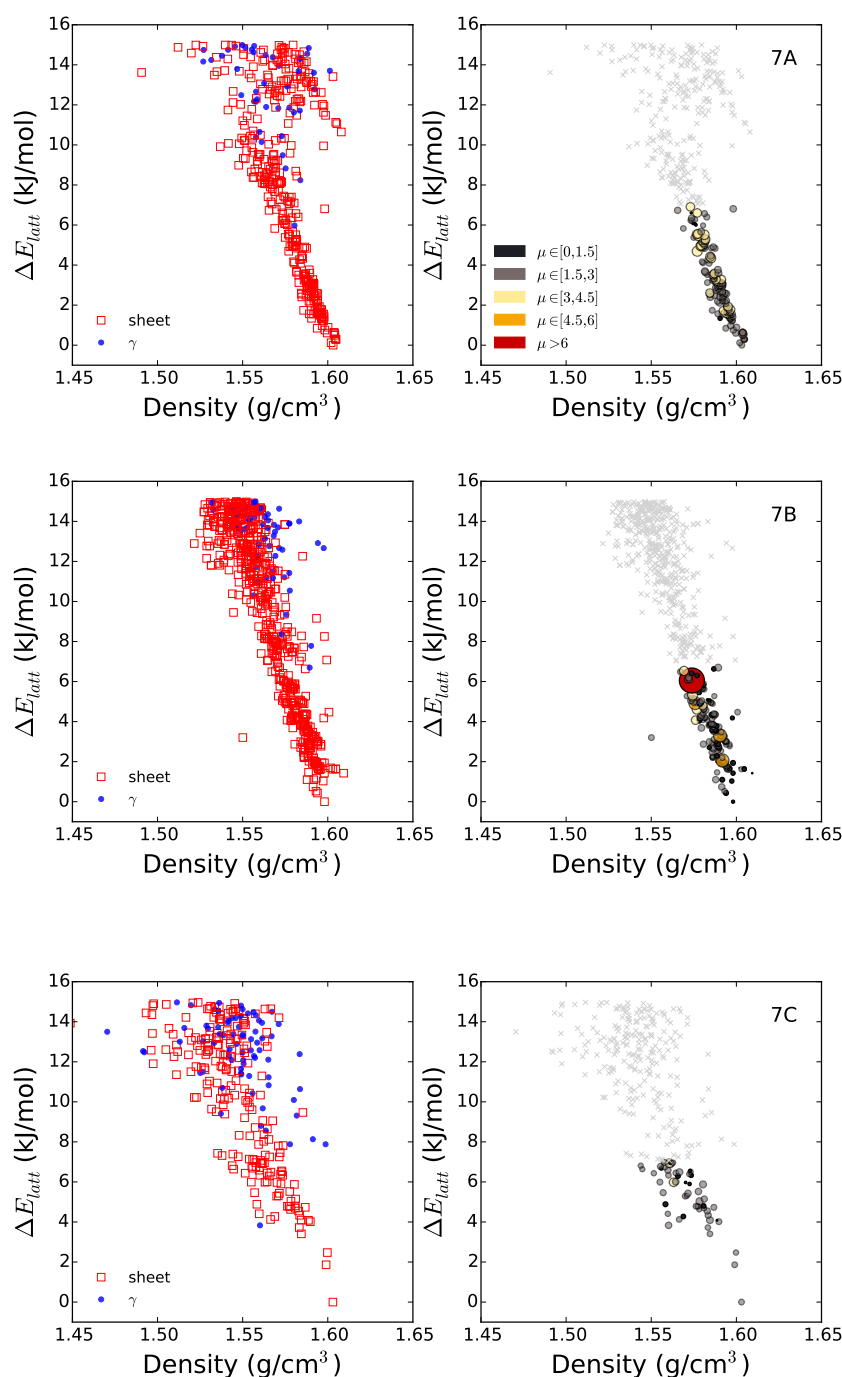


Figure 6.8: Structure–energy–mobility landscapes for the predicted crystal structures of azapentacene with 7N, mobilities are given in cm^2/Vs .

We now move onto the results for the series of hypothetical 5N and 7N azapentacenes. First, we analyse the preferred packing motifs of the series of azapentacenes. Similarly to TT, a further increase of N–substitution to 5N leads to a further change in the packing landscape of the hypothetical azapentacenes. The lattice energy landscape

for the three 5N substituted azapentacenes were dominated by γ and sheet packings. $\text{CH} \cdots \text{C}$ edge to face interactions are mostly replaced with $\text{N} \cdots \text{HC}$ hydrogen bonds along the long axis of the molecule, which directs the sheet-like motifs. Furthermore, sheet-like structures were more apparent in 5N-substituted azapentacenes compared to the validation molecules.

5A exhibits a band of sheet packing structures as in Fig 6.9 (the third lowest energy sheet structure is shown in 6.9a), these sheets show small offsets in regards to other layers approaching a brickwork motif. Fig. 6.7 shows that the number of N substituents is not enough to completely disrupt the γ -motif, such that 5N-substituted azapentacenes exhibit a range of packing motifs. This also led to a crowded, polymorphic landscapes for 5N-substituted azapentacenes. In particular, despite our expectation that 5C might exhibit poorer packing, because the long edges of the molecule are not complementary, 5C shows similar packing preferences to compared to 5A and 5B.

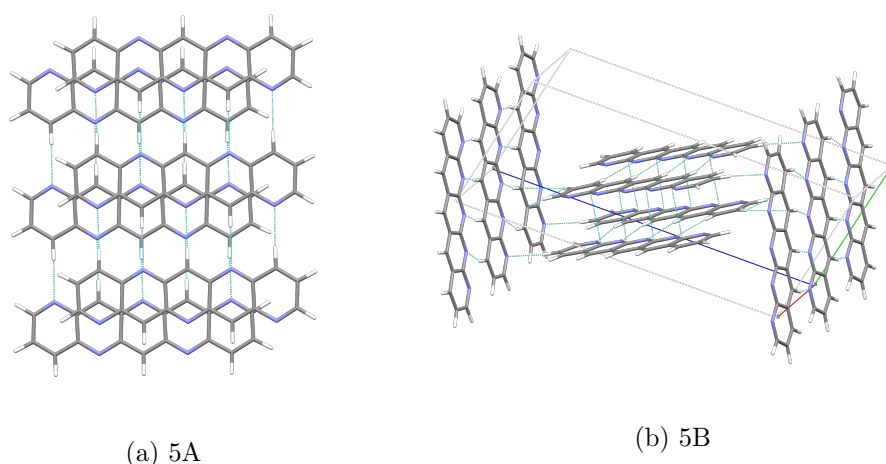


Figure 6.9: Sheet and γ herringbone directing interactions in the 3rd lowest 5A structure and the 5B global minimum

The addition of two extra N atoms has a strong impact on the dominant packing motif seen. All 7N containing molecules exhibit a strong preference for sheet-like packing across the low energy range and show a much sparser lattice energy landscape than the 5N azapentacenes. The addition of nitrogen atoms to the corners of the molecules strongly favours sheet formation, that are held together with hydrogen bonds in triangular patterns (Fig 6.10). This interaction pulls the sheets together, creating more overlap with the sheets above and below. Edge-to-face interactions can no longer direct herringbone packing without the significant expense of losing $\text{C-H} \cdots \text{N}$ hydrogen bonds. The vertical distance between layers is slightly shorter than what is usually seen in π -stacked materials²⁶⁹ (around 3.2 Å for predictions, compared to 3.4-3.6 Å experimentally observed). 7A and 7B sheets differ however, in the vertical displacement between layers, which is larger for 7A structures. This vertical displacement is 0.334 Å for the global minimum of 7B and 1.091 Å for the global minimum of 7A. Again, the unsymmetric substitution pattern in 7C led to little difference on the resulting packing motifs. Low

energy structures for 7C are mostly sheets with slightly more variation than seen in 7A/B as neighbouring molecules are primarily shifted along the long axis (Fig 6.10c).

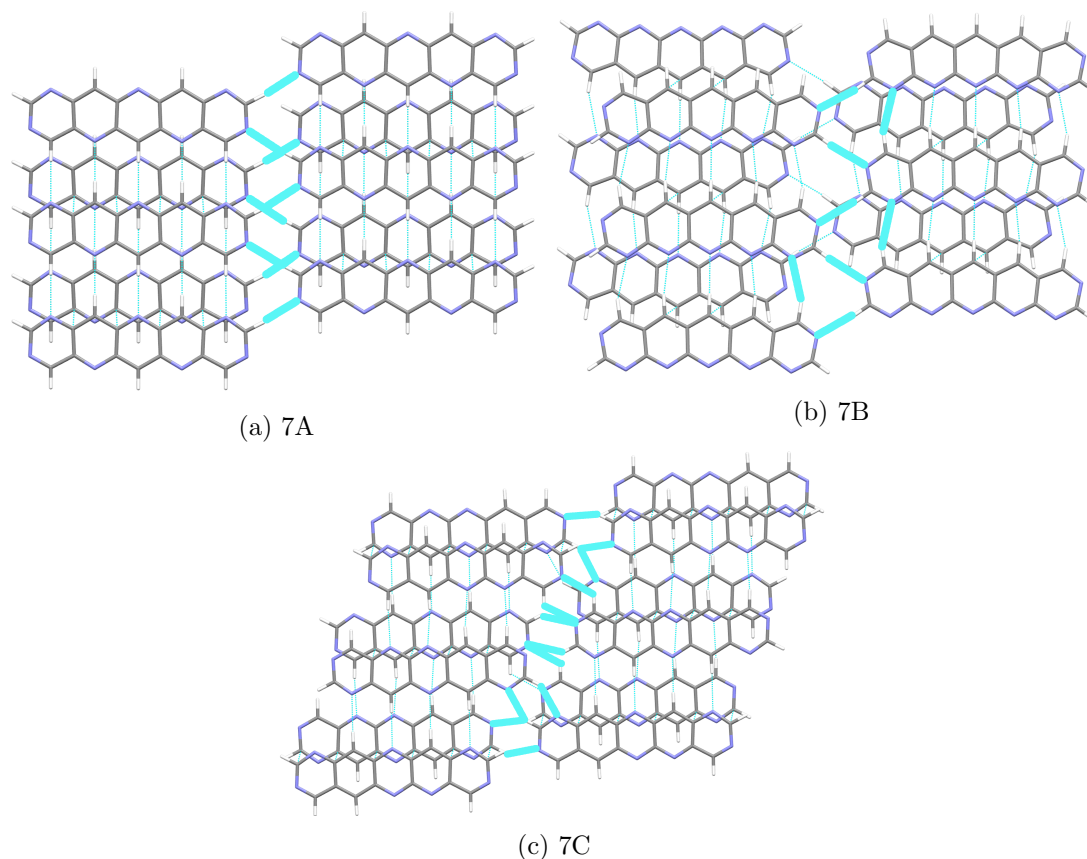


Figure 6.10: Sheet motifs and packing seen in the global minima of 7A, 7B and 7C, showing triangular hydrogen bonding connecting parallel sheets and varying levels of offset.

6.4.2 Charge mobility of predicted structures

The charge transport properties of crystalline organic semiconductors depend on the subtle interplay between the intrinsic molecular properties (characterised by the reorganisation energy of individual molecule in the gas-phase) as well as the packing geometries (characterised by the transfer integral) in the observed polymorphs. The relative weightings of these two parameters is not entirely known however. This section aims to investigate this interplay for the series of hypothetical azapentacenes and their predicted crystal structures.

Table 6.3 summarises the gas-phase electron reorganisation energies for all six azapentacene molecules investigated, from which it can be seen that molecule 5A is the molecule with the lowest reorganisation energy. This leads to the fact that the predicted crystal structures for molecule 5A (up to 7 kJ/mol above the global minimum), are, in general, of higher charge mobilities compared with the other five molecules (Figs. 6.7 and

Table 6.3: Summary of the charge transport parameters for the azapentacene molecules investigated: λ_e is electron reorganisation energies, calculated at B3LYP/6-31G** level of theory. μ_{\max} is the maximum predicted electron mobility among the predicted crystal structures. $\Delta E(\mu_{\max})$ is the lattice energy gap between the crystal structure with the highest charge mobility to the predicted global minimum. $\langle\mu\rangle$ is the ensemble-averaged electron mobility across all crystals with calculated mobilities. μ_{gm} is the mobility of the global minimum.

Molecule	λ_e (eV)	μ_{\max} (cm ² /Vs)	$\Delta E(\mu_{\max})$ (kJ/mol)	$\langle\mu\rangle$ (cm ² /Vs)	μ_{gm} (cm ² /Vs)
5A	0.151	11.4	5.62	3.27	5.36
5B	0.165	5.91	5.34	2.86	3.98
5C	0.157	5.97	4.68	4.29	3.78
7A	0.180	4.22	4.69	2.52	2.10
7B	0.198	6.56	6.05	1.81	0.62
7C	0.184	3.16	5.98	1.91	2.01

6.8). This suggests that reorganisation energy dominates and should be the main target property in the design of novel organic semiconductors. This confirms the relationship between pentacene’s excellent mobility in spite of its poor crystal packing.

The relationship between the crystal packing and the mobility is also difficult to generalise. For the 5N molecules high and low mobilities are seen for both γ and sheet packings. For the 7N azapentacenes despite the prevalence of sheet motifs mobilities are lower than their 5N counterparts. This is principally due to their higher reorganisation energies.

Apart from the practical limitations in predicting charge mobilities solely based on packing-type classifications, it can also be seen in Figs. 6.7 and 6.8 that crystal structures with high mobilities are not necessarily the global minimum structures on the lattice energy landscapes. This is not surprising, as dimers with the largest electronic couplings are those with co-facial π -stacking, with good spatial overlap between the interacting molecular orbitals, the lattice energies of which could be penalised by exchange-repulsion between the interacting molecules. The purpose of performing CSP and mobility calculations in advance of synthesis would be to prioritise the molecules according to which is most likely to lead to a high charge carrier mobility. The question is how do we then rank the molecules studied here? The most obvious choice is the maximum charge mobility μ_{\max} among the low energy structures, and the corresponding energy gap to the predicted global minimum at which μ_{\max} was observed, ($\Delta E(\mu_{\max})$). It can be seen from Table 6.3 that while molecule 5A leads to the crystal structure with the largest μ_{\max} across the whole set, it is energetically penalised by a relatively high $\Delta E(\mu_{\max})$ value, which is the second highest among all six molecules. Molecule 5C could be a better choice, with the third highest predicted μ_{\max} along with the lowest $\Delta E(\mu_{\max})$. Analysing μ_{gm} (the mobility of the most likely structure) points to 5A being the best candidate for further development, though the mobility of the global minimum is less than half that of μ_{\max} .

The ranking function used needs to be able to balance the desire for the maximum mobility molecule, with the stability of its high mobility crystal structures. A expected mobility value can be calculated,

$$\langle \mu \rangle = \frac{\sum_i^N \mu_i \exp(-\Delta E_i/\beta)}{\sum_i^N \exp(-\Delta E_i/\beta)} \quad (6.5)$$

where μ_i is the electron mobility of the i -th crystal structure and ΔE_i is the lattice energy difference of this structure to the predicted global minimum. The decay constant $\beta = 2.696$ kJ/mol is obtained by fitting the probability for observing a pair of molecular crystal polymorphs from reference 1. It can be seen from Table 6.3 that, molecule 5C is now a clear winner, an underdog win for a molecule that was originally included to give poor crystal packing. The high expected mobility is due to a landscape whose lowest energy structures mostly show high mobilities, in contrast with 5A where a large number of low energy structures have low predicted mobilities.

6.5 Conclusions

This chapter outlines the CSP study of potential novel organic semiconductors. The substitution of nitrogen atoms into the ring system of pentacene can modify the transport character and packing motif. Two molecules were chosen as validation for this study, pentacene and tetraazatetracene. CSP was successful for both molecules, with all experimental polymorphs located and the global minimum for both molecules corresponding to an observed polymorph. Though upon optimisation the thin film phase also matches the global minimum suggesting substrate effects are responsible for the difference in d_{001} spacing reported experimentally. CSP was then performed on 6 hypothetical azapentacenes with a range of substitution patterns and numbers of nitrogen. 5N containing azapentacenes show a disruption of the typical herringbone packing of pentacene with γ being the most common packing motif among the low energy crystal structures. However, sheet like structures are also found in low energy regions of our search. In contrast, 7N molecules pack almost exclusively in sheet motifs. All C-H \cdots C edge to face interactions are disrupted and replaced by CH \cdots N hydrogen bonds which promote coplanar molecular arrangements.

The charge transport properties of the predicted structures were then analysed to try and elucidate the relationship between packing motifs and charge mobilities. In terms of maximum charge mobilities, 5A was found to be the best performing molecule in our study in spite of its packing motif not being the most favourable *a priori*, resulting in the relatively large energy gap between structure of highest mobility and the predicted global minimum. A better compromise can be achieved with molecule 5C. Given the rather unusual N-substitution pattern in 5C, it is clear that computer-guided design of novel

organic semiconductors requires simultaneous exploration of both chemical space and crystal packing. The next chapter will introduce the design of genetic algorithm designed to search possible substitution schemes of azapentacenes. Chapter 8 will contain the "production" runs of this genetic algorithm and the CSP and mobility calculations of promising molecules discovered.

Chapter 7

Azapentacene Genetic Algorithm

7.1 Introduction

Chapter 4 introduced machine learning and genetic algorithms (GAs). The parts needed for a successful GA were also discussed. Factors to take into consideration include:

- i) the encoding of the population members; this varies depending on the problem being studied, binary encodings are common but their transferability may be limited
- ii) how the initial population is generated, whether completely random or biased in some way to ensure evolution in the desired direction
- iii) the nature of the fitness function. This is problem dependent. As fitness is calculated possibly 100s of times per generation, this needs to be relatively quick. Obviously this also needs to represent how well a member of the population is at solving the problem in question, the calculation of the fitness also needs to be accurate enough to differentiate between similar population members
- iv) how the population members are ranked and selected for crossover. The selection method can have a large impact on the speed at which the GA reaches a minimum. Generally, slower convergence is better to stop convergence on a local minimum
- v) the choice of how the new generation is made. Elitism transfers a proportion of the fittest members into the new generation. This is useful as it can stop the disappearance of fit members. However if the proportion of elite members is too high they can dominate further generations
- vi) which crossover operators are used. If more than one how likely they are to be chosen. Single point crossover is the simplest but has drawbacks. Extensions

such as two point and uniform crossover also have pros and cons, again this is largely problem dependent. For uses such as genetic programming, it is often advantageous to retain as much of a successful member as possible

- vii) the use of mutation, changing one bit of a new member randomly can help retain diversity in the population.

When these choices have been made there are many parameters inside the GA that need to be optimised. Population size, crossover rate (ratio of crossover operators if more than one are used), numbers inside the crossover operator, elitism rate, mutation rate and various values within the chosen selection method all need to be decided upon. There is no general best set of values for these parameters as the performance of the GA depends on the interplay between encoding, selection method, crossover and mutation. These parameters are usually developed in step with the GA before any "production" runs are attempted.

Chapter 6 contained the CSP of and evaluation as potential organic semiconductors of a number of human designed azapentacenes. The results there showed that what may seem as a strong substitution pattern at the time may not lead to the best performance due to the complex interaction of crystal packing and molecular electronic properties. What is the best way to explore the substitution space then? There are many possibilities: too many to explore systematically. Therefore a GA calculating the fitness of many molecular candidates at once and driving towards the best one, combined with the CSP of promising molecules is one approach to this problem. This chapter will describe the development and testing of a GA designed to accomplish this task. The first half will discuss the structure of the GA, the second half the testing that was carried out to choose the best parameters. We will apply the resulting GA to the exploration of azapentacene molecules as potential n-type organic semiconductors.

7.2 The GA

7.2.1 Encoding

The first step in designing the GA is the decision on the encoding to be used. While binary bitstrings are common their usefulness can be limited. For molecular chemical applications perhaps a more sensible approach is to use an encoding that represents the molecule as a string in an easy to understand way using known chemical symbols. Preferably, this encoding would that includes atomic (for our nitrogen atoms in azapentacenes) and bonding information. Additionally the encoding should be amenable to the action of crossover operators and represent the molecule uniquely.

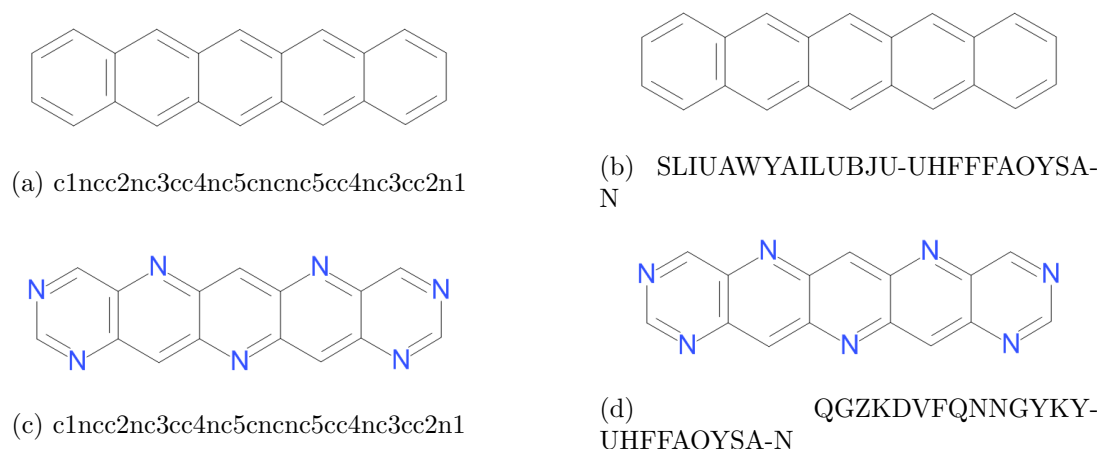


Figure 7.1: Two potential encoding schemes for the GA on pentacene and 7A, (a) and (c) are SMILES strings, (b) and (d) are InChi key strings.

Two common schemes that accomplish this are the simplified molecular-input line-entry system (SMILES)²⁷⁰ and IUPAC International Chemical Identifier (InChi)²⁷¹. SMILES uses short ASCII strings to represent molecules, by the rules of generating the strings, SMILES strings are not unique. For example CCO, OCC and C(O)C all represent ethanol. Algorithms to ensure uniqueness are included in most cheminformatics packages, a process called canonicalisation, as long as one package is stuck to the SMILES will be unique. In terms of information encoded, a SMILES string contains the same information as an extended atom connectivity table. However it is much more compact than the same table would be. A SMILES string is created by printing the nodes encountered in a depth first (selecting an arbitrary atom as the starting point and exploring as far as possible before backtracking) tree traversal (visiting each atom exactly once) of the chemical graph. Hydrogens are removed first and rings are broken into a spanning tree. Where the rings have been broken numbers are used to indicate connected atoms. Branching of the molecule is represented by parentheses. Fig 7.1c shows this for the azapentacene 7A.

InChi another ASCII character encoding scheme can include more information than a SMILES string. Each part of the string includes different information of the molecule, from atom and bond connectivity to tautomeric information, isotope information, stereochemistry, and electronic charge information. In addition each molecule has a unique InChi, without the need to worry about conversion between different programs. While an InChi is unique it is rather unwieldy, a hashed version (an InChi key) has been developed but this can result in the uniqueness being lost. Additionally it is not possible to reconstruct the InChi string from the key, they must be linked somehow.

As can be seen from Fig 7.1d an InChi key is long even for simple molecules, much information is included that is not needed and conversion between InChi keys and the full string can be difficult. SMILES strings however, are easier to read, store and as long

as the same package is used to generate them, uniqueness should not present a problem. In addition, a SMILES string is easier to code crossover operators for. For these reasons SMILES was chosen to encode our molecules and for use in the GA.

After selecting SMILES strings as our encoding method a decision needed to be made about whether to develop our own SMILES generator or use one in a cheminformatics package. RDKit²⁷² is an open-source cheminformatics and machine learning package. It not only has a SMILES generator but many additional features useful for our GA. RDKit contains a molecule class that can be generated from a SMILES string, which includes code for the breaking and reforming of bonds and can identify specific atoms. This functionality is very useful for potential crossover functions. In addition, RDKit can generate molecular fingerprints, save molecule objects with information attached (such as properties) and can integrate with Ipython/Jupyter to depict molecules.

7.2.2 Generating a population

Most GAs begin with a randomly generated population to ensure diversity in the starting point for the GA. Seeding¹⁷⁴ can be used to encourage a specific direction in the search, through inserting population members you would like to see propagate through the GA. Due to our wish to explore the substitution space of pentacene thoroughly an entirely random population was chosen as the starting point. First the SMILES string of pentacene is randomly modified by replacing "c" atoms with "n". The code must ignore the numbers present in the string, and not modify any of the 3-centre carbon atoms in the fusing points of the rings. RDKit contains functions to identify the number of bonds for each atom allowing these to be easily skipped. The pseudocode below shows this in practice,

```
def mutator(smiles_list, mutation_factor, N_number):
    while smiles_list.count('n') != N_number:
        smiles_string = ''.join(smiles_list)
        m_test = Chem.MolFromSmiles(smiles_string)
        bond_list = [len(a.GetBonds()) for a in m_test.GetAtoms()]
        for i, atom in enumerate(smiles_list):
            if atom == int:
                pass
            if len(bond_list[i]) > 2:
                pass
            elif random.randint(0,100) <= mutation_factor:
                if smiles_list.count('n') < N_number:
                    atom = 'n'
                elif smiles_list.count('n') > N_number:
```



```
atom = 'c'
smiles_list[i]=atom
test_string = ''.join(smiles_list)
test_m = Chem.MolFromSmiles(test_string)
if smiles_list.count('n') == N_number:
    return smiles_list
```

where the mutation factor is 50 (to ensure a high chance of atoms being changed), `N_number` is the number of nitrogen atoms to be placed in the molecule and the `smiles_list` is the python list of the SMILES string. For initial testing, a fixed number of N was placed into the molecule, though creating a population of mixed amounts of nitrogen atoms is possible. The output from this function is shown in Fig 7.2 (it should be noted that hydrogen atoms are omitted from all RDKit representations but are added on before 3D coordinates are generated).

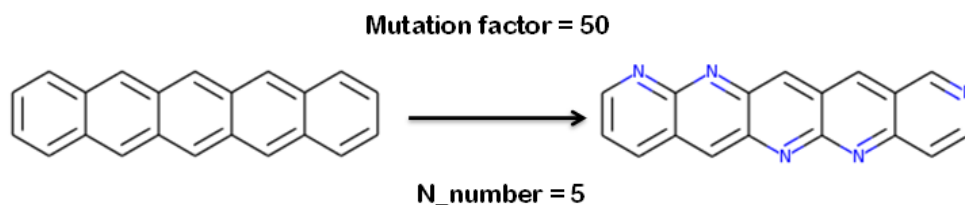


Figure 7.2: The input and output of the mutator function used to generate members of the initial population.

Creating the population then involves running the mutator function on pentacene upto the population size necessary. To check for duplicates within the population a new molecule was checked against the current population using the SMILES string and again using the reversed SMILES (to catch any symmetric matches). Later improvements added the use of a molecular fingerprint at this step (to be discussed in a later chapter). A sample of a generated population for $N=5$ is shown in Fig 7.3.

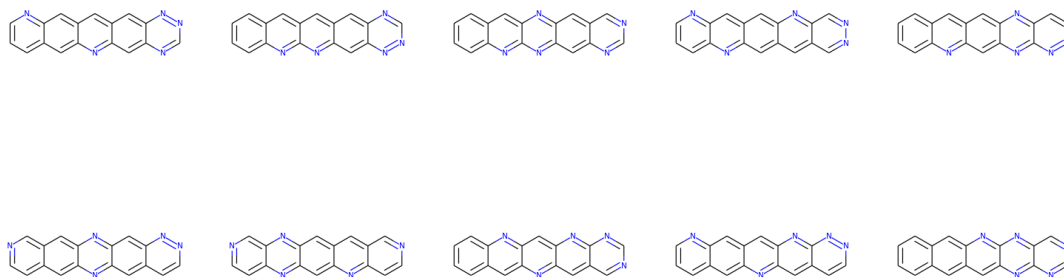


Figure 7.3: Ten members of a randomly generated initial population of 5N substituted pentacenes.

7.2.3 Calculating fitness

The fitness function of a GA is an important part of ensuring the GA locates the best solution. As more and more solutions crossover, the fitness landscape becomes more complicated and it is possible that many solutions are very close to one another. Any accuracy lacking in the fitness function makes it possible that the GA will not find the correct minimum. In testing our GA a simple fitness function was implemented as it was not yet known what final form would be best, so a number of molecular properties were chosen to optimise. The most important one for our application in discovering novel organic semiconductors was the molecular reorganisation energy (λ). In addition runs using the energy of the lowest unoccupied molecular orbital (LUMO), the molecular dipole moment and the energy difference between the LUMO and highest occupied molecular orbital (HOMO) were performed.

Using λ as a measure of fitness formed the bulk of the testing due to the fact that it has a known minimum. Pentacene has an excellent hole λ but also a very good electron λ . Therefore, using the GA to minimise λ should lead to pentacene as the global minimum. λ is calculated as the sum of the energy required for the reorganisation of the vertically ionised neutral geometry to the anion geometry, plus the energy required to reorganise the anion geometry back to the neutral equilibrium. The process of calculating the reorganisation energy requires four calculations: *a*) a geometry optimisation of the neutral molecule; *b*) a geometry optimisation of the molecule with a negative charge; *c*) a single point energy of the *a* geometry with a negative charge and *d*) a single point energy of the geometry of *b* with a neutral charge. The energies from these calculations are then summed as,

$$(c - b) + (d - a) \quad (7.1)$$

or more formally,

$$\lambda_- = E_{0opt}(Q_-) - E_-(Q_-) + E_{-opt}(Q_0) - E_0(Q_0) \quad (7.2)$$

where the subscript on E refers to the geometry of the molecule and Q the charge. For the hole reorganisation a similar procedure would be followed, replacing the negative charges with positive ones.

Calculations were performed using Gaussian09⁷¹ at the B3LYP/6-31G**⁷² level of theory. RDKit allows the generation of 3D coordinates from a molecule object. To give a good starting point for the λ optimisations all molecules are optimised using UFF²⁷³ before the Gaussian09 input file is generated.

7.2.4 Selection methods

Four of the selection methods discussed in Chapter 4 are implemented in our GA code: fitness proportionate sampling (FPS); stochastic universal sampling (SUS); tournament selection and rank selection. While no large scale testing was done tournament selection was chosen as the selection method for the GA. Exploring the substitution space of azapentacenes is our goal, so ensuring diversity throughout the GA run is important. Both FPS and SUS present problems with keeping diversity present in generations. At the beginning of a GA run when fitness variance may be high, both SUS and FPS will select more of the higher fitness members than is wanted reducing diversity and increasing the chance of the GA stopping at a local minimum. Rank selection can mask the fitness variance present in the population although it lacks user control over the selection pressure. Tournament selection, however, masks the fitness variance and has an easy way to control the selection pressure of the GA. Tournament selection involves running tournaments among a number of (usually 2) randomly selected members of the population. A random number between 0 and 1 is selected and if it less than a user defined parameter (usually around 0.75) the fittest individual in the tournament goes through to crossover. Otherwise the less fit member will go through. Selection pressure can be easily adjusted by changing the size of the tournament, as larger tournaments disadvantage weaker members. The implementation of tournament selection in our GA is as shown in the code below,

```
tournament_selection(mol_list_en, elitism_rate, new_pop_size):  
    elite_mols = []  
    crossover_mols = []  
    p0 = 75  
    for i,mol in enumerate(mol_list_en):  
        if i < elitism_rate:
```

```

        elite_mols.append(mol[0])
    while len(crossover_mols) < new_pop_size:
        challenger_1 = random.randint(0,(new_pop_size-1))
        challenger_2 = random.randint(0,(new_pop_size-1))
        print str(challenger_1)+" versus "+str(challenger_2)
        winner = random.randint(0,100)
        if winner <= p0:
            print "stronger won"
            crossover_mols.append(mol_list_en[challenger_1][0])
        if winner > p0:
            print "weaker won"
            crossover_mols.append(mol_list_en[challenger_2][0])
    return crossover_mols,elite_mols

```

where, `mol_list_en` is the list of molecules ranked by their fitness, `elitism_rate` is the amount of that generation to take through to the next unchanged and `new_pop_size` is the total size of the next generation. Elitism (the keeping of fittest members of the current generation unchanged for the next one) is included in our GA, keeping 10% of the total population size. Elitism prevents the loss of fit members from generation to generation and can keep the GA on track towards the global minimum.

7.2.5 Crossover

Two methods of approaching the crossover of our molecules are implemented in our GA, with three crossover operators using these methods. The first crossover method consists of only modifying those atoms on the outside of the ring system and is referred to as atomic site crossover. The genome for crossover then consists of 14 atoms (Fig 7.4).

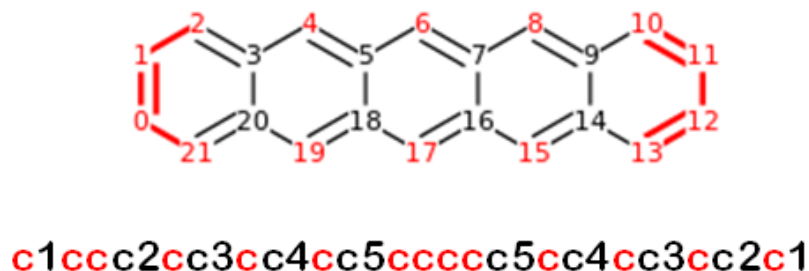


Figure 7.4: The atoms highlighted for the atomic site crossover and their position in the SMILES string.

The crossover genome is a list of these atoms at their positions as illustrated in Fig 7.5. This list is made for each and every parent for use in the crossover operators. Once the crossover operators have been applied, a "blank" pentacene SMILES string is modified using the new genome, changing atoms at the positions specified.

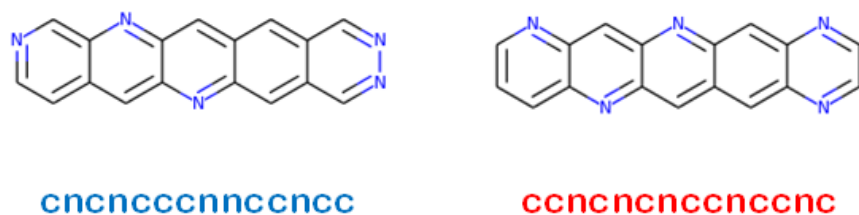


Figure 7.5: Two azapentacene molecules with their crossover genomes.

The three implemented crossover operators are single-point, two-point and uniform. Single-point crossover is the simplest: a random number between one and fourteen is chosen (n) and the parent genomes are cut at this point. The first offspring will contain n of parent one, with the rest of the genome being made up from parent two. The second offspring will be the inverse of the first, with n of the genome coming from parent two and the remaining part from parent one. Fig 7.6 shows this for the two azapentacene molecules in Fig 7.5.

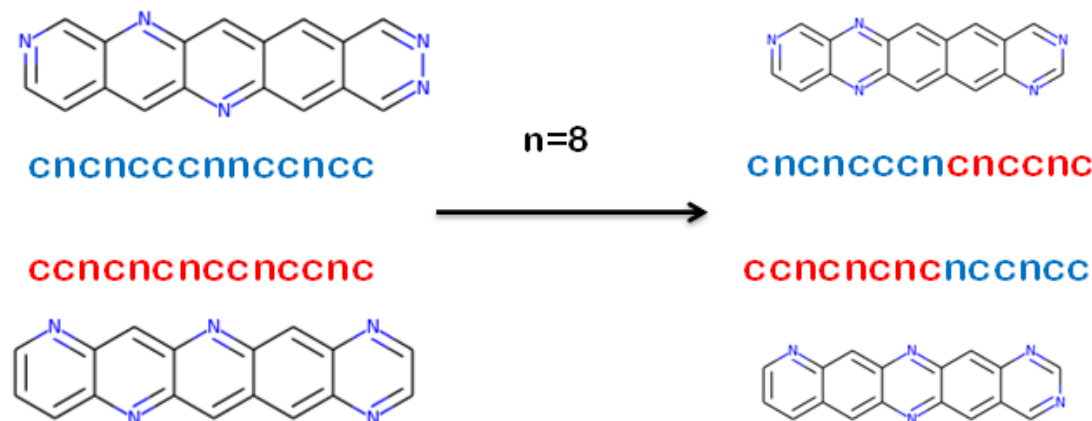


Figure 7.6: Single-point crossover for the two azapentacene molecules from Fig 7.5.

Single-point crossover has known drawbacks. It preferentially saves the "ends" of the genome each time and only short parts of the genome are swapped. It is possible that longer parts of the genome are needed for passing on maximal fitness and the ends that are kept may be low fitness parts of the genome. Two-point crossover attempts to reduce this positional bias. In two-point crossover two random numbers are selected, n_0 and n_1 ,

defining a segment in the parents to be exchanged. This results in child one containing upto n_0 of parent one, between n_0 and n_1 of parent two, with the rest being made up of parent one. Again the process is reversed to produce the inverted offspring. Fig 7.7 shows this in practice.

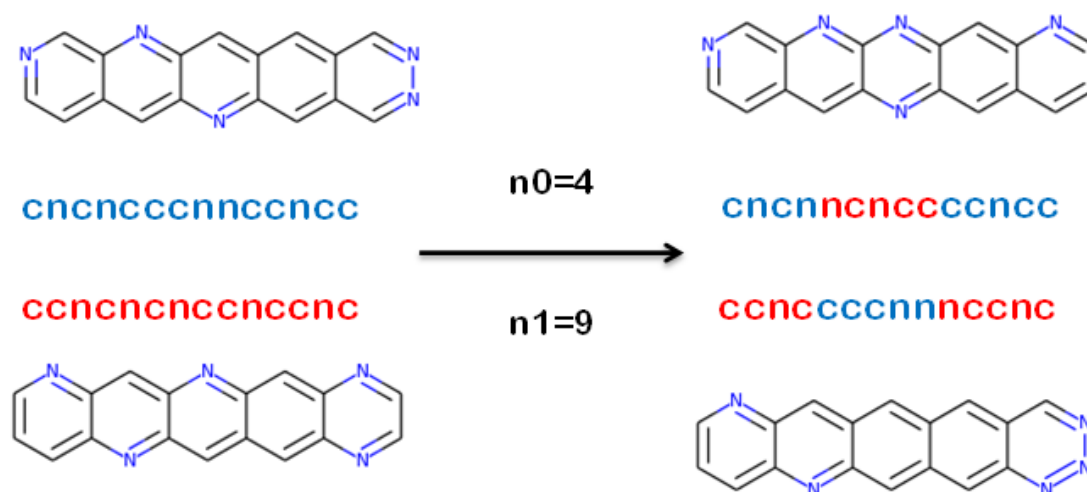


Figure 7.7: Two-point crossover for the two azapentacene molecules from Fig 7.5.

Two-point crossover is less likely to disrupt large parts of the chromosome and can combine more parts of the genome together. The final crossover operator included in our GA is uniform crossover. In this operator, exchange can happen at every point of the parents: for each point in the children a random number is chosen between 0 and 1. If the random number is above 0.5 take the character from parent one, if below 0.5 the character is taken from parent two. The second offspring is then the inverse of the first one (Fig 7.8). Fig 7.8 also shows an additional choice made during the development of the crossover, whether to fix the number of N atoms present in the population members. Initial development of the GA used the mutator function to keep the number of N constant, randomly removing or adding nitrogen atoms as necessary. This was more for convenience of coding a working example rather than a design choice and the number of nitrogens was allowed to vary to more effectively sample the substitution space.

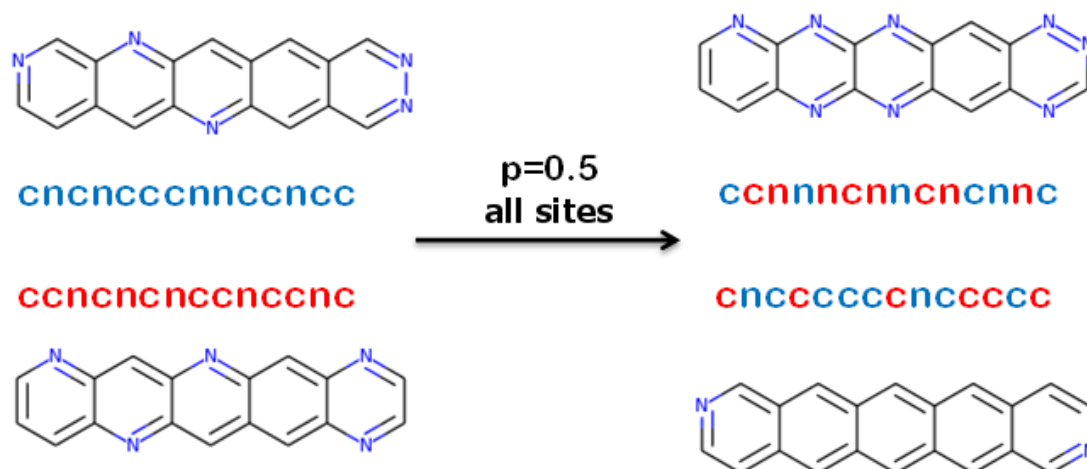


Figure 7.8: Uniform crossover for the two azapentacene molecules from Fig 7.5.

Debate around which crossover operator is best is still ongoing¹⁸⁴ as the success of the GA overall relies on the encoding, fitness function and the suitability of the problem for a GA. All three operators described above are included in our GA; each time a pair of parents is picked to create offspring the operator is selected randomly: two-point has a 60% chance to be chosen, single-point 30% and uniform 10%. Mutation is also included in the GA; once the next generation has been made there is a 5% chance of a member being selected and having one randomly chosen atom modified from C to N or vice versa.

While (as shown later) this crossover method performed well, it lacks the ability to generate new molecular shapes due to relying on a pentacene smiles as the base to which the crossover genome is applied. In addition, the selection of 14 atoms is not the most chemically intuitive way of crossing ring systems over. For these reasons a second crossover method based on fragmenting the molecules at ring joining atoms was developed (Fig 7.9) referred to as the ring crossover method. All three operators are used in this method with the same probabilities, although the numbers inside the operator functions are smaller as there are now only five parts of the entire genome.

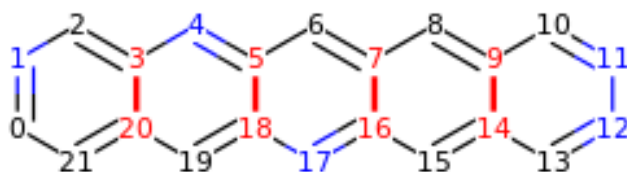


Figure 7.9: Highlighted atoms showing where rings are separated in the ring crossover method.

The ring crossover proceeds as follows. 3-centre carbon atoms are identified based on their number of bonds, the molecule is fragmented by breaking the bonds to the next ring and replacing the atoms with dummy atoms (represented in Fig 7.10 by *) using RDKit's editable molecule functionality. The corresponding fragments from both parents are then placed together as one molecular object, the dummy atoms are then bonded together before being replaced by carbon atoms again.

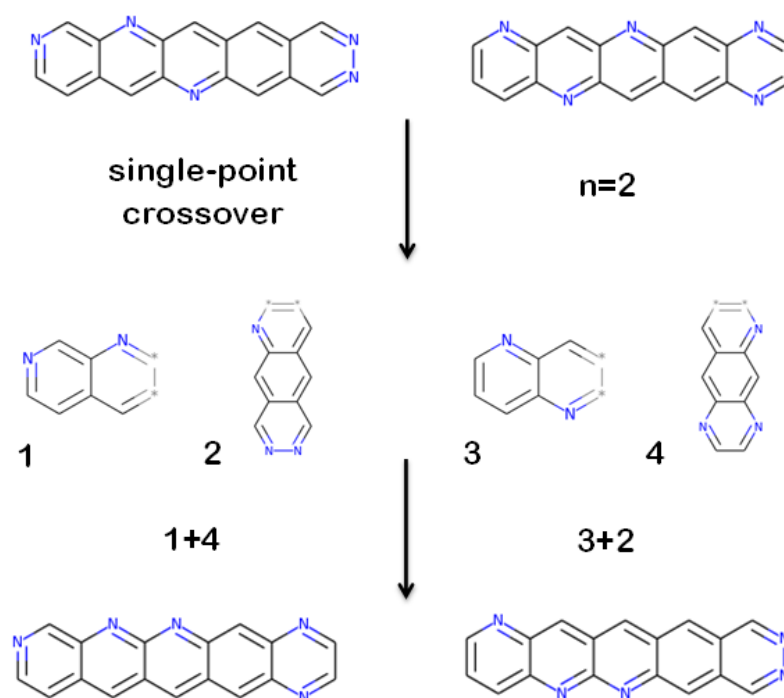


Figure 7.10: Two molecules fragmented for crossover, * represents dummy atoms that are placed into the fragments where the molecule was cut.

In addition to allowing for ring systems to be broken more intuitively this crossover method allows molecules to assume different shapes to a linear backbone. Before the crossover operator is chosen, we allow a 5% chance for the offspring to become branched. This involves moving the dummy atoms from where they were initially placed one atomic position up or down (Fig 7.11a). It is possible for branched molecules to be branched again in further generations leading to shapes vastly different from the original linear backbone (Fig 7.11b). Mutation is still present in this crossover method and is applied in the same way as in the atomic site crossover; applied to a single atom on the molecular periphery. Again having a 5% to occur once offspring have been created.

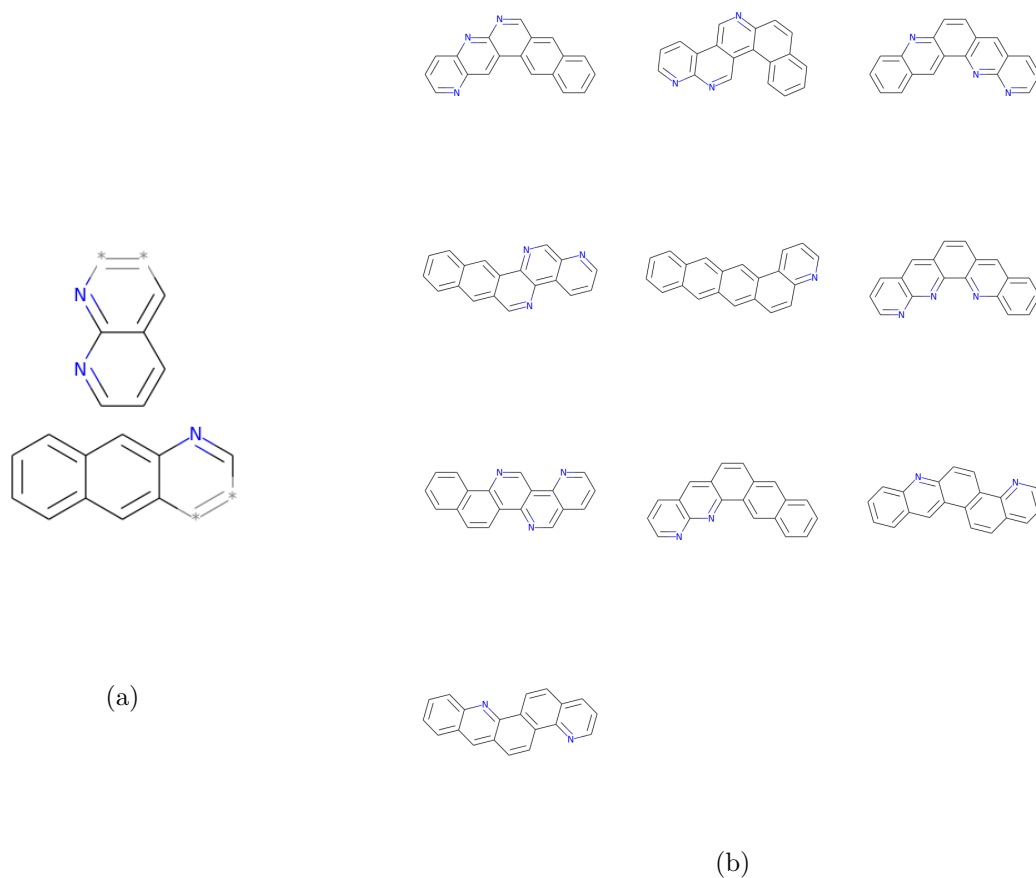


Figure 7.11: (a) shows how dummy atoms are moved to create branched molecules, (b) shows some typical branched molecules produced.

7.3 GA testing

Once the basic framework of the GA was in place, test runs to decide on the best parameters could take place. In the design stage, tournament selection was chosen as the selection method due to its ease of implementation and masking of fitness variance. The balance between crossover and elitism in the production of new generations was chosen at 90% crossover and 10% elitism. These figures are broadly the same as what other researchers have used in their GAs; the crossover percentage is perhaps higher than usual but this was chosen to encourage diversity in our populations and avoid the saturation of the GA with high fitness molecules. Thus the first parameter to be tested was the population size.

7.3.1 Population size

For our population size testing the λ_- (electron) reorganisation energy was used as the fitness function. We know that pentacene represents the global minimum for both the hole and electron reorganisation energy (as long as the linear backbone is maintained). This allowed us to compare population sizes based on how many generations they took to reach pentacene, how many unique molecules were seen before finding pentacene and how many total calculations were performed to reach it. The number of unique molecules seen before pentacene acted as a measure of the diversity of the search; in each new population it is inevitable that some molecules will match to those already seen. All children in a generation are compared to all previous generations using a SMILES string matching procedure. If a child matches a previously seen molecule it has its fitness copied from the last seen match to save running reorganisation energy calculations on identical molecules over and over. The number of unique molecules per generation will drop as the GA progresses but dropping to <10 suggests the GA may end prematurely. The total number of calculations was used as a measure of the efficiency of the GA. While we want the GA to sample as much of the space as possible, computational cost and time must be considered, especially as the GA is producing candidates for CSP, which is a time consuming endeavour in itself.

Four population sizes were chosen at first: 50; 100; 150 and 200. Calculations were performed at the B3LYP/6-31G** level of theory using Gaussian09. Each population was generated with 5 nitrogens randomly in each molecule. The atomic site crossover method was used initially with the amount of nitrogen substitution allowed to vary in successive generations. Each population size was run 10 times due to inherent randomness present in the GA and stopped once the first pentacene was created by the GA.

Fig 7.12 shows the number of generations taken within each run to hit the first occurrence of pentacene for each population size. The randomness of the GA can be seen quite clearly; the most variable population size was 50, taking between 4-10 generations to locate pentacene across the 10 runs. The large population sizes are also variable, both managed to locate pentacene within two generations on run 3 but in other runs take as many as seven. A population size of 100 is our "best behaved" with a small spread between four generations and seven.

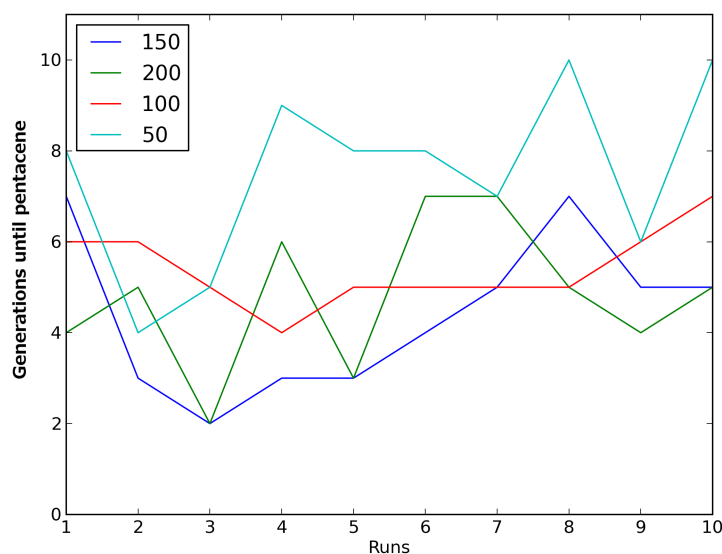


Figure 7.12: The number of generations until pentacene is hit for all population sizes and runs.

Fig 7.13 tells a similar story to Fig 7.12, showing the large variance between population sizes and the number of unique molecules seen before pentacene. Both 200 and 150 population sizes show large differences between their number of unique molecules, although 150 is perhaps smoother than 200. The number of molecules for a population size of 50 remains small irrespective of how many generations it took to reach pentacene, suggesting this population size may not be suitable for effectively sampling the substitution space. Of note is the low number of molecules for population size 100 run 8. This run took the same amount of generations to reach pentacene as the previous three runs before it but sampled 50-80 fewer molecules, highlighting the random nature of the GA even when started with the same conditions.

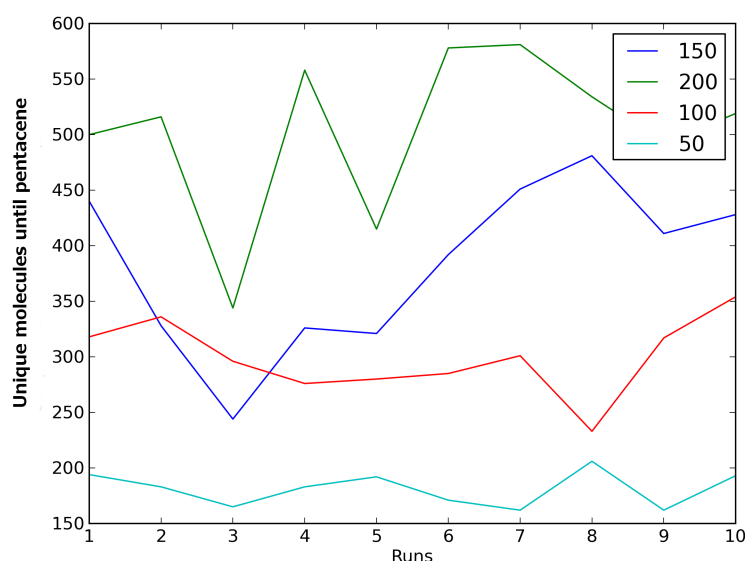


Figure 7.13: The number of unique molecules until pentacene is hit for all population sizes and runs.

While the efficiency of the different population sizes is variable, it is encouraging to see all GA runs finding pentacene quickly with large differences in the number of molecules seen. This suggests that the well surrounding pentacene on the reorganisation energy surface is deep and wide and our crossover and selection methods can rapidly aim the GA in the right direction. Examining the number of unique molecules per generation from each population size's longest run can also shed light on how the space is being explored. Table 7.1 shows these results for our four population sizes. For all population sizes, the number of unique molecules drops off rapidly. This trend is most worrying for a population size of 50. Encountering very low (<10) numbers of new molecules per generation quickly calls into question the effectiveness of the search using this size of population. The fact that the number of unique molecules decreases for all population sizes may be a result of the homogeneity of the initial population (all 5N containing azapentacenes) and this is addressed in the next chapter. For the larger population sizes it can be seen that, after a few generations, the GA is producing similar numbers of new molecules per generation. This shows that we quickly lose any benefit of the larger population sizes (>100).

The number of calculations until pentacene is reached is a measure of the computational cost of the search. For the calculation of reorganisation energy, two optimisations and two single-point energies need to be run. This can rapidly increase the time spent on calculating fitness which is by far the most expensive part of the GA (from timing information everything apart from the fitness calculation takes 1-2 seconds per generation). Table 7.2 shows the averaged results for each population size. No increase in speed is

Population size	50	100	150	200
Generation				
1	50	100	150	200
2	39	67	93	150
3	28	80	73	91
4	28	31	66	50
5	18	37	41	40
6	9	24	36	27
7	14	15	22	23
8	6	-	-	-
9	7	-	-	-
10	7	-	-	-

Table 7.1: Showing the decreasing amount of unique molecules for each population sizes longest run.

seen once a population size of above 100 is used, although the number of calculations performed rises. A population size of 200 requires almost double the amount needed for a population size of 100. Despite the number of calculations rising, the difference in the number of unique molecules is not as stark with an increase of around 40% seen between 100 and 200 versus a near 100% increase in the number of calculations. For these reasons 100 was chosen as the population size for the "production" runs of the GA, offering the best balance between speed, time spent on calculating fitness and number of molecules sampled.

Population size	Ave. GTP	Ave. MTP	Ave. CTP
50	7	181	724
100	5	299	1198
150	4	382	1528
200	5	503	2014

Table 7.2: The averaged results for each population size for atomic site crossover method. GTP is generations until pentacene, MTP unique molecules until pentacene, CTP calculations until pentacene.

With the development of the ring crossover method the tests above were repeated. Table 7.3 shows the averaged results for the ring crossover, again locating pentacene using the reorganisation energy as the fitness of each molecule. The statistics between the methods are broadly similar. All population sizes apart from 50 now take slightly longer to reach pentacene. However, the number of unique molecules sampled is also increased for each population size relative to the atomic site crossover. This is desirable as we want to sample as many molecules as possible. This causes the number of calculations required to rise but this is a small trade off for the increase in sampling. A population size of 100 was again chosen as the best balance between our needs for computational sampling and exploration of the substitution space.

Population size	Ave. GTP	Ave. MTP	Ave. CTP
50	7	220	880
100	8	394	1576
150	7	507	2028
200	6	621	2484

Table 7.3: The averaged results for each population size for the ring crossover method. GTP is generations until pentacene, MTP unique molecules until pentacene, CTP calculations until pentacene.

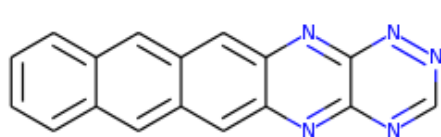
In addition to using reorganisation energy two other molecular properties of interest in the design of novel organic semiconductors were investigated as the fitness function of our molecules. While not necessarily part of any future fitness function, they provided more confidence in our testing of the GA. The HOMO/LUMO energy difference is essentially the energy difference between the conduction and valence bands of an inorganic semiconductor. Minimising this difference would result in the easier transfer of charge carriers into the LUMO and increase charge mobility. In a similar vein, the LUMO energy of the molecule as a whole was also tested as a measure of fitness. Again lowering the energy level of the LUMO should result in easier excitation of charge carriers into a conduction level. It was hoped that each molecular property would have a global minimum for the GA to reach. Calculations were performed at the same level of theory as before, with ten runs for each of the four population sizes. Both properties (LUMO energy or HOMO-LUMO difference) can be calculated with one calculation. Runs were judged to stop when all elite molecules were the same for two successive generations. Tables 7.4 and 7.5 show the averaged results for each population size. Fig 7.14 shows the minimum reached across all searches for both molecular properties.

Population size	Ave. GTM	Ave. MTM
50	4	180
100	7	407
150	8	606
200	6	596

Table 7.4: The averaged results for each population size for the ring crossover method, using HOMO/LUMO energy difference as the fitness. GTM is generations until the minimum and MTM unique molecules until the minimum.

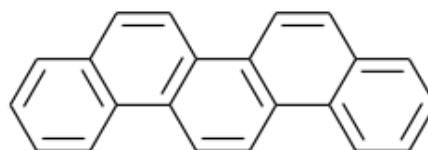
Population size	Ave. GTM	Ave. MTM
50	13	317
100	14	565
150	13	778
200	12	918

Table 7.5: The averaged results for each population size for the ring crossover method, using LUMO energy as the fitness. GTM is generations until the minimum and MTM unique molecules until the minimum.



-0.06677

(a) The HOMO/LUMO minimum



-0.04723

(b) The LUMO energy minimum

Figure 7.14: The minima located in our the searches of minimising the HOMO/LUMO energy difference and minimising the LUMO energy, values are in eV.

Both additional molecular properties show similar population statistics to the reorganisation energy. The HOMO/LUMO energy difference has a similar convergence rate to reorganisation energy with all population sizes reaching the minimum within 8 generations. Of note is that the 150 molecule population GA taking longer than the 200, which can be attributed to the randomness of the GA. Not shown in the table is the fact that population size 50 failed to locate the global minimum for both molecular properties twice, all times stopping on the 2nd ranked molecule from the other population sizes. This observation confirms our earlier point that smaller population sizes sometimes result in premature convergence. The quickness of the searches can also be attributed to the ease at which the minimum can be made by our crossover.

The LUMO energy searches took longer to locate the global minimum, which has been branched multiple times, which can take a few generations to create. This is a clear advantage of the ring crossover method compared to the atomic site crossover, which

could not have found this molecule. The final list of molecules produced by these searches are mostly branched, including almost discotic molecules higher in the ranking. The statistics from both searches confirm our earlier decision to use 100 as a population size, this being the best trade-off between speed and sampling.

7.4 Conclusions

This chapter introduced a GA for the design and evaluation of azapentacenes as novel organic semiconductors. Molecules are encoded using a SMILES string due to ease of use in crossover and readability. The initial population is randomly generated by substituting nitrogen atoms into a pentacene molecule. The number of nitrogen is fixed but this can be changed. Fitness during testing was calculated as a molecular property; reorganisation energy initially followed by the HOMO/LUMO energy difference and the energy level of the LUMO. Tournament selection is used as the selection method of molecules for crossover, due to the masking of fitness variance and easy way to modify selection pressure if necessary. Elitism is included in the GA with a ratio of 10% and 90% for crossover. Two crossover methods are included. The first (atomic site crossover) uses the 14 atoms around the outside of the ring system of pentacene as the crossover genome. An additional more chemically intuitive crossover method (ring crossover) was implemented using the individual rings of the molecule and fragmenting molecules at ring-joining points. Three crossover operators are included: single-point, two-point and uniform. Rather than use just one of these, we use all three with ratios of 30%, 60% and 10% respectively. When using the second crossover method branching can take place with a probability of 5% when single and two-point crossover operators are used. Mutation is also included, occurring with a 5% chance when a new generation has been made.

The effect of population size on the performance of our GA was investigated. Initially using atomic site crossover population sizes of 50, 100, 150 and 200 were tested using reorganisation energy as the fitness. This leads to pentacene as the known global minimum of the search. Population sizes were assessed on the number of generations until the first pentacene was found, the number of unique molecules sampled on the way and the total calculations performed. A population size of 100 offered the best compromise between computational cost and sampling ability. This test was repeated for the ring crossover method using reorganisation energy, HOMO/LUMO difference and the LUMO energy. Results from this matched to our earlier work and 100 was chosen as the population size for production runs of the GA. The next chapter will describe how the GA was used to generate candidate molecules for CSP and mobility calculations.

Chapter 8

Genetic Algorithm Production Runs

8.1 Introduction

Chapter 7 introduced our genetic algorithm (GA) for exploring the substitution space of azapentacenes. Our GA generates an initial population of randomly substituted azapentacenes and can use a range of molecular properties as the fitness. Molecules are selected for crossover using tournament selection. New generations are made through one of two crossover methods, either based on a crossover genome of the 14 atoms on the outside of the rings (atomic site crossover) or based on fragmenting the rings at 3-centre carbon atoms (ring crossover). Initial testing focused on population size and 100 was chosen as the value for our production runs.

This chapter will introduce some additional improvements to the GA made before the production runs began. A more complex fitness function is implemented to better select promising molecules for semiconducting applications. In addition a molecular fingerprinting scheme is introduced to better stop molecules being run again if they have been seen by the GA. Then the results of 5 GA runs will be presented, with results of CSP and mobility calculations on promising molecules from the GA.

8.2 Improvements to the GA

An issue that appeared early on in the testing of our GA was crossover generation molecules that had been sampled by the GA in previous generations. Removing them completely would result in a slowly dropping population size and damage the diversity of the search. Therefore, we keep previously seen molecules, but want to avoid repeating expensive fitness function calculations unnecessarily. Each new generation is checked

against all previous generations; if a molecule matches, the properties are updated and it is not sent for the calculation of fitness, but is still present for selection and crossover.

In the first builds of the GA this matching procedure worked off the SMILES strings of the molecules, checking both forwards and the reverse string (to catch any symmetry related molecules). However with the introduction of the ring crossover method and branched molecules, a more robust method of identifying identical molecules needed to be implemented. RDKit²⁷² includes functionality for the generation and similarity calculation of molecular fingerprints. The fingerprinting algorithm identifies and hashes topological paths (e.g. along bonds) in the molecule and then uses them to set bits in a fingerprint. After all paths have been identified, the fingerprint is typically folded down until a particular density of set bits is obtained. Each bit in the fingerprint represents the presence or absence of a feature in the molecule.

The fingerprint can then be used for similarity matching between molecules. The similarity measure used is the Tanimoto similarity²⁷⁴. As our fingerprints are composed of binary bits, similarity can be determined via the overlap, or intersection, of the sets. Simply put, the Tanimoto Coefficient uses the ratio of the intersecting set to the union set as the measure of similarity as below,

$$T(a, b) = \frac{N_c}{N_a + N_b - N_c} \quad (8.1)$$

where $T(a, b)$ is the similarity score, N represents the number of attributes in each object (a, b) and c the intersection set. A value of 1.0 indicates identical molecules, while anything less than that means the molecules are distinct, Table 8.1 shows the Tanimoto similarity scores for the azapentacenes studied in Chapter 6. Other metrics are available in RDKit though as a class their performance is very similar²⁷⁵.

Molecule	5A	5B	5C	7A	7B	7C
5A	1	0.374126	0.607029	0.855856	0.423333	0.574468
5B	0.374126	1	0.531157	0.407285	0.85654	0.508523
5C	0.607029	0.531157	1	0.677116	0.58046	0.92176
7A	0.855856	0.407285	0.677116	1	0.466454	0.671733
7B	0.423333	0.85654	0.58046	0.466454	1	0.578212
7C	0.574468	0.508523	0.92176	0.671733	0.578212	1

Table 8.1: A table showing the Tanimoto similarity scores for the six azapentacene molecules studied in Chapter 6.

The fingerprinting is used at multiple points in the code. In the generation of the initial population molecules, are checked against all ones currently in the population using this scheme so no duplicate molecules are allowed in the initial population. After each new generation, it is used to check for matches in previous generations, and it is used to remove duplicates when the final list of all unique molecules is created.

In addition to the implementation of molecular fingerprinting, a new fitness function was developed including the electron affinity of each population member. The electron affinity controls how efficient it is to inject charge carriers into the conduction band of an organic semiconductor from a metal electrode. For n-character organic semiconductors a lower bound of 3.0 eV has been suggested¹³¹ with an upper bound of 4.0 eV; above this the molecule will be too electrophilic to remain stable in atmospheric moisture.

Electron affinity was not measured in the initial testing runs of the GA (Chapter 6) but is easily calculated within the reorganisation energy framework as the energy difference between the neutral optimised molecule and the negatively charged optimised molecule. To analyse the electron affinity distribution of the molecules generated by our GA, a run using reorganisation energy as the fitness was performed. The ring crossover method was used, nitrogen amount was allowed to vary and the GA was judged to have finished with the first occurrence of pentacene, as in Chapter 6. The main difference was the initial population had a mixed amount of nitrogen substitution. Rather than 100 population members with 5N, the initial population was created with 20 population members with 6,7,8,9 and 10 N atoms each to better elucidate the spread of electron affinity with varying substitution. Accurate calculation of electron affinity requires larger basis sets than just reorganisation energy so calculations were performed at the B3LYP/6-311G**⁷² level of theory using Gaussian09⁷¹.

Fig 8.1 shows a plot of the electron affinity versus reorganisation energy for this run. The red lines on the graph show the lower and upper bounds as suggested in reference 131. Pentacene, as the minimum of our search, is annotated in the bottom left of our graph. A strong diagonal trend can be seen showing that as electron affinity decreases so does reorganisation energy and the number of N atoms present in the molecule. While there are many molecules with low reorganisation energies to the left of the first red line their low electron affinity suggests that they will not be feasible as n-type semiconductors.

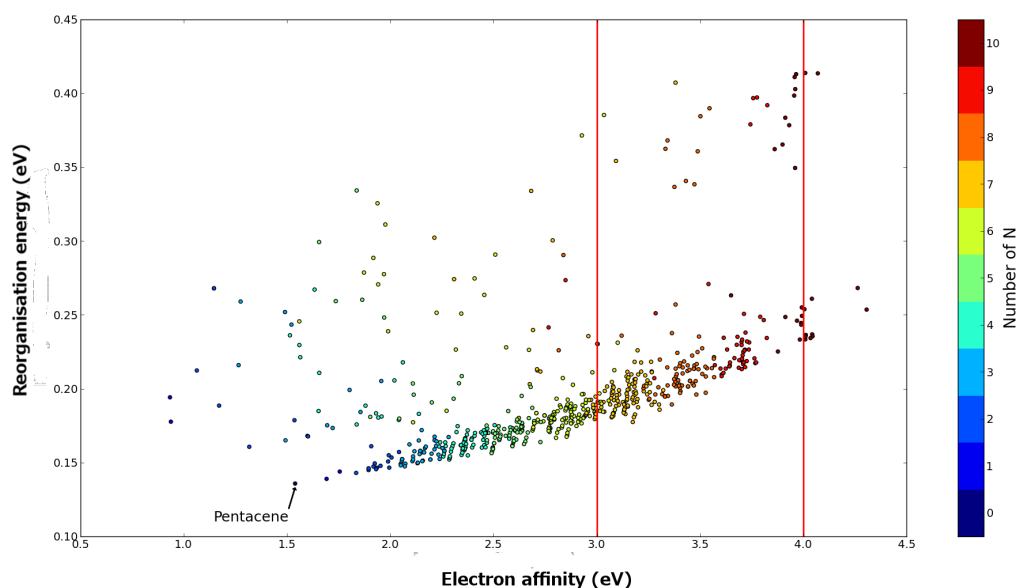


Figure 8.1: A plot of electron affinity versus reorganisation energy for all molecules encountered in a run of our GA. Points are coloured according to the amount of N atoms present in the molecule.

Most of the molecules with electron affinities within our bounds have 6-10 nitrogen atoms present, Fig 8.2 shows the best molecule (as judged by reorganisation energy) for 10 bins of 0.1 eV between 3.0 and 4.0 electron affinity. All the molecules show that as the amount of N atoms increases so does the reorganisation energy. Three of the molecules exhibit 2 N atoms on the end of the ring system, possibly a promising feature. While these molecules may be the best as chosen just by reorganisation energy, some of the substitution patterns (especially in the later bins) are infeasible for synthesis. This suggests the leading edge of the diagonal trend through our electron affinity range should be sampled. Using the electron affinity as a selection stage at the end of the GA means the GA does not take electron affinity into account when driving towards a minimum. A better approach that may make more realistic candidates is to use the electron affinity with the reorganisation energy in the calculation of the fitness of population members.

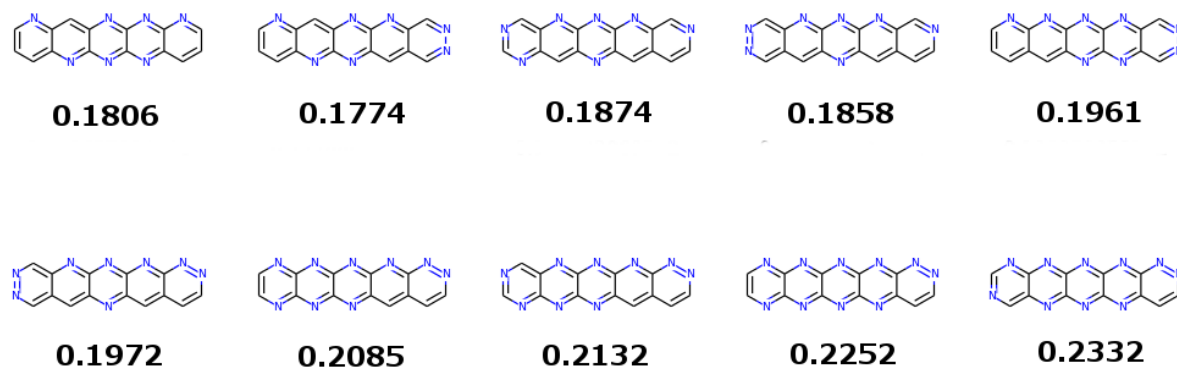


Figure 8.2: The best molecules for our search from 10 bins of 0.1 eV width across our preferred range of electron affinities (3.0 - 4.0 eV). The legend of each molecule is its reorganisation energy in eV.

To achieve this a simple step function was developed as below,

$$F = k_1\lambda + k_2Ea \quad (8.2)$$

where F is the fitness, $k_1 = 1$, λ is the reorganisation energy, Ea is the electron affinity. We set $k_2 = 0$ if $3 < Ea < 4$, $k_2 = (Ea - 4)$ if $Ea > 4$, or $k_2 = 3 - Ea$ if $Ea < 3$. Fig 8.3 shows electron affinity plotted against the new fitness for the molecules sampled in the run above. As can be seen, molecules are now penalised for having an electron affinity outside of our bounds. The large number of molecules below our lower bound is due to the run only being performed with reorganisation energy as the fitness. This function, while simple, shows good behaviour for our purpose and was used as the new fitness function for our production runs. The next section will describe these runs and their CSP and mobility calculations.

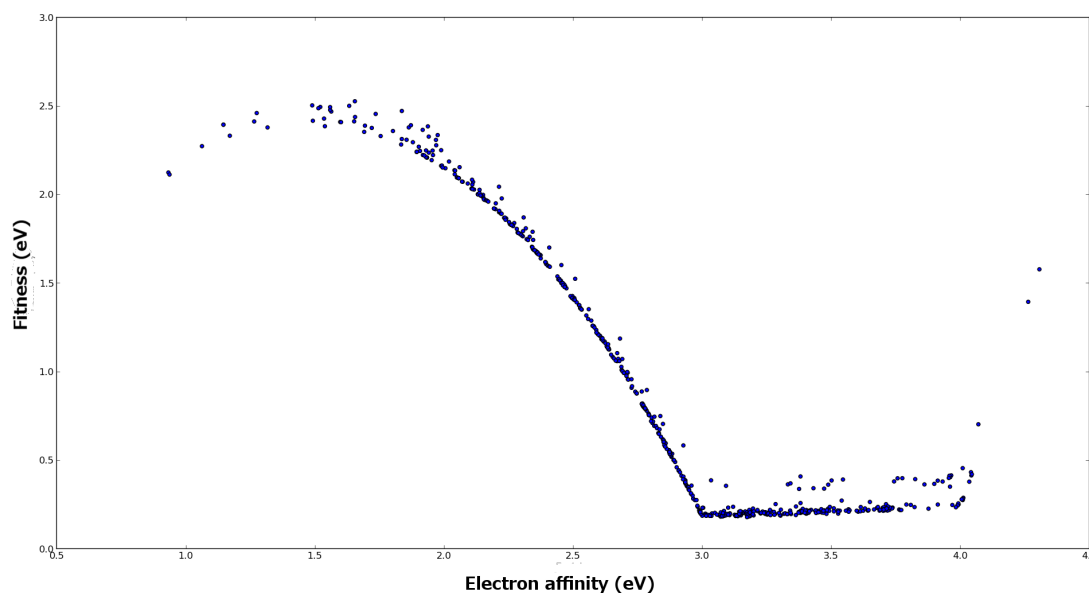


Figure 8.3: A plot showing electron affinity versus fitness (as calculated from 8.2) for the molecules sampled in our initial run.

8.3 Production runs of our GA

With a new fitness function in place the first production runs of the GA could begin. Five runs in total were performed due to the inherent randomness of GAs, with the only stopping condition set to a maximum of 150 generations. A mixed population of 100 molecules (20 with 6,7,8,9,10 N) was generated randomly each time. Fitness was calculated as described above (equation 8.2). Molecules were selected for crossover using tournament selection, with the fitter molecule having a 75% chance to win the tournament. Elitism was used with 10 molecules per generation, 90 members of the next generation being generated using the ring crossover method. Three crossover operators were used: single-point, two-point and uniform. Calculations of molecular electronic properties were performed at the B3LYP/6-311G** level of theory using Gaussian09.

8.3.1 GA results

The 10 best molecules from each GA run are shown in Figs 8.4, 8.5, 8.6, 8.7 and 8.8. The homogeneity of the top 10s are pleasing; each top 10 is broadly similar with three of the runs locating the same global minimum. Runs 2 (Fig 8.5) and 3 (Fig 8.6) both stopped on the 2nd ranked structure from the other three searches. Therefore their top 10s are slightly different. A common pattern in the best molecule from each search is the pair

of neighbouring nitrogen atoms at each end of the molecule, with two nitrogen atoms on central rings. Each top 10 can be seen as a progression towards this minimum, removing and moving nitrogen to reach the next fittest molecule. Compared to the molecules from Fig 8.2, the substitution schemes are much more sensible. The difference in molecules demonstrates the impact of including electron affinity in the fitness function of the GA.

Table 8.2 shows the number of unique molecules sampled per run. The numbers are broadly similar, but not the same, across the five runs. Even with uneven sampling our GA can locate the same minima each time.

GA Run	Number of unique molecules
1	1811
2	1974
3	1762
4	1829
5	1943

Table 8.2: The number of unique molecules sampled per run for our five runs.

Once all five runs had completed, their final lists of unique molecules were clustered together using the fingerprinting scheme described above. Once clustered, 4203 unique molecules were sampled in total, with 1127 of these being branched. Fig 8.9 shows the top 10 of this clustered list. The progression towards the features in the global minimum is more clear here. Eight of the top 10 contain a nitrogen pair on the end of the ring system, with half of them also having two nitrogen on the opposite end of the molecule. Each molecule in the top 10 contained at least six nitrogen atoms. Compared to our original six azapentacenes the substitution patterns are quite different. The GA has led to different molecules whose properties should be improved. This shows the advantage of using a GA for the exploration of substitution space, it can point in new, unexpected directions for possible synthetic targets.

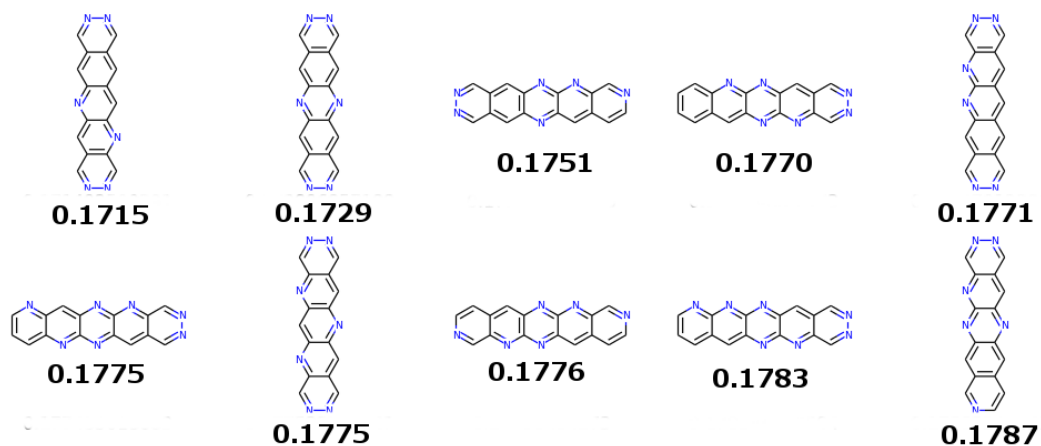


Figure 8.4: GA run 1 top 10 molecules. The number given below for each molecule is its fitness in eV (see equation 8.2).

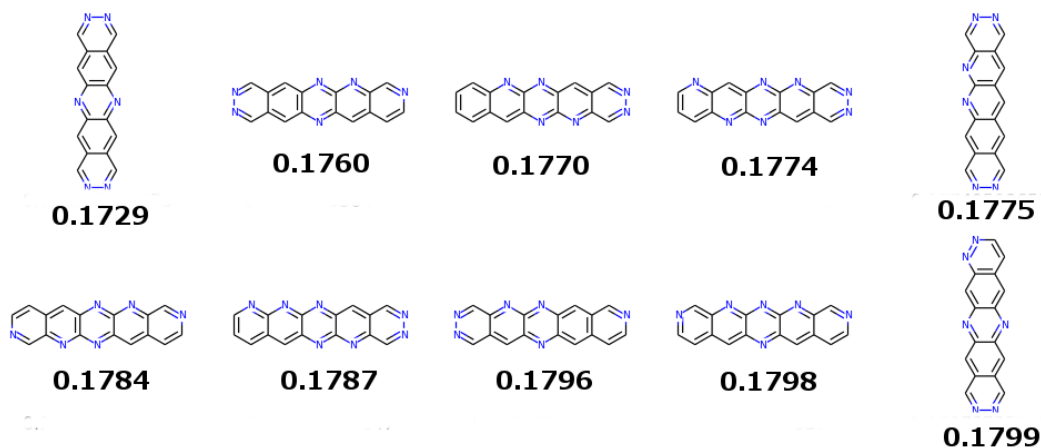


Figure 8.5: GA run 2 top 10 molecules. The number given below for each molecule is its fitness in eV (see equation 8.2).

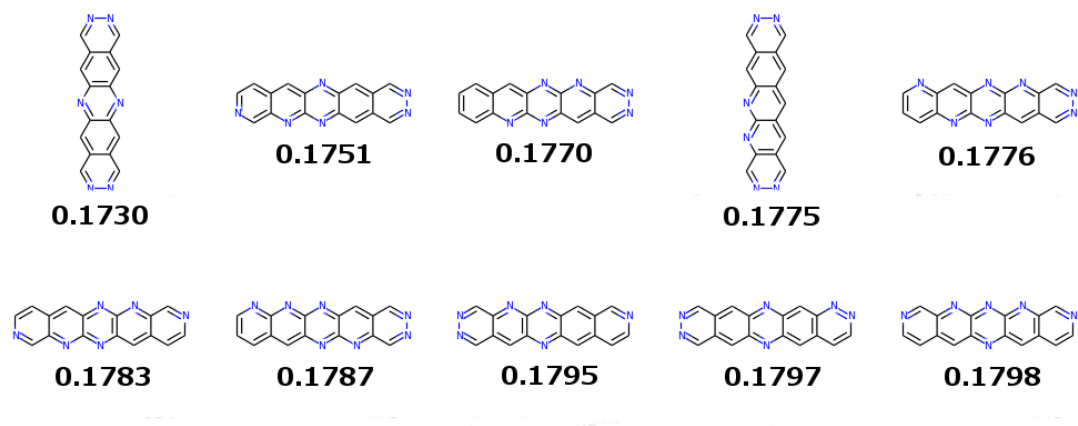


Figure 8.6: GA run 3 top 10 molecules. The number given below for each molecule is its fitness in eV (see equation 8.2).

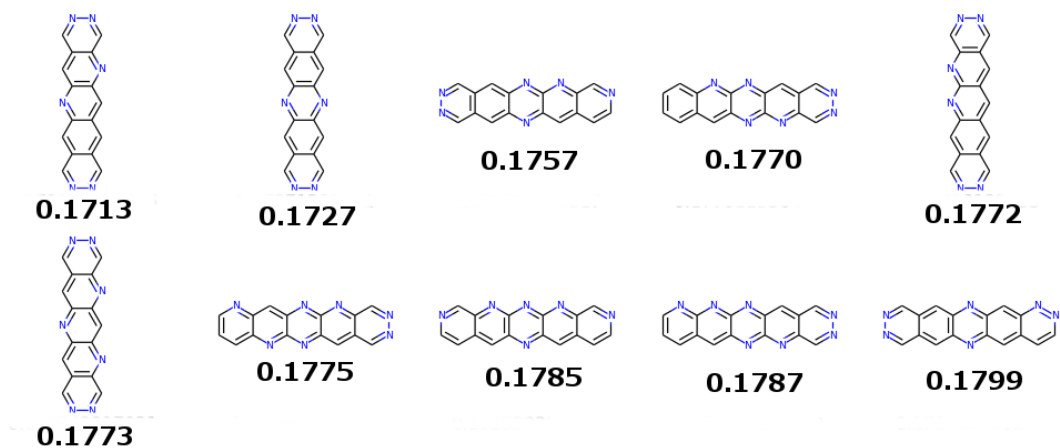


Figure 8.7: GA run 4 top 10 molecules. The number given below for each molecule is its fitness in eV (see equation 8.2).

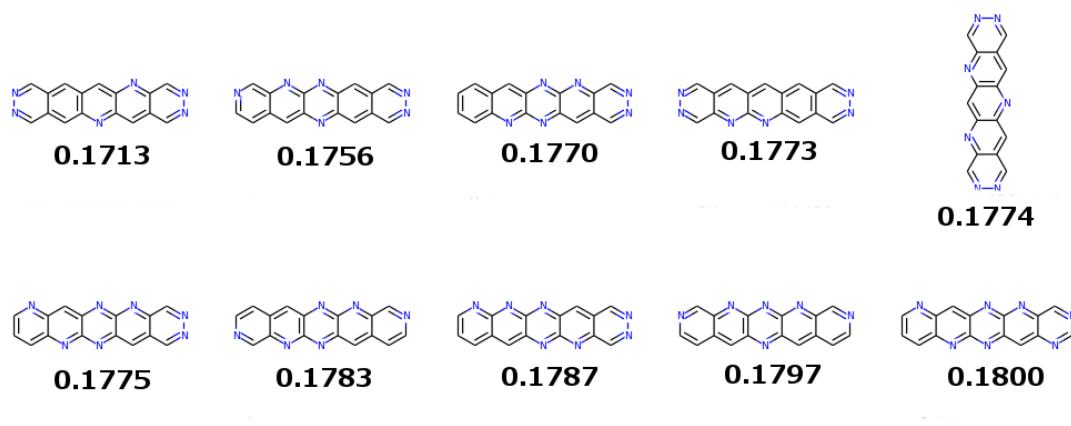


Figure 8.8: GA run 5 top 10 molecules. The number given below for each molecule is its fitness in eV (see equation 8.2).

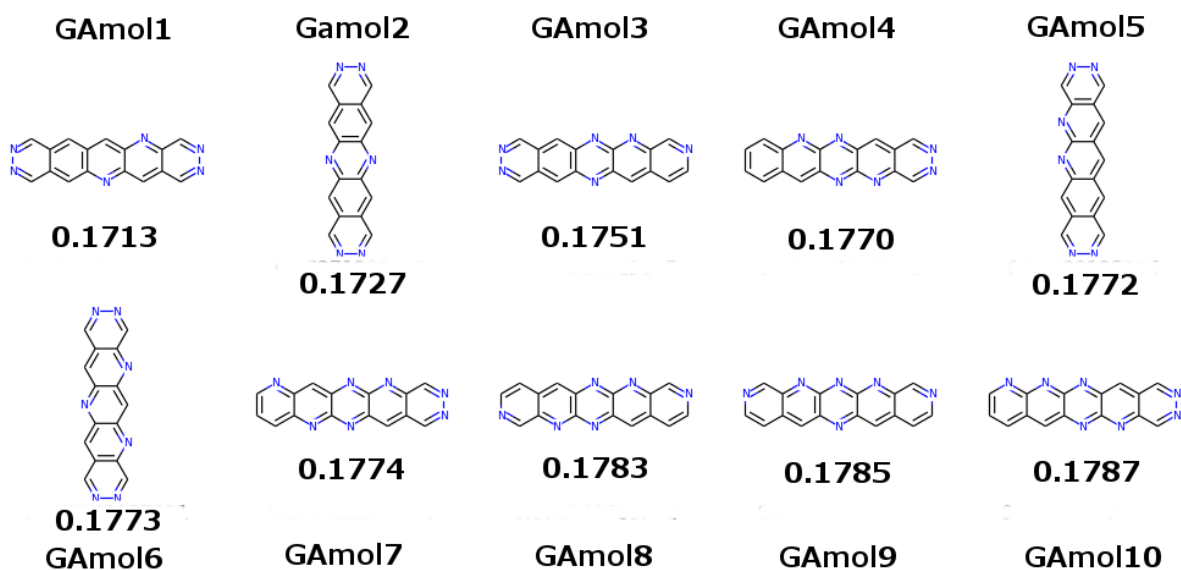


Figure 8.9: The clustered best 10 molecules from all of our GA runs. The number below each molecule is the molecules fitness in eV (equation 8.2).

Figure 8.10 shows the plot of electron affinity versus reorganisation energy for our clustered results. The plot is similar to Fig 8.1 but better sampled. The strong diagonal trend is still seen. Fig 8.11 shows the region of electron affinities from 3.0-4.0 eV in more detail, highlighting where our top 10 molecules lie on this plot. All 10 molecules are found close to our lower electron affinity bound (3.0 eV), with the global minimum lying almost exactly on the lower electron affinity boundary. Each of the top 10 is separated from the main band of molecules on the plot of reorganisation energy against electron affinity, either as the tip of a downward protrusion or grouped with other top 10

molecules. The reason for the lack of any branched molecules among the fittest molecules can be seen from Fig 8.10; many branched molecules have a low electron affinity so are penalised by our fitness function. This is shown more clearly by plotting the fitness function of only the branched molecules (Fig 8.12).

From looking at our set of fittest molecules two molecules that may be expected to be seen are missing (Fig 8.13). One, related to the best GA molecule is similar but with the nitrogen atoms separated by an unsubstituted ring, the second is related to our 5th best molecule, both nitrogen atoms are on the same side of the molecule but again separated by an empty ring. Fig 8.13 shows these molecules and their calculated electron affinities. Both were sampled by the GA but their electron affinity falls just below our lower bound, if their electron affinity was slightly higher they would have come 2nd and 4th in our final ranking. These two molecules, plus the top 10 from the GA were chosen for CSP and mobility calculations, which are presented in the next section.

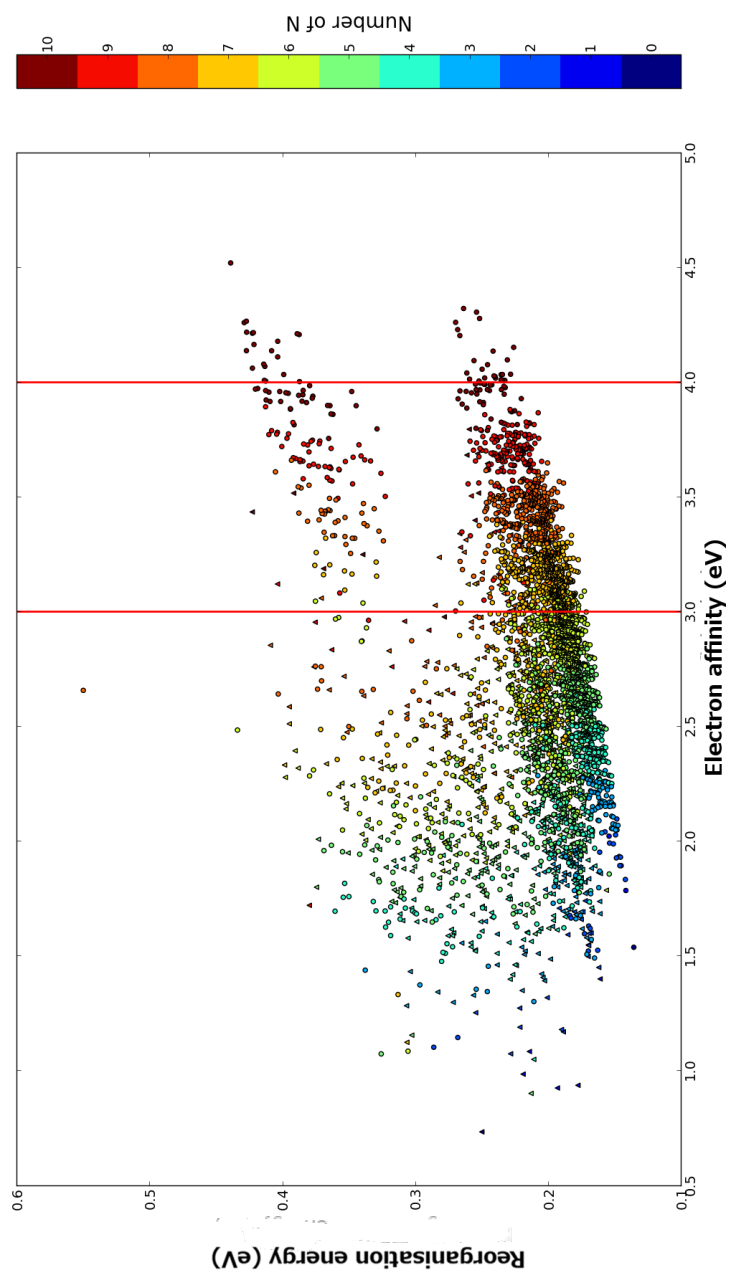


Figure 8.10: A plot showing electron affinity versus reorganisation energy for the clustered results of our 5 runs, circular points are linear molecules, triangles branched.

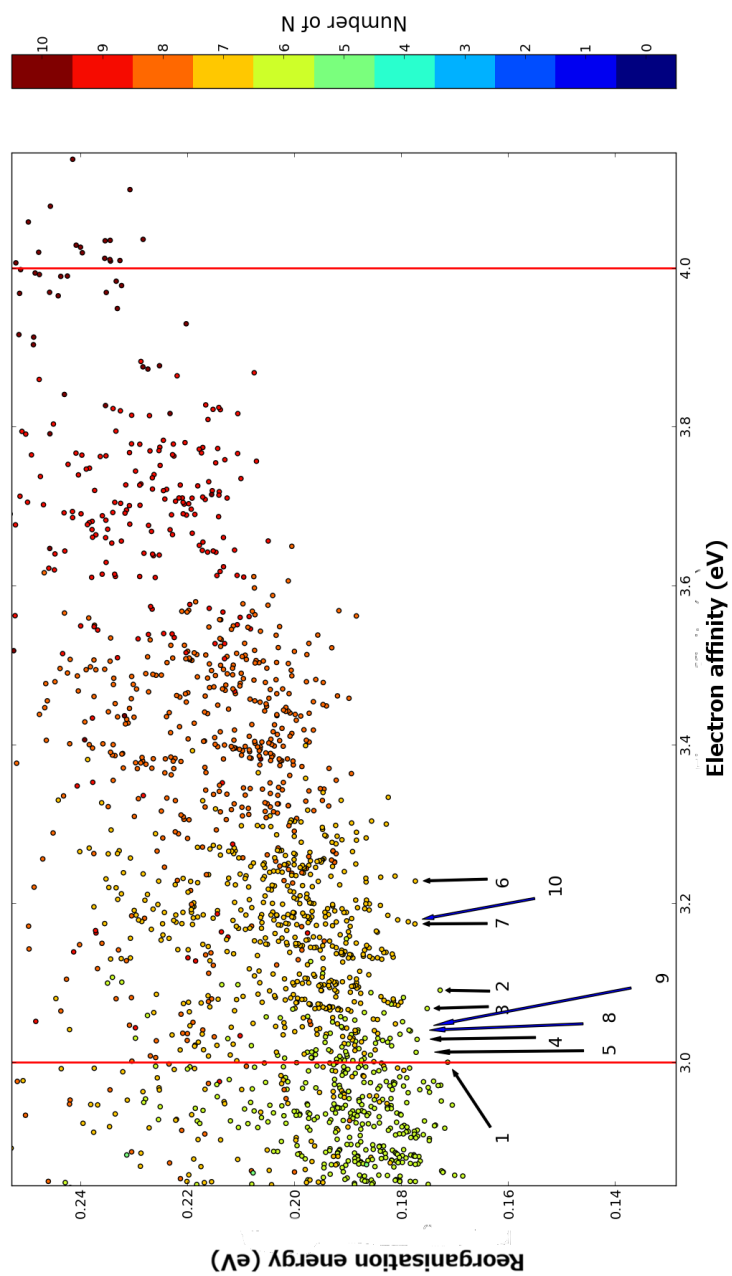


Figure 8.11: A zoomed-in version of Fig 8.10 showing where our top 10 molecules are found.

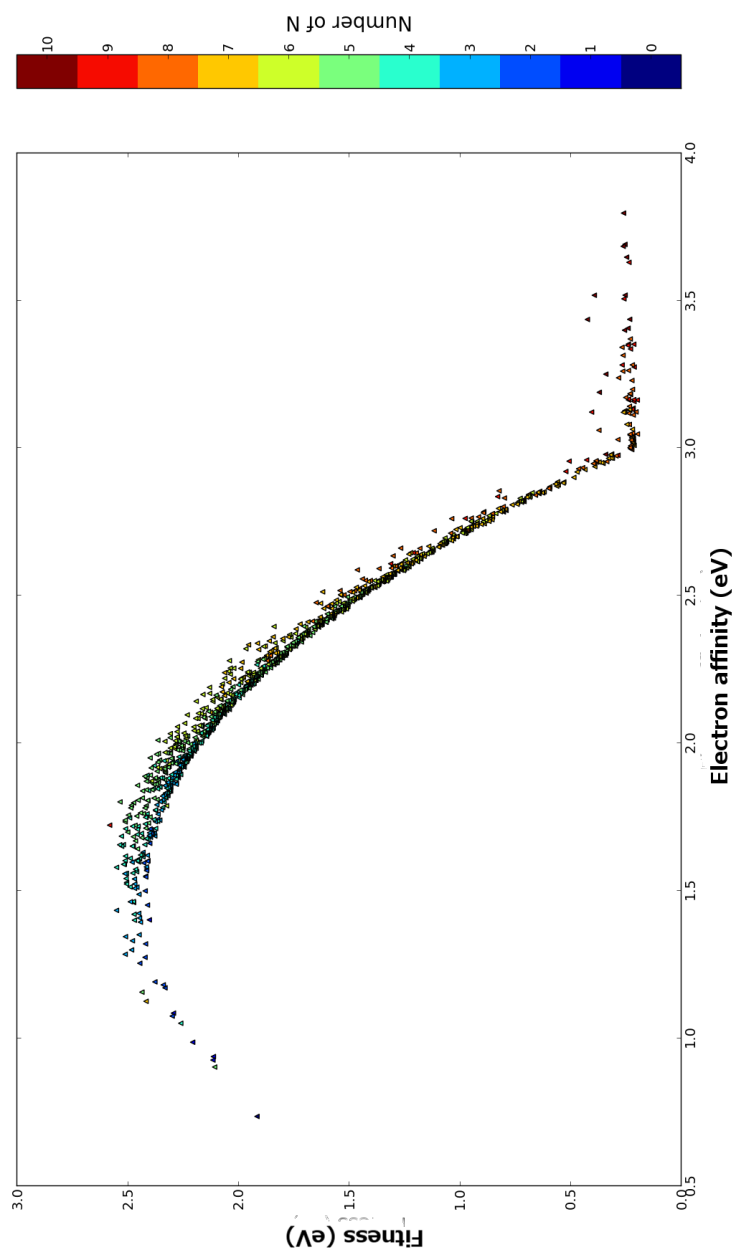


Figure 8.12: The fitness of the branched molecules sampled during the GA against their electron affinity. Most branched molecules have too low an electron affinity (<3.0 eV) and are penalised by our fitness function.

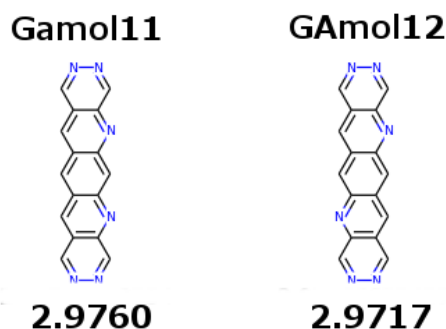


Figure 8.13: Two molecules expected to be sampled by the GA but were not in our top 10. They were located in the GA but fell just outside our electron affinity lower bound. The number under the molecule is their electron affinity in eV.

8.3.2 GA molecules CSP and mobility calculations

The CSP of our 12 molecules was performed as in Chapter 6 but the search was truncated for the quick evaluation of their potential as organic semiconductors. Eventually the GA may be able to do small CSPs like this as a part of the fitness function, but this has not yet been implemented. The geometry of each molecule was kept rigid throughout the crystal structure calculations, at the optimised gas phase structure from a B3LYP⁷²/6-311G** calculation using Gaussian 09⁷¹. Trial crystal structures were generated in 11 of the most common space groups, only considering one molecule in the asymmetric unit. Searches were performed using the Global Lattice Energy Explorer (GLEE) software, which is described in Chapter 5 and a recent paper²²³. 4000 valid crystal structures were generated in each of $P1$, $P2_1$, $P2_1/c$, $C2/c$, $P2_12_12_1$ and $Pbca$. 2000 structures were generated in each of $P\bar{1}$, $C2$, Cc , $Pca2_1$ and $Pna2_1$.

All lattice energy minimisations were performed using DMACRYS⁶⁶, using the W99⁶³? ? model potential for all intermolecular atom-atom interactions. Electrostatic interactions were described using atomic multipoles (up to hexadecapole on each atom) derived from a distributed multipole analysis of the calculated molecular electron density. Ewald summation was used for charge-charge, charge-dipole and dipole-dipole interactions, while all higher order electrostatics and repulsion-dispersion interactions were summed to a 25 Å cutoff. Lattice energy minimisation was initially performed within the space group of the generated structure. In cases where this led to a saddle point, lattice energy minimisation was continued after removing the the space group symmetry operators that allowed minimisation from the saddle point. This process led to some structures of higher Z' in the final structure sets.

Clustering was performed to identify and remove duplicate crystal structures. An initial screen was performed using the clustering method described in Chapter 5 (and reference

223) within individual space groups. An overall clustering across all space groups was then performed using COMPACT⁵⁴.

As in Chapter 6 an analysis of the packing motifs in the predicted crystal structures was performed. First, we classified all dimers formed between the molecule in the asymmetric unit and all molecules within a 20 Å distance cutoff according to the angle between their principal moments-of-inertia. Crystal structures in which all the angles are between 0–9° are classified as sheet structures (the β packing type). For structures where some dimers are not co-planar, the packing type is assigned using the four nearest neighbouring molecules. Structures in which none of the four nearest neighbours are co-planar are classed as herringbone packing. Where only one of the nearest neighbours is co-planar, the structure is classified as sandwich herringbone. Two or more co-planar neighbours indicates a stack of molecules, so these structure were classed as the γ packing type. This last category contains traditional γ structures and more sheet-like structures, where parallel sheets are tilted along the short axis of the molecule (usually 3–10°).

Mobility calculations also took the same form as Chapter 6, with the hopping rates of charge carriers calculated using Marcus Theory,

$$k_{et} = \frac{t^2}{\hbar} \sqrt{\frac{\pi}{\lambda_{\pm} k_B T}} \exp \left[-\frac{\lambda_{\pm}}{4k_B T} \right]. \quad (8.3)$$

where t is the transfer integral (a measure of the overlap of molecular wavefunctions), λ the reorganisation energy (the reaction of the molecular geometry to a charge carrier landing on the molecule), λ_{\pm} is used in this case to assess the azapentacenes are electron transporters. For the calculation of the transfer integrals between unique dimer pairs in the crystal the nearest-neighbouring molecular dimer electronic coupling matrix elements were calculated using subsystem density functional theory¹⁴⁴ at PW91²⁶⁵/DZ level of theory, as implemented in the Amsterdam Density Functional²⁶⁶ (ADF) package. The intramolecular reorganisation energies were taken from the calculations performed during the GA.

8.3.3 Results and discussion

In this section molecules will be referred to as GAmol n , where n is their rank from our fitness function, GAmol11 and GAmol12 are our additional two molecules. As seen in Chapter 6 the amount of nitrogen substitution present in our molecules has made structures with CH edge to face interactions directing packing unfavourable, with sheet and γ motifs dominating all the predicted crystal landscapes (Fig 8.13, landscapes for all other molecules can be found at the end of this chapter in Fig 8.20). In Chapter 6 the jump from 5N to 7N accomplished this; herringbone packing is completely absent from the predicted low energy structures of 7N azapentacenes. From the results of

CSP on the GA molecules (which contain 6 to 7N) we can see that 6N is sufficient to eliminate herringbone packing from the low energy structures. Once six nitrogen atoms are substituted into the pentacene molecule only sheet and γ packing is predicted. CH edge-to-face interactions are replaced with C-H \cdots N hydrogen bonds along the long axis of the molecule, which directs both motifs seen.

Some molecules show different preferences for their low energy structures, with GAmol2 having only γ structures present within 15 kJ/mol of its global minimum (Fig ??, Fig 8.16a shows the motif present in the global minimum). Unlike our previous 7N containing azapentacenes not all sheets are "true" sheet like packing, some more closely resemble flattened herringbone or γ motifs (Fig 8.16b shows this for the 2nd ranked crystal structure of GAmol1). More traditional sheet structures are seen throughout the searches however, including some approaching true brickwork packing such as the 4th ranked structure of GAmol4 (Fig 8.16c).

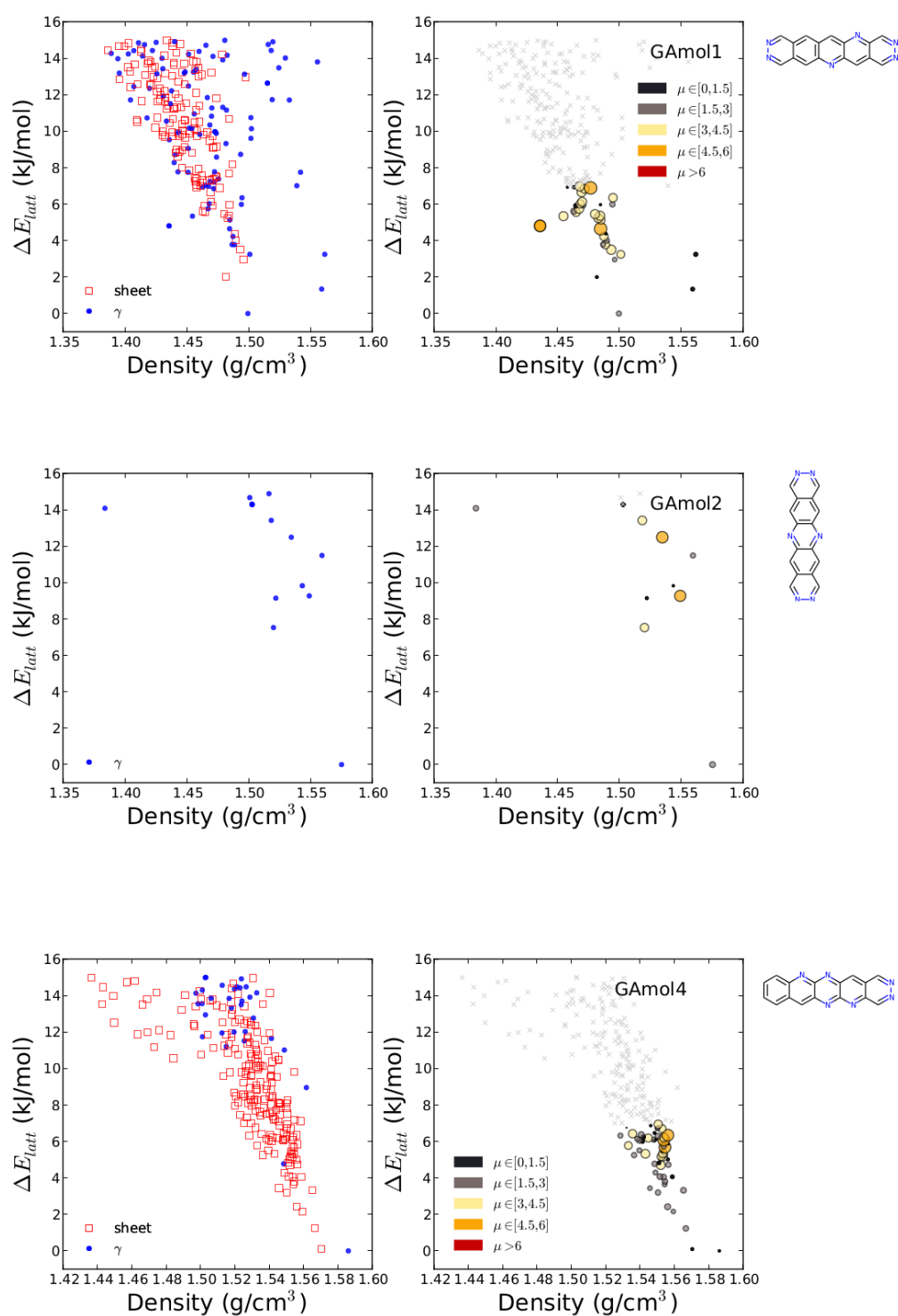
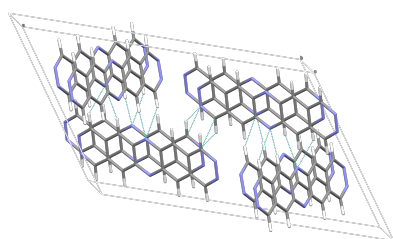
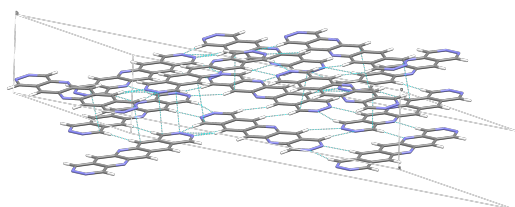


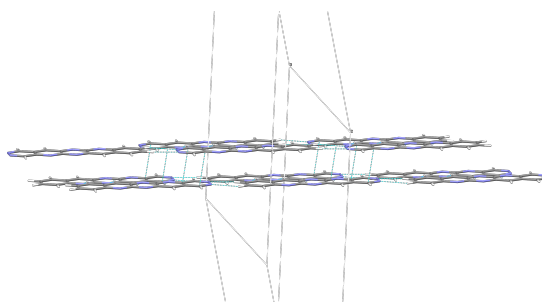
Figure 8.13: Structure–energy–mobility landscapes for the predicted crystal structures of GAMol1, GAMol2 and GAMol4, the number after GAMol refers to the rank of the molecule from the fitness function. Colouring and the size of circles on the right–hand–side correspond to the magnitudes of calculated electron mobilities.



(a) γ packing as seen in the global minimum of GAmol2.



(b) A flattened motif classified as a sheet by our classifier for the 2nd ranked structure of GAmol1.



(c) Brickwork sheet packing as seen in the 4th ranked structure of GAmol4.

Figure 8.14: Three common packing motifs seen across our predicted structures.

As before the predicted electron mobilities for crystals of each molecule were analysed through examining their averaged mobility over the predicted structures within 7 kJ/mol of their global minimum. For GAmol 2 and 8 however both lattice energy landscapes are sparse (with no structures within 7 kJ/mol of the global minimum for GAmol2). While not expected from our previous work on azapentacenes it is not infeasible for a molecule to show such large energy gaps between the global minimum and all other structures. For the purposes of investigating mobilities more predicted crystal structures were examined for these molecules, but these were not included in the final calculation of averaged mobilities. Table 8.3 summarises the gas-phase electron reorganisation energies for all 12 molecules investigated, from which it can be seen that, as expected, GAmol1 has the best molecular properties. This does not necessarily mean that it is the best

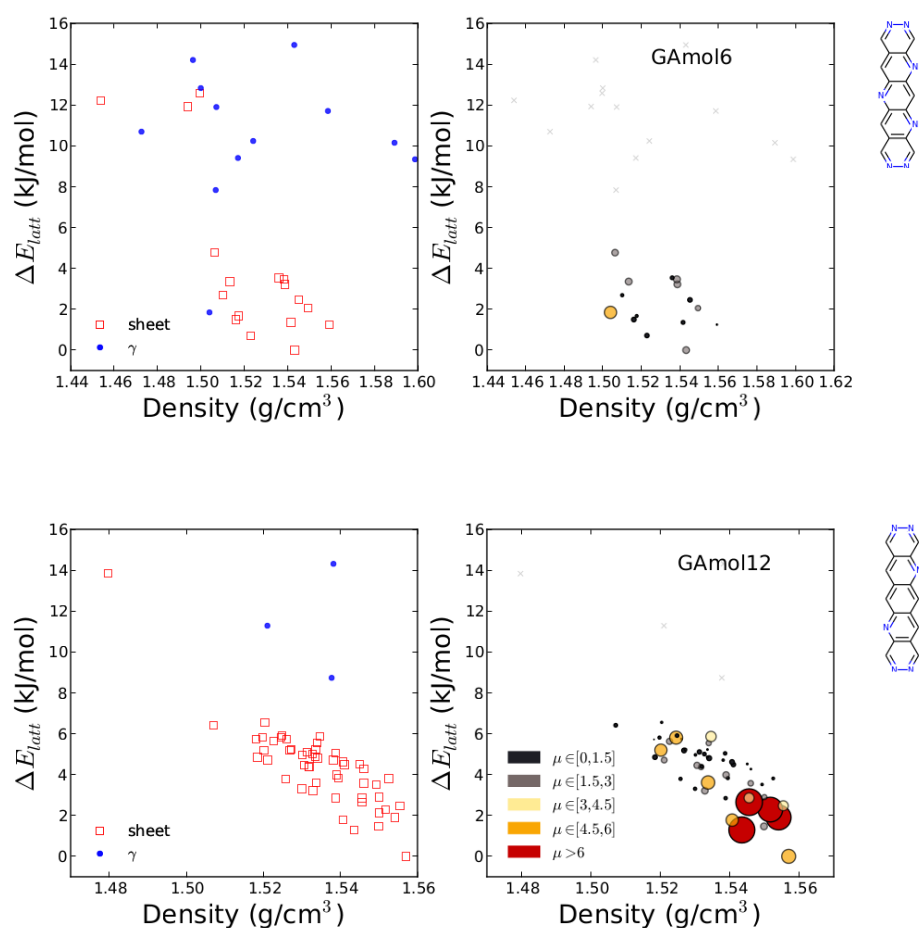
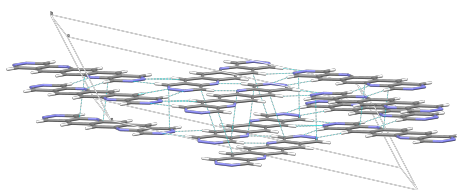


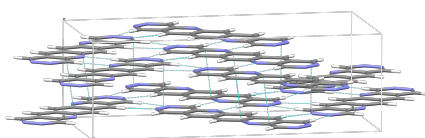
Figure 8.15: Structure–energy–mobility landscapes for the predicted crystal structures of GAMol6 and GAMol12, the number after GAMol refers to the rank of the molecule from the fitness function. Colouring and the size of circles on the right–hand–side correspond to the magnitudes of calculated electron mobilities.

molecule overall, as its highest mobility crystal structure is only 3rd overall across the 12 molecules and the averaged mobilities is also the 3rd best.

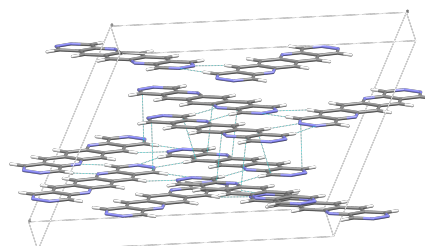
For most of our molecules the structure with the highest mobility lies at least 5 kJ/mol above the global minimum. This is expected as the highest mobility structure may have features that are unfavourable for a low lattice energy. GAMols 6 and 12 break this pattern, with GAMol12 in particular having a handful of high mobility structures close to its global lattice energy minimum (Fig 8.15). Fig 8.16 shows three of the high mobility structures predicted, all are classified as sheets but are similar to a flattened γ motif, allowing for high mobility vertically through stacked molecules. GAMol12 also has the highest averaged mobility seen in our set of molecules and using the logic from Chapter 6 is the most promising molecule for development.



(a) The global minimum of the GAmol12 CSP (mobility $5.90 \text{ cm}^2/\text{Vs}$).



(b) The 2nd ranked structure from the GAmol12 CSP (mobility $6.82 \text{ cm}^2/\text{Vs}$).



(c) The 5th ranked structure from the GAmol12 CSP (mobility $6.29 \text{ cm}^2/\text{Vs}$).

Figure 8.16: Three of the high mobility crystal structures predicted for GAmol12.

Table 8.3: Summary of the charge transport parameters for the azapentacene molecules investigated: λ_- is electron reorganisation energies, calculated at B3LYP/6-311G** level of theory. μ_{\max} is the maximum predicted electron mobility among the predicted crystal structures. $\Delta E(\mu_{\max})$ is the lattice energy gap between the crystal structure with the highest charge mobility to the predicted global minimum. $\langle\mu\rangle$ is the ensemble-averaged electron mobility across all crystals with calculated mobilities.

Molecule	λ_e (eV)	μ_{\max} (cm ² /Vs)	$\Delta E(\mu_{\max})$ (kJ/mol)	$\langle\mu\rangle$ (cm ² /Vs)
GAmol1	0.1713	5.94	6.89	2.55
GAmol2	0.1727	2.67	0.00	2.67
GAmol3	0.1751	5.15	6.16	2.06
GAmol4	0.1770	5.36	6.08	1.93
GAmol5	0.1772	8.71	5.26	2.54
GAmol6	0.1773	4.54	1.85	1.70
GAmol7	0.1774	5.02	5.88	1.99
GAmol8	0.1783	1.10	5.34	0.24
GAmol9	0.1785	3.34	5.60	1.86
GAmol10	0.1787	3.86	6.25	1.62
GAmol11	0.1758	5.85	6.98	1.44
GAmol12	0.1726	7.15	2.65	2.89

8.4 GA extensions

After completion of the above study it was decided to extend the GA to azaheptacenes. In the original code the length of the ring size is always fixed at five rings, but the change to six was largely simple. The initial SMILES string now supplied is that of unsubstituted heptacene and the range of random numbers inside the crossover functions are now larger by one. The runs proceeded the same way as before with five runs in total performed due to the inherent randomness of GAs, with the only stopping condition set to a maximum of 150 generations. A mixed population of 100 molecules (20 with 6,7,8,9,10 N) was generated randomly each time. Fitness was calculated as in equation (8.2). Molecules were selected for crossover using tournament selection, with the fitter molecule having a 75% chance to win the tournament. Elitism was used with 10 molecules per generation, 90 members of the next generation being generated using the ring crossover method. Three crossover operators were used, single-point, two-point and uniform. Calculations of molecular electronic properties were performed at the B3LYP/6-311G**⁷² level of theory using Gaussian09⁷¹.

Once the five runs had completed the results were clustered together and the top 10 unique molecules are shown in Fig 8.17. As can be seen the GA located similar minima to our azapentacene search; two nitrogens on each end of the molecule is again a favourable feature, with the top 10 molecules showing a clear progression towards this. The reorganisation energies as a whole are lower than those of the azapentacenes, which is due to the larger delocalisation offered with an extra ring system and is a known

feature of hexacenes when compared to pentacenes. This comes at a cost, however, of hexacenes and large PAHs being very unstable compared to their smaller brethren, though it is hoped the introduction of nitrogen atoms will make these longer chains more stable.

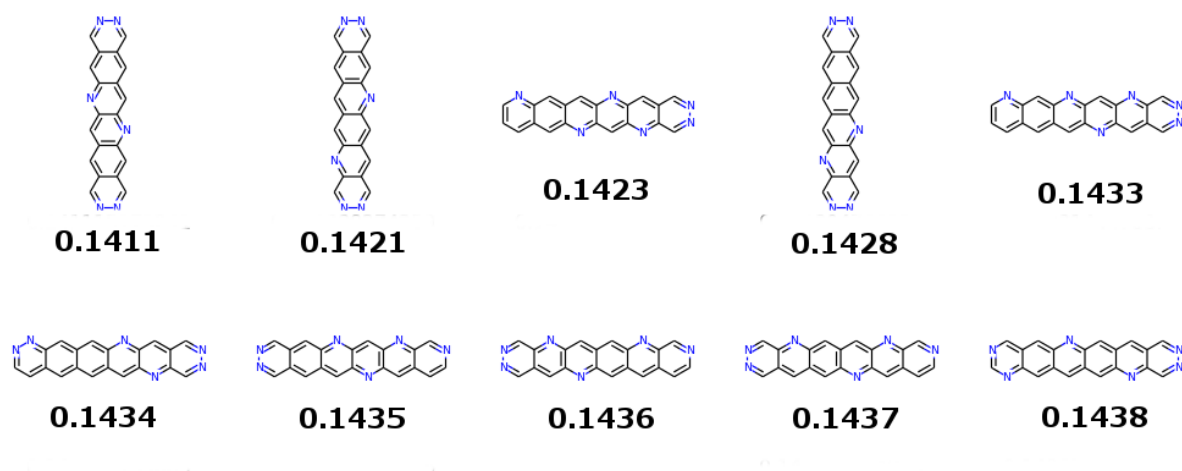


Figure 8.17: The clustered best 10 molecules from all of our hexacene runs. The number underneath is the molecules fitness in eV.

Fig 8.18 shows the plot of electron affinity versus reorganisation energy for our clustered results. More unique molecules were sampled overall, with the final list containing 7669 molecules in total, of which 1649 were branched. This reflects the increased number of possibilities compared to pentacene for both nitrogen substitution and branching from an extra ring, though the proportion of branched molecules is similar. Like for our pentacene-based search a strong diagonal trend is observed, with electron affinity and reorganisation energy falling with the number of nitrogen atoms in the molecule. Fig 8.19 shows the zoomed section of our graph with the most promising molecules highlighted. These would be promising targets for future CSP and electron mobility calculations.

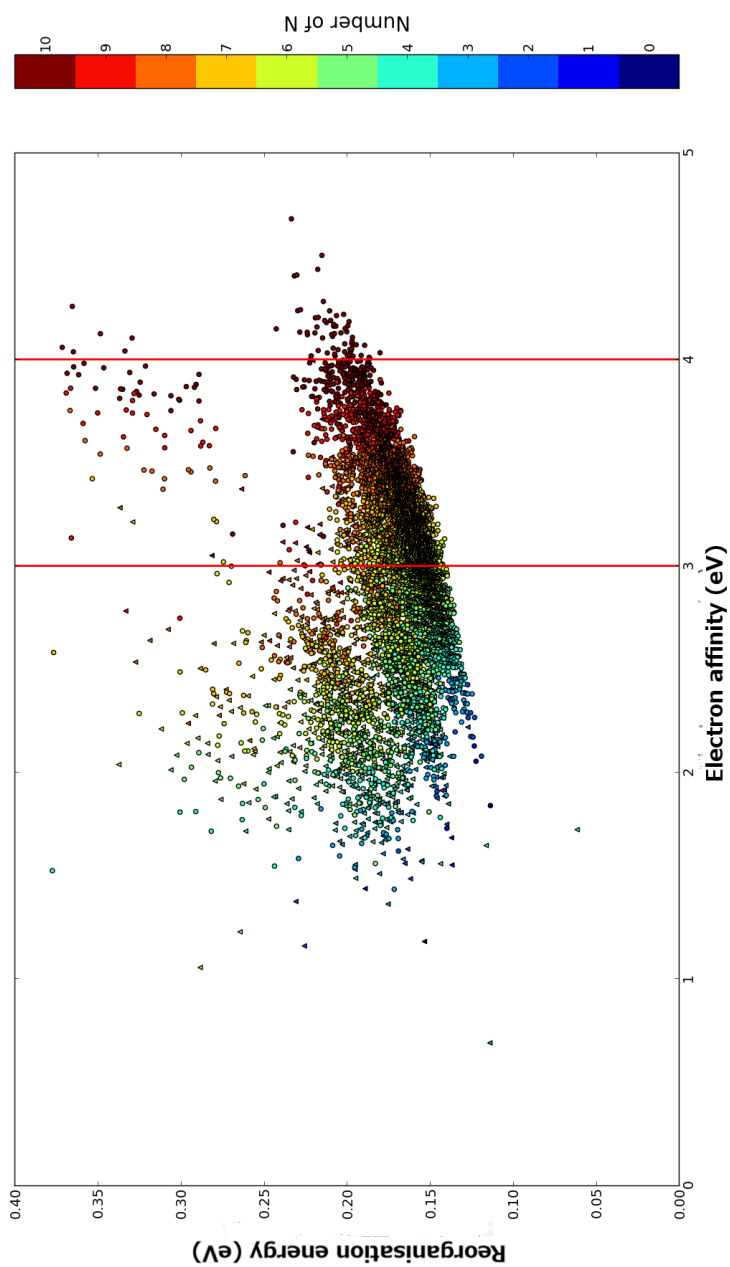


Figure 8.18: A plot showing electron affinity versus reorganisation energy for the clustered results of our five hexacene runs, circular points are linear molecules, triangles branched.

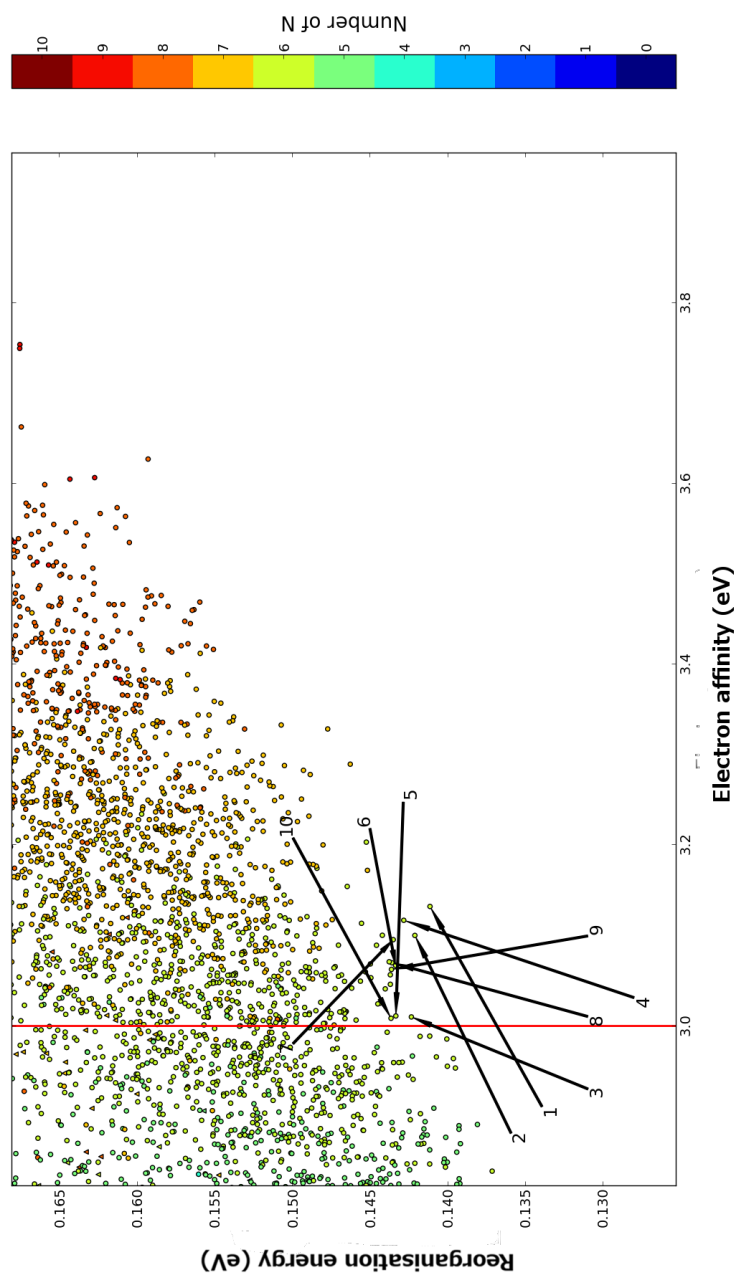


Figure 8.19: A zoomed in version of Fig 8.18 showing where our top 10 molecules are found.

8.5 Conclusions

This chapter presented changes to our GA before production runs began, the results from these production runs and an extension to the GA to be able to handle azaheptacenes. Molecular fingerprinting was introduced to better remove molecules already sampled by the GA from calculation lists using the Tanimoto similarity metric. A new fitness function was developed incorporating electron affinity into a molecules fitness. Electron affinity is an important property for n-character semiconductors as it is a measure of how easy it is to inject charge carriers into the crystal from an electrode. The new fitness function was used in five runs of the GA, starting from a mixed population. Once all five runs had completed, the final unique molecule lists were clustered and an overall unique list created. The top 10 molecules by fitness were chosen for CSP, plus an additional two with similar substitution patterns to our best molecules. A truncated CSP was performed for each molecule with matching mobility calculations. As in Chapter 6, CH edge to face interactions were disrupted and all molecules exhibited γ and sheet packing motifs throughout their low energy crystal structures. The tendency for the highest mobility structure to be far away from the global minimum was repeated. GAMol12 showed a number of high mobility structures close to its global minimum. The GA was then extended to sample azaheptacenes, initial results are similar to those of the azapentacene searches with similar substitution patterns seen in their respective top 10s. The top 10 also occupy a similar place on the electron affinity reorganisation energy landscape as the azapentacene top 10.

These results show that nitrogen substitution is a promising way to promote favourable crystal packing of pentacene-like molecules. The nitrogen substitution patterns can be optimised with respect to a fitness function using a GA to evolve a population of azapentacenes. Coupling this GA with CSP methods and electron mobility calculations gives a powerful method for proposing new molecules as n-type organic semiconductors. Here, we find GAMol12 as particularly promising due to a large number of low energy, high mobility, predicted crystal structures. In summary the design guides elucidated by this investigation are quite simple. The number of nitrogen atoms should be increased to encourage sheet and γ motifs to form, with the knowledge that too many will have a negative impact on the λ , which has been shown to have the largest impact on the overall mobility. The improved packing achieved by hydrogen bonded networks will partially offset a high λ but cannot ameliorate this entirely. The next chapter is the final conclusions of this thesis.

8.6 Additional figures

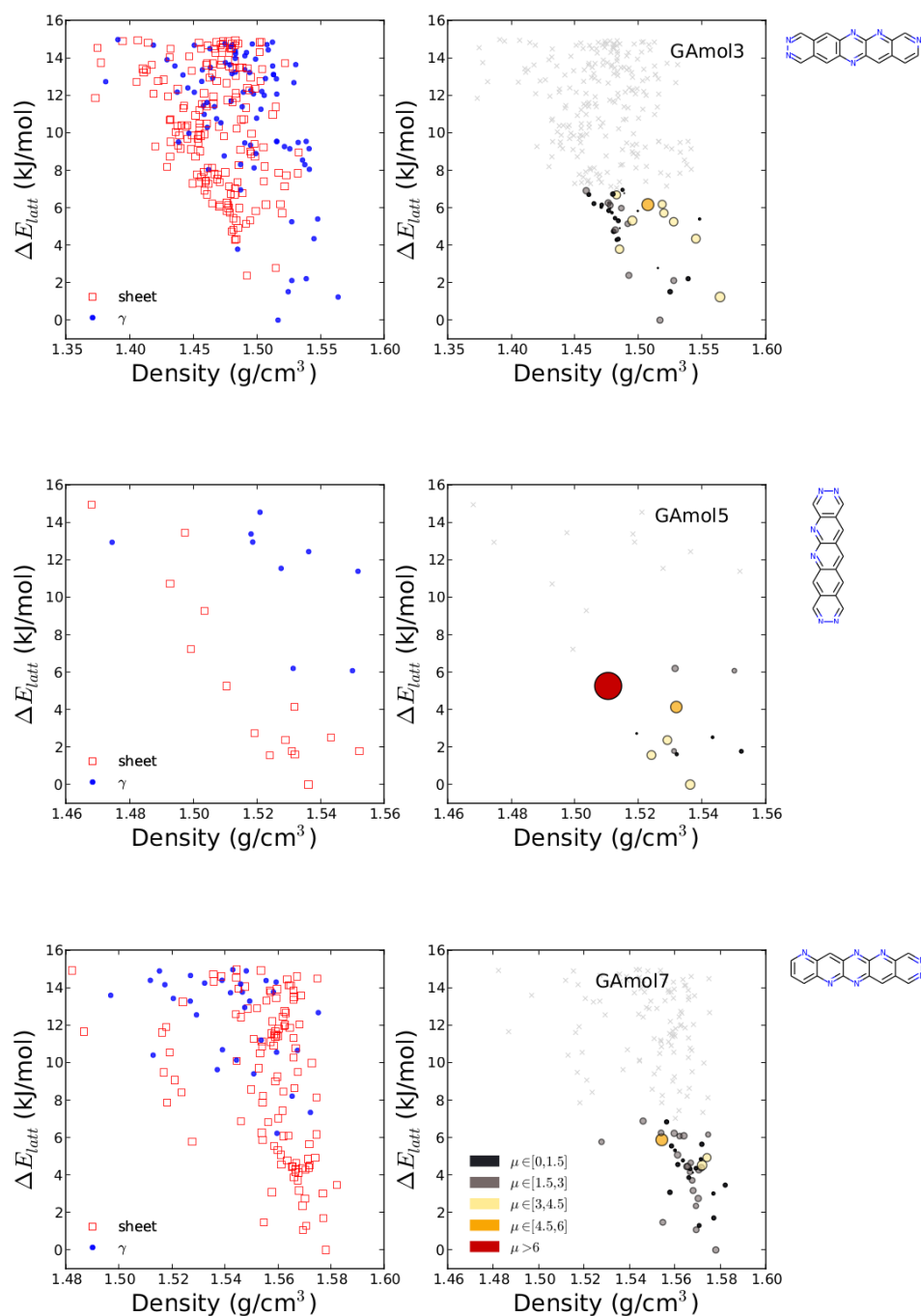


Figure 8.20: Structure–energy–mobility landscapes for the predicted crystal structures of GAmol3, GAmol5 and Gamol7, the number after GAmol refers to the rank of the molecule from the fitness function. Colouring and the size of circles on the right–hand–side correspond to the magnitudes of calculated electron mobilities.

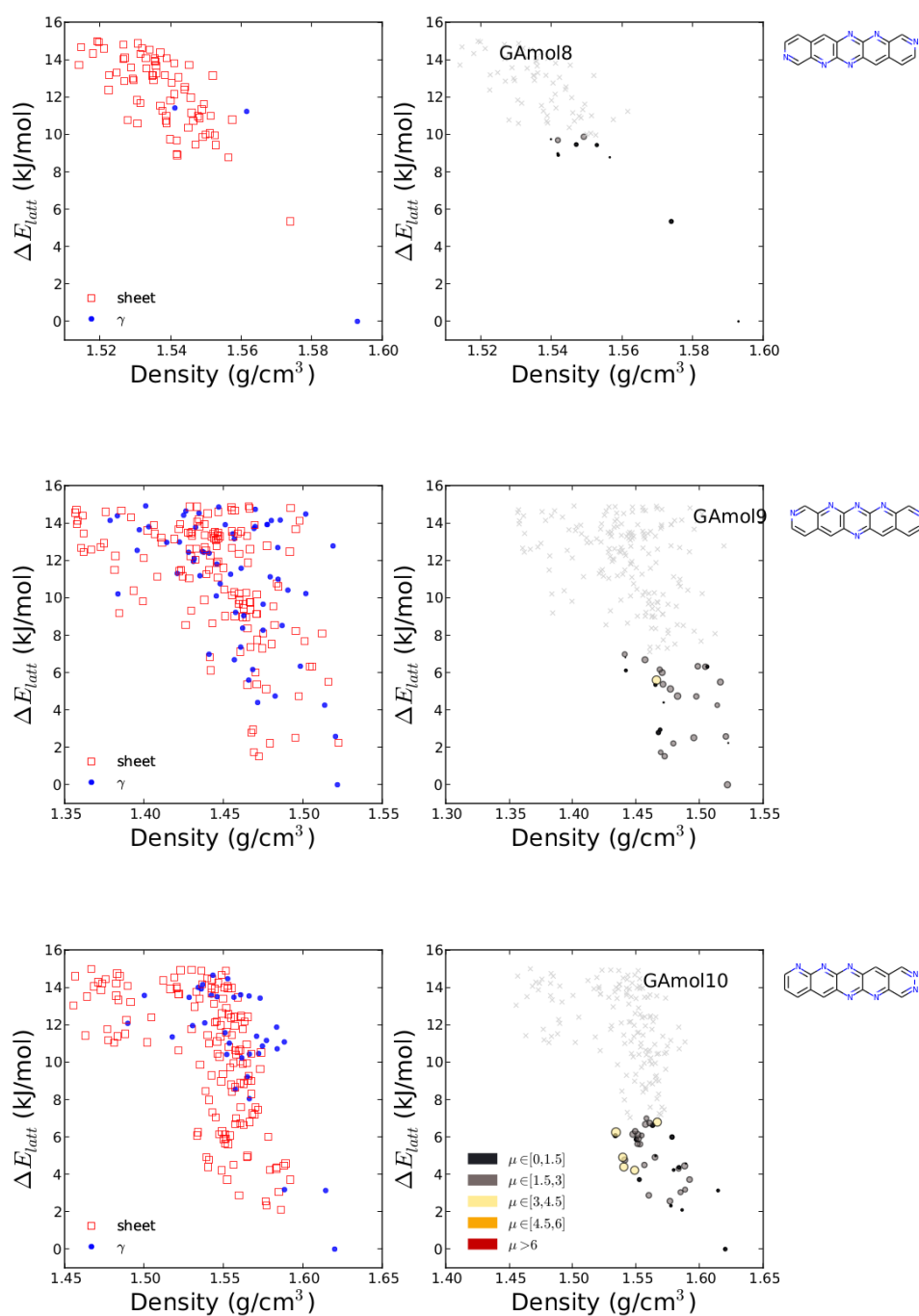


Figure 8.21: Structure–energy–mobility landscapes for the predicted crystal structures of GAmol8, GAmol9 and Gamol10, the number after GAmol refers to the rank of the molecule from the fitness function. Colouring and the size of circles on the right–hand–side correspond to the magnitudes of calculated electron mobilities.

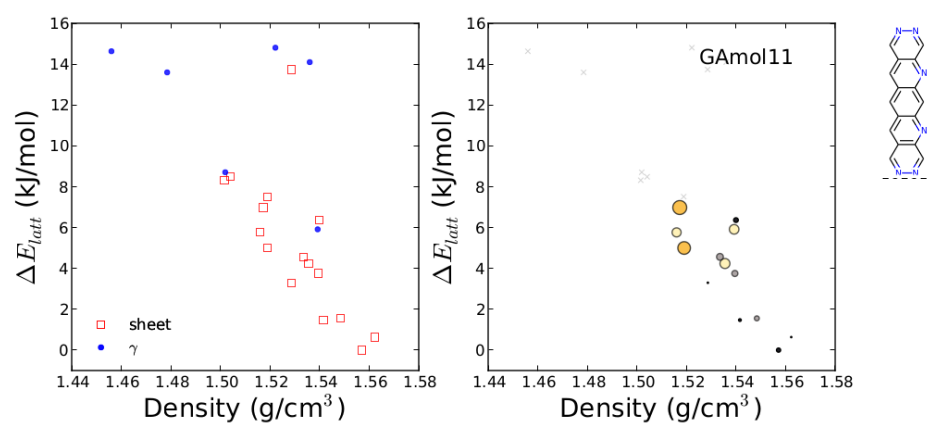


Figure 8.22: Structure–energy–mobility landscapes for the predicted crystal structures of GAMol11, the number after GAMol refers to the rank of the molecule from the fitness function. Colouring and the size of circles on the right–hand–side correspond to the magnitudes of calculated electron mobilities

Chapter 9

Conclusions

The work in this thesis has made contributions to the understanding of how substitution affects the packing and mobilities of pentacene derived semiconductors. More importantly the work presented here has explored how evolutionary searches of chemical space can be coupled to crystal structure and mobility prediction methods to propose new molecules for a target function. The chapters in this thesis explore and develop the required ingredients for this approach to the computational discovery of molecular materials.

CSP was performed for a series of hypothetical azapentacenes and their electron mobilities calculated. To better explore the substitution space of azapentacenes, a GA was developed and combined with CSP of promising molecules. The results of these CSP studies provided further information on the relationship between molecular electronic properties and crystal packing. In addition, work was undertaken on the design and implementation of a new structure generator for the generation of trial crystal structures in CSP. This was tested and a the CSP of a blind test molecule attempted as further validation of the CSP methodology.

9.1 Azapentacene CSP and GA

A major part of this thesis was the CSP of six hypothetical azapentacenes (Chapter 6). Four were taken from Winkler and Houk¹⁴² and two were designed to explore unfavourable substitution schemes. The substitution of heteroatoms into the ring system of pentacene offers an attractive way to change its transport character (from holes to electrons) and modify the crystal packing. The excellent mobility of pentacene is often attributed to its low reorganisation energy and we found that the balance between molecular electronic properties and crystal packing to be weighted towards reorganisation energy being the most important parameter. All six of our molecules presented

”better” motifs for charge transport compared to unsubstituted pentacene. CH edge-to-face interactions are disrupted and herringbone packing is rarely seen. 5N containing molecules present a range of packing motifs while 7N exhibit mostly γ and sheet.

The analysis of the charge mobility of the low energy region of the lattice energy landscape was performed. In the end, we want to rank a set of hypothetical molecules by their likelihood to yield high electron mobility crystal structures. However, choosing which parameter to rank the molecules by is not straightforward. The maximum mobility among the low energy structures can be used, but this ignores the lattice energy difference between the highest mobility structure and the global minimum. For all our molecules the highest mobility structure is at least 4 kJ/mol away from the global minimum. This is expected; good mobility requires a large overlap between molecular wavefunctions, which is penalised by the exchange-repulsion contribution to the intermolecular interaction. We chose to rank our molecules using the averaged mobilities of the structures within our energy window of interest. Even with more favourable packing, the charge transport properties of our hypothetical azapentacenes are not as good as would be expected. The reorganisation energy of all six molecules is larger than that of pentacene, as charge can not as easily delocalise over the entire molecule. For our 7N containing molecules this difference in reorganisation energy could not be overcome with good packing; all exhibited lower charge mobilities overall than their 5N containing analogues. Our best molecule overall was 5C, a molecule whose substitution pattern was expected to be unfavourable for good packing. Even with this unfavourable pattern packing motifs and lattice energies were similar to the other 5N containing molecules, highlighting the need for CSP to understand a packing landscape; it is almost impossible to do so by eye beforehand. The performance of 5C also highlights the subtle interplay between molecular properties and crystal packing and the need for both to be examined in the design of novel organic semiconductors.

The azapentacene CSP laid the foundation for the computational evaluation of organic semiconducting molecules, but the design of said molecules was still done by hand. The desire to better sample the possible chemical space of aza-substituted pentacenes led to the other major work in this thesis: the design and use of a genetic algorithm (GA) to discover promising molecules for semiconducting applications. A GA was particularly suited for this task. While we had some information on how substitution affects molecular properties it was not complete enough to design the best molecule *a priori*. GAs offer the ability to evaluate many molecules at once, and drive towards minima on the property surface of interest. While the substitution space for azapentacenes is limited searching exhaustively by hand would still take far too long.

Therefore, a GA was developed to generate promising molecular candidates for CSP and mobility calculations. Choices made during the design process (Chapter 7) included how to encode our molecules, how the population was generated, how fitness was measured, how to select population members for crossover and which crossover operators to apply.

The GA was eventually coded to use SMILES strings as the encoding method, due to readability and ease of applying crossover operators. Initial populations were generated by randomly placing a fixed amount of N atoms into the smiles string of pentacene, though later (Chapter 8) this was adapted to allow the creation of mixed populations, with differing amounts of nitrogen substitution present. In initial testing, fitness was simply the value of various molecular properties. At first we used reorganisation energy but later the HOMO/LUMO energy difference and the LUMO energy level were used. In Chapter 8 a fitness function was developed incorporating reorganisation energy and electron affinity. No large scale testing was done on which selection method to use, however the positives of tournament selection in regards to diversity and selection pressure far outweighed other methods implemented in the code.

Two crossover methods were designed and implemented. The atomic site crossover method is perhaps truer to the ideal GA, being a binary string of two values (n or c) representing atoms at certain positions around the outside of the ring system. However, this limited us to maintaining the same molecular shape and is not the most chemically intuitive way of crossover over fused ring systems. To this end the ring crossover method was developed that works by fragmenting the rings of a molecule and joining them back together, the creation of branched molecules is possible. This method was used in the production runs of our GA, it was perhaps slower than crossover method one, but sampled more molecules on the way to the same minimum in the fitness function. In addition, the true minimum of our LUMO energy testing would not have been found using atomic site crossover with it being wedded to the shape of pentacene. The number of crossover operators in the literature today is almost bewildering; for our GA the three most common seen (single-point, two-point and uniform) are used with differing probabilities. Mutation and elitism are both included, mutation allowing the GA to "jump" to another part of the search space, while elitism maintains the GAs direction, stopping the loss of fit population members.

The parameters of the GA were chosen empirically. Most of our testing was focused on population size. The runtime of the GA was an important issue as we wished to perform CSPs and mobility calculations, both time-costly endeavours, and indeed in Chapter 8 truncated CSPs were performed for our 12 molecules. Population size then is a trade off between the need for speed and the confidence in how well the space is being sampled. We tested four population sizes (50, 100, 150 and 200), using their efficiency in locating pentacene as our metric, with reorganisation energy as the fitness. Statistics between the four were broadly similar, though the smallest population size showed worrying tendencies to sample a small part of the space and in later testing on other molecular properties missed the global minimum more than once. The two largest population sizes showed no real increase in performance so a population size of 100 was chosen for our production runs. Additional parameters such as the crossover/elitism rate, the value inside tournament selection and mutation rate were chosen rather than

experimented with. From our initial population testing and the speed in which the GA located pentacene it was decided that these values sufficed.

Our production runs of the GA (Chapter 8) revealed 10 molecules with similar levels of nitrogen substitution. The substitution patterns of these molecules are quite different to those of our original six azapentacenes. This is an advantage of using a GA to evolve solutions pointing us towards patterns that may not be considered when molecules are designed by humans. Examining the molecules from ten to one the evolution of features can be seen. Two molecules with similar features were at first assumed to be missing, calling the sampling of the GA into question, but were found to have been sampled by the GA, but penalised by our fitness function. These 12 molecules were then taken into CSP and mobility calculations. While the CSPs were short, the crystal packing landscapes showed similar information to our original azapentacene CSP. The amount of nitrogen substitution present is enough to enforce γ and sheet motifs amongst the low energy crystal structures. Mobility calculations also offered similar information. Our fittest molecule using the GA fitness measure was not the best performing, again highlighting the need for CSP and electronic property calculations to be pursued together. As before, many of the highest mobility structures are located away from the global minimum, though GAmol12 offers numerous high mobility structures close to its global minimum suggesting a promising candidate for synthesis. While the short CSPs served their purpose, better understanding of the packing mobility landscape could be gained with more complete structure generation, especially for those two molecules with large gaps to their global minima. Extensions to the GA involved a similar search of azahexacenes though the resultant molecules have not yet been assessed using CSP or mobility calculations.

It would also be interesting to apply the GA to design molecules for different applications, as the overall framework is a sound one. We could apply these methods to target any property where we can assess a fitness function. For organic semiconductors there is a range of molecular properties that can be used to define the fitness function. Allowing the GA to be used solely in the generation of good candidates for CSP.

However, for other target properties the fitness may depend completely on the crystal structure, so can only be assessed from a predicted crystal packing. One approach to including CSP in the fitness function is to perform very short CSPs on population members in each generation. Shorter CSPs were used on the 12 GA molecules examined in Chapter 8, and we know from the testing of the structure generator that the global minimum (and other low energy crystal structures) are sampled early on in the Sobol' sequence. This gives us confidence that a shorter CSP can give a good picture of the low energy crystal packing landscape. The CSPs presented in Chapter 8 can be run in a day, so applying this to all population members may make the GA prohibitively expensive. A different approach is to perform CSP on only those molecules that fall within the elite section of the population. If a molecule remains within the elite the

CSP is extended, creating a fuller picture of the crystal packing landscape over time. We know from the best 10 molecules of the azapentacene and azahehexacene GAs that they often share similar features and can be viewed as a progression towards the global minimum. Using this "elite only" approach from the beginning of the GA will then give information over a wide range of local minima and limit the computational expense.

Using CSP as part of the fitness function requires a fitness that accounts for the range of possible crystal structures. In Chapters 6 and 8 we used the ensemble averaged mobilities of all structures within the energy window where likely polymorphs are found. Using this within the GA would require yet more calculations to be performed increasing computational cost. The ability to predict properties arising from the crystal structure (such as the transfer integral) without performing the full calculation would be useful. Machine learning techniques show promise in this area. In this thesis we have generated a large set of data on the crystal packing and mobility of 18 azapentacenes. This dataset could be used to train a suitable learner to predict parts of the fitness function and lower the overall cost of the GA. Incorporating these features within the GA would improve this approach to computer guided molecular design and allow the full exploration of property space from an initial molecular population.

9.2 Other CSP work

In addition to CSP on a large number of hypothetical molecules, some real life crystal systems were studied in this thesis. Chapter 6 contains the validation molecules for our CSP approach, pentacene and tetraazatetracene. Pentacene is a well studied polymorphic molecule, the polymorphs differ only slightly, however, so the location of all three in our final list of crystal structures is promising. In addition the only known crystal structure of tetraazatetracene matched to the global minimum of our search. The ranking of pentacene's single crystal form is perhaps not as good as expected, though this could be improved by periodic DFT with a dispersion correction or the development of a molecule specific forcefield.

The testing of our structure generator (Chapter 5) was not as smooth. The structure generator itself performed admirably, for all three molecules all known experimental polymorphs were located, including the $Z' = 4$ polymorph of artemisinin, which is an excellent result. However, the lattice energy ranking of the quinacridone polymorphs was poor, and showed a worrying sensitivity to the quality of the basis set used in the calculation of the atomic multipoles. Partly this can be explained with the balance between our *ab initio* electrostatics and the empirically fitted parameters of the other terms in the forcefield, but it is a strange result in the light of our successes with similar molecular interactions and shapes. Whether this can be fixed with a more specific forcefield needs investigation as the correct ranking of quinacridone polymorphs has been

predicted using an isotropic forcefield in the past²⁷⁶. The importance of the energy model was also highlighted in the CSP of the blind test molecule undertaken, while successful initially (the experimental structure matched the 3rd lowest lattice energy structure) the global minimum did not match the experimental structure until free energy was included. The improvement and refinement of energy models away from just static lattice energies is an important one and should be included wherever possible.

References

- [1] Nyman, J.; Day, G. M. *CrystEngComm* **2015**, *17*, 5154.
- [2] Price, S. L. *CrystEngComm* **2004**, *6*.
- [3] Day, G. M. *Crystallography Reviews* **2011**, *17*.
- [4] Walter, M. *chapter on Polymorphism in Physics and Chemistry of the Organic Solid State*; Wiley Interscience: New York, NY, U.S.A, 1965; Vol. 2.
- [5] Aurora, C.; Bernstein, J. *Chemical reviews* **2014**, *114*, 2170–2191.
- [6] Cruz-Cabeza, A. J.; Reutzel-Edens, S. M.; Bernstein, J. *Chem. Soc. Rev.* **2015**, *44*, 8619–8635.
- [7] Kempf, D. J.; Marsh, K. C.; Denissen, J. F.; McDonald, E.; Vasavanonda, S.; Flentge, C. A.; Green, B. E.; Fino, L.; Park, C. H.; Kong, X. P. *Proceedings of the National Academy of Sciences* **1995**, *92*, 2484–2488.
- [8] Sanjay R. Chemburkar, .; et al. *Organic Process Research & Development* **2000**, *4*, 413–417.
- [9] Lang, D.; Chi, X.; Siegrist, T.; Sergent, A.; Ramirez, A. *Physical review letters* **2004**, *93*, 86802.
- [10] Gavezzotti, A. *Accounts of Chemical Research* **1994**, *27*.
- [11] Dunitz, J. *Chemical communications (Cambridge, England)* **2003**, 545–548.
- [12] Forrest, S. *Nature* **2004**, *428*, 911–918.
- [13] Muccini, M. *Nature materials* **2006**, *5*, 605–613.
- [14] Oana, D. J.; Jacob, B.; Thomas, T. M. P. *Applied Physics Letters* **2004**, *84*.
- [15] Melanie, M. *An Introduction to Genetic Algorithms*; MIT Press: Cambring MA, 1996.
- [16] Whler; Liebig *Annalen der Pharmacie* **1832**, *3*, 249–282.
- [17] Penfold, B. R.; White, J. C. B. *Acta Crystallographica* **1959**, *12*, 130–135.

- [18] David, W. I. F.; Shankland, K.; Pulham, C. R.; Blagden, N.; Davey, R. J.; Song, M. *Angewandte Chemie* **2005**, *117*, 7194–7197.
- [19] Thun, J.; Seyfarth, L.; Senker, J.; Dinnebier, R.; Breu, J. *Angewandte Chemie International Edition* **2007**, *46*, 6729–6731.
- [20] Day, G. M.; Trask, A. V.; Motherwell, W. D. S.; Jones, W. *Chem. Commun.* **2006**, 54–56.
- [21] Thallapally, P. K.; Jetty, R. K. R.; Katz, A. K.; Carrell, H. L.; Singh, K.; Lahiri, K.; Kotha, S.; Boese, R.; Desiraju, G. R. *Angewandte Chemie International Edition* **2004**, *43*, 1149–1155.
- [22] Bauer, J.; Spanton, S.; Henry, R.; Quick, J.; Dziki, W.; Porter, W.; Morris, J. *Pharmaceutical research* **2001**, *18*, 859–866.
- [23] Chen, S.; Guzei, I.; Yu, L. *Journal of the American Chemical Society* **2005**, *127*, 9881–9885.
- [24] Yu, L. *Accounts of chemical research* **2010**, *43*, 1257–1266.
- [25] Govedarica, B.; Injac, R.; Srcic, S.; et al. *African Journal of Pharmacy and Pharmacology* **2009**, *5*, 31–41.
- [26] Karki, S.; Frii, T.; Fbin, L.; Laity, P. R.; Day, G. M.; Jones, W. *Advanced Materials* **2009**, *21*, 3905–3909.
- [27] Diao, Y.; Lenn, K. M.; Lee, W.-Y.; Blood-Forsythe, M. A.; Xu, J.; Mao, Y.; Kim, Y.; Reinspach, J. A.; Park, S.; Aspuru-Guzik, A.; Xue, G.; Clancy, P.; Bao, Z.; Mannsfeld, S. C. B. *Journal of the American Chemical Society* **2014**, *136*, 17046–17057; PMID: 25333565.
- [28] Matsukawa, T.; Yoshimura, M.; Sasai, K.; Uchiyama, M.; Yamagishi, M.; Tominari, Y.; Takahashi, Y.; Takeya, J.; Kitaoka, Y.; Mori, Y.; Sasaki, T. *Journal of Crystal Growth* **2010**, *312*, 310 – 313.
- [29] Jiang, H.; Yang, X.; Cui, Z.; Liu, Y.; Li, H.; Hu, W.; Liu, Y.; Zhu, D. *Applied Physics Letters* **2007**, *91*, 123505.
- [30] Hiszpanski, A. M.; Baur, R. M.; Kim, B.; Tremblay, N. J.; Nuckolls, C.; Woll, A. R.; Loo, Y.-L. *Journal of the American Chemical Society* **2014**, *136*, 15749–15756; PMID: 25317987.
- [31] Herbst, W.; Hunger, K.; Wilker, G.; Ohleier, H.; Winter, R. *General*; Wiley-VCH Verlag GmbH Co. KGaA, 2005; pp 1–181.
- [32] Schmidt, M. U.; Hofmann, D. W. M.; Buchsbaum, C.; Metz, H. J. *Angewandte Chemie International Edition* **2006**, *45*, 1313–1317.

- [33] Bernstein, J. *Polymorphism in Molecular Crystals*; IUCr monographs on crystallography; OUP Oxford, 2007.
- [34] Miller, G.; Garroway, A.; *A Review of the Crystal Structures of Common Explosives. Part I: RDX, HMX, TNT, PETN, and Tetryl*; Tech. Rep.; DTIC Document; 2001.
- [35] Gallagher, H. G.; Roberts, K. J.; Sherwood, J. N.; A. Smith, L. *J. Mater. Chem.* **1997**, *7*, 229–235.
- [36] Golovina, N.; Titkov, A.; Raevskii, A.; Atovmyan, L. *Journal of Solid State Chemistry* **1994**, *113*, 229 – 238.
- [37] Brock, C. P.; Dunitz, J. D. *Chem. Mater.* **1994**, *6*, 1118–1127.
- [38] Steiner, T. *Acta crystallographica. Section B, Structural science* **2000**, *56*, 673–676.
- [39] Van Eijck, B. P.; Kroon *Acta crystallographica. Section B, Structural science* **2000**, *56 (Pt 3)*, 535–542.
- [40] Alexander, V. D.; Viatcheslav, A.; Valery, A. D. *The Journal of Physical Chemistry A* **1999**, *103*.
- [41] James, R. H.; Zuyue, D.; Herman, L. A. *Journal of Computational Chemistry* **1993**, *14*.
- [42] Gavezzotti, A. *Journal of the American Chemical Society* **1991**, *113*.
- [43] Hayes, B. *American Scientist* **2011**, *99*, 282–287.
- [44] Dellavalle, R. *Organic Electronics* **2004**, *5*.
- [45] Karamertzanis, P.; Pantelides, C. *Journal of computational chemistry* **2005**, *26*, 304–324.
- [46] Sobol, I. M. *U.S.S.R. Computational Mathematics and Mathematical Physics* **1967**, *7*, 784–802.
- [47] Case, D. H.; Campbell, J. E.; Bygrave, P. J.; Day, G. M. *Journal of Chemical Theory and Computation* **2015**.
- [48] Inc, A.; *Materials Studio*; 2012; Release 6.1 ed.
- [49] Heinrich, R. K.; Frank, J. J. L.; Robert, J. G. *Journal of Computer-Aided Materials Design* **1994**, *1*.
- [50] Pillardy, J.; Arnautova, Y.; Czaplewski, C.; Gibson, K.; Scheraga, H. *Proceedings of the National Academy of Sciences of the United States of America* **2001**, *98*, 12351–12356.

- [51] Motherwell, W. D. S. *Molecular Crystals and Liquid Crystals Science and Technology. Section A. Molecular Crystals and Liquid Crystals* **2001**, 356.
- [52] Rainer, S.; Kenneth, P. *Journal of Global Optimization* **1997**, 11.
- [53] Glass, C. W.; Oganov, A. R.; Hansen, N. *Computer Physics Communications* **2006**, 175.
- [54] Chisholm, J. A.; Motherwell, S. *J. Appl. Cryst* **2005**, 38, 228–231.
- [55] Willighagen, E.; Wehrens, R.; Verwer, P.; de Gelder, R.; Buydens, L. *Acta crystallographica. Section B, Structural science* **2005**, 61, 29–36.
- [56] Dzyabchenko, A. *Acta Crystallographica Section B: Structural* **1994**.
- [57] Gelder, R. d.; Wehrens, R.; Hageman, J. A. *Journal of Computational Chemistry* **2001**, 22.
- [58] Karfunkel, H. R.; Rohde, B.; Leusen, F. J. J.; Gdanitz, R. J.; Rihs, G. *Journal of Computational Chemistry* **1993**, 14.
- [59] Stone, A. *Chemical Physics Letters* **1981**, 83, 233–239.
- [60] Richard, F. W. B. *Chemical Reviews* **1991**, 91.
- [61] J.E.Lennard-Jones *Proc. R. Soc. London* **1924**, 106, 463.
- [62] Buckingham, R. *Proceedings of the Royal Society of London. Series A. Mathematical and Physical Sciences* **1938**, 168, 264–283.
- [63] Williams, D. E. *J. Mol. Struct.* **1999**, 485-486, 321–347.
- [64] Coombes, D. S.; Price, S. L.; Willock, D. J.; Leslie, M. *The Journal of Physical Chemistry* **1996**, 100, 7352–7360.
- [65] Curt, M. B.; Kenneth, B. W. *Journal of Computational Chemistry* **1990**, 11.
- [66] Price, S. L.; Leslie, M.; Welch, G. W. A.; Habgood, M.; Price, L. S.; Karamertzanis, P. G.; Day, G. M. *Phys. Chem. Chem. Phys.* **2010**, 12, 8478–8490.
- [67] Willock, D. J.; Price, S. L.; Leslie, M.; Catlow, C. R. A. *Journal of Computational Chemistry* **1995**, 16.
- [68] Carole, O.; Sarah, L. P. *Crystal Growth & Design* **2004**, 4.
- [69] Brodersen, S.; Wilke, S.; Leusen, F. J. J.; Engel, G. *Physical Chemistry Chemical Physics* **2003**, 5, 4923–4931.
- [70] Gavezzotti, A. *CrystEngComm* **2003**, 5.
- [71] Frisch, M. J.; et al.; *Gaussian 09*; 2009.

- [72] Becke, A. D. *J. Chem. Phys.* **1993**, *98*, 5648.
- [73] Petersson, G.; Al-Laham, M. A. *The Journal of chemical physics* **1991**, *94*, 6081.
- [74] Williams, D. *Journal of molecular structure* **1999**.
- [75] Williams, D. E. *J. Comp. Chem.* **2001**, *22*, 1–20.
- [76] Williams, D. E. *J. Comp. Chem.* **2001**, *22*, 1154–1166.
- [77] Stone*, A. J. *Journal of Chemical Theory and Computation* **2005**, *1*, 1128–1132; PMID: 26631656.
- [78] Lommerse, J.; Motherwell, W.; Ammon, H.; Dunitz, J.; Gavezzotti, A.; Hofmann, D.; Leusen, F.; Mooij, W.; Price, S.; Schweizer, B.; Schmidt, M.; van Eijck, B. P.; Verwer, P.; Williams, D. *Acta crystallographica. Section B, Structural science* **2000**, *56*, 697–714.
- [79] Motherwell, W.; et al. *Acta crystallographica. Section B, Structural science* **2002**, *58*, 647–661.
- [80] Day, G.; et al. *Acta crystallographica. Section B, Structural science* **2005**, *61*, 511–527.
- [81] Day, G.; et al. *Acta crystallographica. Section B, Structural science* **2009**, *65*, 107–125.
- [82] Bardwell, D.; et al. *Acta crystallographica. Section B, Structural science* **2011**, *67*, 535–551.
- [83] Reilly, A. M.; et al..
- [84] Willock, D. J.; Price, S. L.; Leslie, M.; Catlow, C. R. A. *Journal of Computational Chemistry* **1995**, *16*, 628–647.
- [85] Day, G. M.; Chisholm, J.; Shan, N.; Motherwell, W. D. S.; Jones, W. *Crystal Growth & Design* **2004**, *4*.
- [86] Day, G. M.; Motherwell, W. D. S.; Jones, W. *Crystal Growth & Design* **2005**, *5*.
- [87] Mayo, S. L.; Olafson, B. D.; Goddard, W. A. *The Journal of Physical Chemistry* **1990**, *94*.
- [88] Hwang, M. J.; Stockfisch, T. P.; Hagler, A. T. *Journal of the American Chemical Society* **1994**, *116*.
- [89] Sun, H. *The Journal of Physical Chemistry B* **1998**, *102*.
- [90] Gavezzotti, A. *Modelling Simul. Mater. Sci. Eng* **2002**.

- [91] Brodersen, S.; Wilke, S.; Leusen, F. J. J.; Engel, G. *Physical Chemistry Chemical Physics* **2003**, *5*.
- [92] Karamertzanis, P. G.; Pantelides, C. C. *Molecular Physics* **2007**, *105*, 273–291 pages = 273–291.
- [93] Mooij, W. T. M.; van Eijck, B. P.; Kroon, J. *Journal of the American Chemical Society* **2000**, *122*, 3500–3505.
- [94] van Eijck, B. P.; Mooij, W. T. M.; Kroon, J. *Journal of Computational Chemistry* **2001**, *22*, 805–815.
- [95] van Eijck, B. P.; Mooij, W. T. M.; Kroon, J. *The Journal of Physical Chemistry B* **2001**, *105*, 10573–10578.
- [96] Karamertzanis, P. G.; Price, S. L. *Journal of Chemical Theory and Computation* **2006**, *2*, 1184–1199.
- [97] A. V. Kazantsev, C. C. P., Panagiotis G. Karamertzanis; Adjiman, C. S. *Molecular Systems Engineering, Volume 6, chapter 1*; Wiley-VCH, 2010.
- [98] Lancaster, R.; Karamertzanis, P.; Hulme, A.; Tocher, D.; Covey, D.; Price, S. *Chemical communications* **2006**, 4921–4923.
- [99] Polito, M.; D’Oria, E.; Maini, L.; Karamertzanis, P. G.; Grepioni, F.; Braga, D.; Price, S. L. *CrystEngComm* **2008**, *10*.
- [100] Ouvrard, C.; Price, S. L. *Crystal Growth & Design* **2004**, *4*.
- [101] Nowell, H.; Frampton, C.; Waite, J.; Price, S. *Acta crystallographica. Section B, Structural science* **2006**, *62*, 642–650.
- [102] Day, G. M.; Cooper, T. G. *CrystEngComm* **2010**, *12*.
- [103] Bruno, I.; Cole, J.; Kessler, M.; Luo, J.; Motherwell, W.; Purkis, L.; Smith, B.; Taylor, R.; Cooper, R.; Harris, S.; Orpen, A. *Journal of chemical information and computer sciences* **2004**, *44*, 2133–2144.
- [104] Bruno, I.; Cole, J.; Edgington, P.; Kessler, M.; Macrae, C.; Patrick, M.; Pearson, J.; Taylor, R. *Acta crystallographica. Section B, Structural science* **2002**, *58*, 389–397.
- [105] Brameld, K.; Kuhn, B.; Reuter, D.; Stahl, M. *Journal of chemical information and modeling* **2008**, *48*, 1–24.
- [106] Kazantsev, A.; Karamertzanis, P.; Adjiman, C.; Pantelides, C.; Price, S.; Galek, P.; Day, G.; Aurora, C. *International journal of pharmaceuticals* **2011**, *418*, 168–178.

- [107] Chisholm, J. A.; Motherwell, S.; Tulip, P. R.; Parsons, S.; Clark, S. J. *Crystal Growth & Design* **2005**, *5*.
- [108] Byrd, E. F. C.; Scuseria, G. E.; Chabalowski, C. F. *The Journal of Physical Chemistry B* **2004**, *108*.
- [109] Neumann, M.; Perrin, M. *The journal of physical chemistry. B* **2005**, *109*, 15531–15541.
- [110] Neumann, M. *The journal of physical chemistry. B* **2008**, *112*, 9810–9829.
- [111] Wallace, P. R. *Phys. Rev.* **1947**, *71*, 622–634.
- [112] Inokuchi, H. *Bulletin of the Chemical Society of Japan* **1951**, *24*, 222–226.
- [113] Akamatu, H.; Inokuchi, H.; Matsunaga, Y. *Bulletin of the Chemical Society of Japan* **1956**, *29*, 213–218.
- [114] Kallmann, H.; Pope, M. *The Journal of Chemical Physics* **1960**, *32*.
- [115] Kallmann, H.; Pope, M. *Nature* **1960**, *186*, 31–33; 10.1038/186031a0.
- [116] Pope, M.; Kallmann, H. P.; Magnante, P. *The Journal of Chemical Physics* **1963**, *38*.
- [117] Koezuka, H.; Tsumura, A.; Ando, T. *Synthetic Metals* **1987**, *18*, 699 – 704.
- [118] Ghosh, A. K.; Morel, D. L.; Feng, T.; Shaw, R. F.; Rowe, C. A. *Journal of Applied Physics* **1974**, *45*.
- [119] Podzorov, V. *MRS Bulletin* **2013**, *38*.
- [120] Podzorov, V.; Menard, E.; Borissov, A.; Kiryukhin, V.; Rogers, J.; Gershenson, M. *Physical review letters* **2004**, *93*.
- [121] Jurchescu, O. D.; Baas, J.; Palstra, T. T. M. *Applied Physics Letters* **2004**, *84*.
- [122] Nielsen, C.; Turbiez, M.; Iain, M. *Advanced materials Fla.)* **2013**, *25*, 1859–1880.
- [123] Molinari, A.; Alves, H.; Chen, Z.; Facchetti, A.; Morpurgo, A. *Journal of the American Chemical Society* **2009**, *131*, 2462–2463.
- [124] Islam, M.; Pola, S.; Tao, Y. *Chemical communications England)* **2011**, *47*, 6356–6358.
- [125] Podzorov, V.; Menard, E.; Borissov, A.; Kiryukhin, V.; Rogers, J.; Gershenson, M. *Physical review letters* **2004**, *93*.
- [126] Podzorov, V.; Menard, E.; Rogers, J.; Gershenson, M. *Physical review letters* **2005**, *95*.

- [127] Minder, N.; Ono, S.; Chen, Z.; Facchetti, A.; Morpurgo, A. *Advanced materials Fla.* **2012**, *24*, 503–508.
- [128] Pope, M.; Swenberg, C. E. *Electronic Processes in Organic Crystals and Polymers*, 2nd ed.; Oxford University Press, 1982.
- [129] Marcus, R. A. *The Journal of Chemical Physics* **1956**, *24*.
- [130] Norton, J.; Brédas, J. *Journal of the American Chemical Society* **2008**, *130*, 12377–12384.
- [131] Brédas, J.; Calbert, J.; da Silva Filho, D.; Cornil, J. *Proceedings of the National Academy of Sciences of the United States of America* **2002**, *99*, 5804–5809.
- [132] Valeev, E.; Coropceanu, V.; da Silva Filho, D.; Salman, S.; Brédas, J. *Journal of the American Chemical Society* **2006**, *128*, 9882–9886.
- [133] Troisi, A. *Chem. Soc. Rev.* **2011**, *40*, 2347–2358.
- [134] Desiraju, G. R.; Gavezzotti, A. *Acta Crystallographica Section B: Structural Science* **1989**, *45*, 473–482.
- [135] Coropceanu, V.; Cornil, J.; da Silva Filho, D.; Olivier, Y.; Silbey, R.; Brdas, J.-L. *Chemical reviews* **2007**, *107*, 926–952.
- [136] Wang, C.; Dong, H.; Hu, W.; Liu, Y.; Zhu, D. *Chemical reviews* **2012**, *112*, 2208–2267.
- [137] Newman, C.; Frisbie, C. D.; da Silva Filho, D. A.; Brédas, J.-L.; Ewbank, P. C.; ; Mann, K. R. *Chem. Mater.* **2004**, *16*, 4436–4451.
- [138] Stolar, M.; Baumgartner, T. *Physical chemistry chemical physics : PCCP* **2013**, *15*, 9007–9024.
- [139] Sakamoto, Y.; Suzuki, T.; Kobayashi, M.; Gao, Y.; Fukai, Y.; Inoue, Y.; Sato, F.; Tokito, S. *Journal of the American Chemical Society* **2004**, *126*, 8138–8140.
- [140] Chen, H.; Chao, I. *Chemical Physics Letters* **2005**, *401*.
- [141] Sakamoto, Y.; Suzuki, T.; Kobayashi, M.; Gao, Y.; Inoue, Y.; Tokito, S. *Molecular Crystals and Liquid Crystals* **2006**, *444*.
- [142] Sheraw, C. D.; Jackson, T. N.; Eaton, D. L.; Anthony, J. E. *Advanced Materials* **2003**, *15*.
- [143] Anthony, J. *Chemical reviews* **2006**, *106*, 5028–5048.
- [144] Anthony, J.; Eaton, D.; Parkin, S. *Organic letters* **2002**, *4*, 15–18.
- [145] Bunz, U. H. F. *Chemistry* **2009**, *15*, 6780–6789.

- [146] Bunz, U. H. F.; Engelhart, J. U.; Lindner, B. D.; Schaffroth, M. *Angew. Chem. Int. Ed.* **2013**, *52*, 3810–3821.
- [147] Chen, H.; Chao, I. *ChemPhysChem* **2006**, *7*, 2003–2007.
- [148] Winkler, M.; Houk, K. *J. Amer. Chem. Soc* **2007**, *129*, 1805–1815.
- [149] Chen, X.-K.; Guo, J.-F.; Zou, L.-Y.; Ren, A.-M.; Fan, J.-X. *J Phys. Chem. C* **2011**, *115*, 21416–21428.
- [150] Pavanello, M.; Neugebauer, J. *J. Chem. phys.* **2011**, *135*, 234103.
- [151] te Velde, G.; Bickelhaupt, F. M.; Baerends, E. J.; Fonseca Guerra, C.; van Gisbergen, S. J.; Snijders, J. G.; Ziegler, T. *J. Comput. Chem.* **2001**, *22*, 931.
- [152] Wesolowski, T. A.; Warshel, A. *The Journal of Physical Chemistry* **1993**, *97*, 8050–8053.
- [153] Mitchell, T. M.; et al. *Machine learning*. WCB; McGraw-Hill Boston, MA; 1997.
- [154] Strachey, C. S. In *Proceedings of the 1952 ACM National Meeting (Toronto)*; ACM: New York, NY, USA; ACM '52; pp 46–49.
- [155] Samuel, A. L. *Advances in computers* **1960**, *1*, 165–192.
- [156] Hughes, G. *IEEE Transactions on Information Theory* **1968**, *14*, 55–63.
- [157] Caruana, R.; Niculescu-Mizil, A. In *Proceedings of the 23rd international conference on Machine learning*; ACM; pp 161–168.
- [158] Garcia-Gimeno, R. M.; Herv'as-Martinez, C.; de Sil'oniz, M. I. *International Journal of Food Microbiology* **2002**, *72*, 19–30.
- [159] Vapnik, V.; Lerner, A. *Automation and Remote Control* **1963**, *24*.
- [160] Cortes, C.; Vapnik, V. *Machine Learning* **1995**, *20*, 273–297.
- [161] Ivanciuc, O. *Reviews in computational chemistry* **2007**, *23*, 291.
- [162] Li, H.; Liang, Y.; Xu, Q. *Chemometrics and Intelligent Laboratory Systems* **2009**, *95*, 188–198.
- [163] Burbidge, R.; Trotter, M.; Buxton, B.; Holden, S. *Computers —& Chemistry* **2001**, *26*, 5 – 14.
- [164] Borin, A.; Ferro, M. F.; Mello, C.; Maretto, D. A.; Poppi, R. J. *Analytica Chimica Acta* **2006**, *579*, 25 – 32.
- [165] Fatemi, M. H.; Gharaghani, S.; Mohammadkhani, S.; Rezaie, Z. *Electrochimica Acta* **2008**, *53*, 4276 – 4282.

- [166] Riu, J.; Rius, F. X. *Analytical Chemistry* **1996**, *68*, 1851–1857; PMID: 21619096.
- [167] Kokaly, R. F.; Clark, R. N. *Remote Sensing of Environment* **1999**, *67*, 267 – 287.
- [168] Pomerleau, D. A.; *Alvinn: An autonomous land vehicle in a neural network*; Tech. Rep.; DTIC Document; 1989.
- [169] Dahl, G. E.; Yu, D.; Deng, L.; Acero, A. *IEEE Transactions on Audio, Speech, and Language Processing* **2012**, *20*, 30–42.
- [170] Carpenter, G. A.; Grossberg, S. *Computer vision, graphics, and image processing* **1987**, *37*, 54–115.
- [171] Zhou, Z.-H.; Jiang, Y.; Yang, Y.-B.; Chen, S.-F. *Artificial Intelligence in Medicine* **2002**, *24*, 25–36.
- [172] Gasteiger, J.; Zupan, J. *Angewandte Chemie International Edition in English* **1993**, *32*, 503–527.
- [173] Hansen, K.; Biegler, F.; Ramakrishnan, R.; Pronobis, W.; von Lilienfeld, O. A.; Mller, K.-R.; Tkatchenko, A. *The Journal of Physical Chemistry Letters* **2015**, *6*, 2326–2331; PMID: 26113956.
- [174] Tetko, I. V.; Tanchuk, V. Y. *Journal of Chemical Information and Computer Sciences* **2002**, *42*, 1136–1145; PMID: 12377001.
- [175] Sumpter, B. G.; Noid, D. W. *Macromolecular Theory and Simulations* **1994**, *3*, 363–378.
- [176] Kovatcheva, A.; Golbraikh, A.; Oloff, S.; Xiao, Y.-D.; Zheng, W.; Wolschann, P.; Buchbauer, G.; Tropsha, A. *Journal of Chemical Information and Computer Sciences* **2004**, *44*, 582–595; PMID: 15032539.
- [177] Khne, R.; Ebert, R.-U.; Schrmann, G. *Journal of Chemical Information and Modeling* **2006**, *46*, 636–641; PMID: 16562993.
- [178] Chavan, S.; Friedman, R.; Nicholls, I. A. *International Journal of Molecular Sciences* **2015**, *16*, 11659.
- [179] Kowalski, B. R.; Bender, C. F. *Analytical Chemistry* **1972**, *44*, 1405–1411.
- [180] Davis, L. *Handbook of genetic algorithms*, 6th ed.; Van Nostrand Reinhold., 1991.
- [181] Julstrom, B. A. In *Proceedings of the 1994 ACM symposium on Applied computing*; ACM; pp 222–226.
- [182] Holland, J. H. *Adaptation in natural and artificial systems: an introductory analysis with applications to biology, control, and artificial intelligence.*; U Michigan Press, 1975.

- [183] Baker, J. E. In *Proceedings of the Second International Conference on Genetic Algorithms on Genetic Algorithms and Their Application*; L. Erlbaum Associates Inc.: Hillsdale, NJ, USA; pp 14–21.
- [184] Goldberg, D. E. *Genetic Algorithms in Search, Optimization and Machine Learning*, 1st ed.; Addison-Wesley Longman Publishing Co., Inc.: Boston, MA, USA, 1989.
- [185] Tanese, R.; Ph.D. thesis; Ann Arbor, MI, USA; 1989; AAI9001722.
- [186] Baker, J. E. In *Proceedings of the 1st International Conference on Genetic Algorithms*; L. Erlbaum Associates Inc.: Hillsdale, NJ, USA; pp 101–111.
- [187] De Jong, K. A.; Ph.D. thesis; Ann Arbor, MI, USA; 1975; AAI7609381.
- [188] John, H. *Machine learning, an artificial intelligence approach* **1986**, *2*, 593–623.
- [189] Booker, L. B.; Goldberg, D. E.; Holland, J. H. *Artificial intelligence* **1989**, *40*, 235–282.
- [190] Koza, J. R. *Genetic programming: A paradigm for genetically breeding populations of computer programs to solve problems*; Stanford University, Department of Computer Science, 1990.
- [191] Spears, V. M.; Jong, K. A. D. In *Proceedings of the Fourth International Conference on Genetic Algorithms*; pp 230–236.
- [192] Spears, W. M.; et al. *Foundations of genetic algorithms* **1992**, *2*, 221–237.
- [193] Mühlenbein, H. In *PPSN*; pp 15–25.
- [194] Goldberg, D. E.; Richardson, J. In *Proceedings of the Second International Conference on Genetic Algorithms on Genetic Algorithms and Their Application*; L. Erlbaum Associates Inc.: Hillsdale, NJ, USA; pp 41–49.
- [195] Smith, R.; Forrest, S.; Perelson, A. S. In *Foundations of Genetic Algorithms 2*; Morgan Kaufmann; pp 153–166.
- [196] Deb, K.; Goldberg, D. E. In *Proceedings of the 3rd International Conference on Genetic Algorithms*; Morgan Kaufmann Publishers Inc.: San Francisco, CA, USA; pp 42–50.
- [197] Eshelman, L. J. *Foundations of Genetic Algorithms 1991 (FOGA 1)* **2014**, *1*, 265.
- [198] Eshelman, L. J.; Schaffer, J. D. In *ICGA*; pp 115–122.
- [199] Grefenstette, J. J. *IEEE Transactions on Systems, Man, and Cybernetics* **1986**, *16*, 122–128.

- [200] Schaffer, J. D.; Caruana, R. A.; Eshelman, L. J.; Das, R. In *Proceedings of the Third International Conference on Genetic Algorithms*; Morgan Kaufmann Publishers Inc.: San Francisco, CA, USA; pp 51–60.
- [201] Davis, L. In *Proceedings of the Third International Conference on Genetic Algorithms*; Morgan Kaufmann Publishers Inc.: San Francisco, CA, USA; pp 61–69.
- [202] Barricelli, N. A. *Methodos* **1957**, *9*, 143–182.
- [203] Gardner, M. *Scientific American* **1970**, *223*, 120–123.
- [204] Fraser, A. S. *Australian Journal of Biological Sciences* **1960**, *13*, 150–162.
- [205] Rechenberg, I. **1965**.
- [206] Schwefel, H.-P. *Evolutionstrategie und numerische Optimierung*; Technische Universität Berlin, 1975.
- [207] Fogel, D. B.; Anderson, R. W. In *Proceedings of the 2000 Congress on Evolutionary Computation. CEC00 (Cat. No.00TH8512)*; pp 1204–1209 vol.2.
- [208] Forrest, S.; et al. *Science* **1993**, *261*, 872–878.
- [209] Deaven, D. M.; Ho, K. M. *Phys. Rev. Lett.* **1995**, *75*, 288–291.
- [210] Blommers, M. J. J.; Lucasius, C. B.; Kateman, G.; Kaptein, R. *Biopolymers* **1992**, *32*, 45–52.
- [211] Judson, R.; Jaeger, E.; Treasurywala, A.; Peterson, M. *Journal of Computational Chemistry* **1993**, *14*, 1407–1414.
- [212] Leardi, R. *Journal of Chemometrics* **2001**, *15*, 559–569.
- [213] Venkatasubramanian, V.; Chan, K.; Caruthers, J. M. *Computers & Chemical Engineering* **1994**, *18*, 833–844.
- [214] Venkatasubramanian, V.; Chan, K.; Caruthers, J. In *Proc. PSE*; pp 1001–1006.
- [215] Burden, F. R.; Rosewarne, B. S.; Winkler, D. A. *Chemometrics and Intelligent Laboratory Systems* **1997**, *38*, 127 – 137.
- [216] Sundaram, A.; Ghosh, P.; Caruthers, J. M.; Venkatasubramanian, V. *AIChE Journal* **2001**, *47*, 1387–1406.
- [217] Xiao, Y. L.; Williams, D. E. *Computers & Chemistry* **1994**, *18*, 199–201.
- [218] Unger, R.; Moul, J. *Journal of molecular biology* **1993**, *231*, 75–81.
- [219] Ebeling, M.; Nadler, W. *Biopolymers* **1997**, *41*, 165–180.
- [220] Tsay, J.-J.; Su, S.-C. *Proteome science* **2013**, *11*, 1.

- [221] Mannhold, R.; Kubinyi, H.; Timmerman, H.; Clark, D. E. *Evolutionary algorithms in molecular design*; John Wiley & Sons, 2008; Vol. 8.
- [222] Devillers, J. *Genetic algorithms in molecular modeling*; Academic Press, 1996.
- [223] Mitchell, M. *An introduction to genetic algorithms*; MIT press, 1998.
- [224] BP, V. E.; Kroon *Acta crystallographica. Section B, Structural science* **2000**, *56*, 535–542.
- [225] Valle, R. G. D.; Venuti, E.; Brillante, A.; Girlando, A. *The Journal of Chemical Physics* **2003**, *118*.
- [226] Shoemake, K. *ACM SIGGRAPH computer graphics* **1985**, *19*, 245–254.
- [227] Misquitta, A. J.; Welch, G. W.; Stone, A. J.; Price, S. L. *Chemical Physics Letters* **2008**, *456*, 105–109.
- [228] Coutsiias, E. A.; Seok, C.; Dill, K. A. *Journal of computational chemistry* **2004**, *25*, 1849–1857.
- [229] Pidcock, E.; Motherwell, W. D. S. *Crystal Growth & Design* **2004**, *4*, 611–620.
- [230] Case, D. H.; Campbell, J. E.; Bygrave, P. J.; Day, G. M. *J. Chem. Theory Comput.* **2015**.
- [231] Gottschalk, S.; *Separating axis theorem*; Tech. Rep. Technical Report TR96-024; Department of Computer Science, UNC Chapel Hill; 1996.
- [232] O'Rourke, J. *Computational Geometry in C*, 2nd ed.; Cambridge University Press, 2013.
- [233] Paulus, E. F.; Leusen, F. J. J.; Schmidt, M. U. *CrystEngComm* **2007**, *9*.
- [234] Hiramoto, M.; Kawase, S.; Yokoyama, M. *Japanese Journal of Applied Physics* **1996**, *35*.
- [235] Gowacki, E.; Mihai, I.; Kaltenbrunner, M.; Gsiorowski, J.; White, M.; Monkowius, U.; Romanazzi, G.; Suranna, G.; Mastroilli, P.; Sekitani, T.; Bauer, S.; Someya, T.; Torsi, L.; Sariciftci, N. *Advanced materials (Deerfield)*.
- [236] Gao, H. *International Journal of Quantum Chemistry* **2012**, *112*.
- [237] White, N. J. *Journal of Clinical Investigation* **2004**, *113*, 1084–1092.
- [238] Hou, J.; Wang, D.; Zhang, R.; Wang, H. *Clinical Cancer Research* **2008**, *14*, 5519–5530.
- [239] Pyzer-Knapp, E. O.; Thompson, H. P. G.; Schiffmann, F.; Jelfs, K. E.; Chong, S. Y.; Little, M. A.; Cooper, A. I.; Day, G. M. *Chem. Sci.* **2014**, *5*, 2235.

- [240] Pyzer-Knapp, E. O.; Ph.D. thesis; University of Cambridge; 2014.
- [241] Thompson, H. P. G.; Ph.D. thesis; University of Cambridge; 2014.
- [242] Williams, D. E. *J. Comput. Chem.* **2001**, *22*, 1154–1166.
- [243] Becke, A. D. *J. Chem. Phys.* **1993**, *20*, 5648–5652.
- [244] Lee, C.; Yang, W.; Parr, R. G. *Physical review B* **1988**, *37*, 785.
- [245] Nyman, J.; Day, G. M. *CrystEngComm* **2015**, *17*, 5154–5165.
- [246] Grimme, S. *Angew. Chem. Int. Ed.* **2008**, *47*, 3430–3434.
- [247] Reilly, A. M.; Cooper, R. I.; Adjiman, C. S.; Bhattacharya, S.; Boese, A. D.; Brandenburg, J. G.; Bygrave, P. J.; Bylsma, R.; Campbell, J. E.; Car, R.; et al. *Acta Crystallographica Section B* **2016**, 1–59.
- [248] Kolossváry, I.; Guida, W. C. *J. Am. Chem. Soc.* **1996**, *118*, 5011–5019.
- [249] Kolossváry, I.; Guida, W. C. *J. Comp. Chem.* **1999**, *20*, 1671–5019.
- [250] Mohamadi, F.; Richards, N. G.; Guida, W. C.; Liskamp, R.; Lipton, M.; Caufield, C.; Chang, G.; Hendrickson, T.; Still, W. C. *Journal of Computational Chemistry* **1990**, *11*, 440–467.
- [251] Jorgensen, W. L.; S., M. D.; Tirado-Rives, J. *J. Am. Chem. Soc.* **1996**, *118*, 11225–11236.
- [252] Kaminski, G. A.; Friesner, R. A.; Tirado-Rives, J.; Jorgensen, W. L. *The Journal of Physical Chemistry B* **2001**, *105*, 6474–6487.
- [253] *Schrodinger LLC, New York, NY, MacroModel, V9.9.013*; 2014.
- [254] Thompson, H. P. G.; Day, G. M. *Chemical Science* **2014**, *5*, 3173–3182.
- [255] Grimme, S.; Antony, J.; Ehrlich, S.; Krieg, H. *J. Chem. Phys.* **2010**, *132*, 154104–154104–19.
- [256] Grimme, S.; Ehrlich, S.; Goerigk, L. *Journal of computational chemistry* **2011**, *32*, 1456–1465.
- [257] Abraham, N.; Probert, M. *Physical Review B* **2006**, *73*, 224104.
- [258] Kazantsev, A.; Karamertzanis, P.; Pantelides, C.; Adjiman, C.; *CrystalOptimizer: An Efficient Algorithm for Lattice Energy Minimization of Organic Crystals Using Isolated-Molecule Quantum Mechanical Calculations*; 2010.
- [259] Anthony, J. *Angewandte Chemie (International ed. in English)* **2008**, *47*, 452–483.
- [260] Kitamura, M.; Arakawa, Y. **2008**.

- [261] Mas-Torrent, M.; Rovira, C. *Chem. Rev.* **2011**, *111*, 4833–4856.
- [262] Bunz, U. H. F. *Acc. Chem. Res.* **2015**, *48*, 1676–1686; PMID: 25970089.
- [263] Miao, Q. *Adv. Mater.* **2014**, *26*, 5541–5549.
- [264] Campbell, R. B.; Robertson, J. M.; Trotter, J. *Acta Crystallographica* **1961**, *14*.
- [265] Campbell, R. B.; Robertson, J. M.; Trotter, J. *Acta Crystallographica* **1962**, *15*.
- [266] Holmes, D.; Kumaraswamy, S.; Matzger, A. J.; Vollhardt, K. P. C. *Chem. Eur. J.* **1999**, *5*, 3399–3412.
- [267] Schiefer, S.; Huth, M.; Dobrinevski, A.; Nickel, B. *J. Am. Chem. Soc.* **2007**, *129*, 10316–10317.
- [268] Christine, C. M.; Anne, B. D.; Jacob, B.; Gert, T. O.; Auke, M.; Jan, L. d. B.; Thomas, T. M. P. *Synth. Met.* **2003**, *138*, 475 – 481.
- [269] Isoda, K.; Nakamura, M.; Tatenuma, T.; Ogata, H.; Sugaya, T.; Tadokoro, M. *Chem. Lett.* **2012**, *41*, 937–939.
- [270] Cicoira, F.; Santato, C.; Dinelli, F.; Murgia, M.; Loi, M. A.; Biscarini, F.; Zamboni, R.; Heremans, P.; Muccini, M. *Advanced functional materials* **2005**, *15*, 375–380.
- [271] Desiraju, G. R. *Angew. Chem. Int. Ed* **1995**, *34*, 2311–2327.
- [272] Perdew, J. P.; Chevary, J. A.; Vosko, S. H.; Jackson, K. A.; Pederson, M. R.; Singh, D. J.; Fiolhais, C. *Phys. Rev. B* **1992**, *46*, 6671–6687.
- [273] te Velde, G.; Bickelhaupt, F. M.; Baerends, E. J.; Fonseca Guerra, C.; van Gisbergen, S. J.; Snijders, J. G.; Ziegler, T. *J. Comput. Chem.* **2001**, *22*, 931.
- [274] Sokolov, A. N.; Atahan-Evrenk, S.; Mondal, R.; Akkerman, H. B.; Sánchez-Carrera, R. S.; Granados-Focil, S.; Schrier, J.; Mannsfeld, S. C.; Zoombelt, A. P.; Bao, Z.; Aspuru-Guzik, A. *Nature Comm.* **2011**, *2*, 437.
- [275] Valle, R. G. D.; Venuti, E.; Brillante, A.; Girlando, A. *The Journal of Chemical Physics* **2003**, *118*.
- [276] Curtis, M.; Cao, J.; Kampf, J. *Journal of the American Chemical Society* **2004**, *126*, 4318–4328.
- [277] Bunz, U. H. F. *Accounts of Chemical Research* **2015**, *48*, 1676–1686; PMID: 25970089.
- [278] Heller, S.; McNaught, A.; Stein, S.; Tchekhovskoi, D.; Pletnev, I. *Journal of Cheminformatics* **2013**, *5*, 7.

- [279] Landrum, G. *Online*). [http:// www. rdkit. org](http://www.rdkit.org). Accessed **2006**, 3, 2012.
- [280] Rappé, A. K.; Casewit, C. J.; Colwell, K.; Goddard Iii, W.; Skiff, W. *Journal of the American chemical society* **1992**, 114, 10024–10035.
- [281] Tanimoto, T. T. **1958**.
- [282] Bajusz, D.; Rácz, A.; Héberger, K. *Journal of cheminformatics* **2015**, 7, 20.
- [283] Paulus, E. F.; Leusen, F. J. J.; Schmidt, M. U. *CrystEngComm* **2007**, 9.