

University of Southampton

Day Research Group
Department of Chemistry
Faculty of Natural & Environmental Sciences

Incorporating Molecular Flexibility and Conformational Variability into Crystal Structure Prediction

Thomas Simon Gee

*A thesis submitted in fulfilment of the requirements for the degree of
Doctor of Philosophy*

April 2017

Declaration of Authorship

I, Thomas Simon Gee, declare that this thesis and the work presented within are my own. I confirm that:

- This work was done wholly while in candidature for a PhD research degree at the University of Southampton.
- Where I have consulted the published work of others, reference 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 of the main sources of help.
- Where the thesis is based on work that has been performed jointly with others, I have made clear exactly what was done by others and what I have contributed myself.

Signed:

Date:

The difference between a quantum mechanic and a classical mechanic is that the former can get their car into the garage without opening the door. In crystal structure prediction, we can morph into either whenever it suits us...

Acknowledgements

First and foremost my thanks go to my supervisor, Professor Graeme Day, for his extraordinary wisdom and remarkable ability to keep cool in the moments where I was questioning my methods (and sanity). I also thank him for being the most relaxed academic I have met in my short academic career where allowing for freedom and flexibility (no pun intended) to prosper actually increases performance; perhaps something that could seem counter-intuitive to other academics...

I cannot pass this section without thanking Dr Peter Bygrave; perhaps the most competent post-doctoral scientist and “Python-er” I have (and will) ever meet. He has aided me throughout my time in the Day research group; particularly when I first started when I was getting used to the many acronyms associated with the field (DMACRYS, DMAREL, DMAFLEX, NEIGHCRYS, GDMA, COMPACK, UPACK and RANCEL to name a few). On the subject of software, an extended thanks also go to the many users of the StackOverflow and Mathematica forums who have already asked almost of all the many programming questions that I needed answering.

My thanks also go out to the rest of the Day research group including: Dave (for helping me discover the best Chinese restaurant in the UK), Josh (for being a committed gym buddy and for riveting discussions about non-work related matters) and the other group members (past and present): Jonas, Julien, Dave, Chris, Jack, Srhidar and Angeles.

Finally, I would like to thank Jenni for being my rock and without her motivation and focus I would probably never have even gone into science at undergraduate level, let alone this PhD research project.

Abstract

The ability to predict the properties of a crystal structure before any empirical analysis or laboratory work has commenced offers the opportunity for vast reductions in research and development costs for new products. This research focuses on a novel methodology to incorporate molecular flexibility into crystal structure prediction (CSP).

Chapter 3: The quantification of the effect of incorporating molecular flexibility and a revised Williams99 potential in the lattice energy minimisation finds that 5.5% and 6.3% more observed crystal structure matches, respectively, lie within 1 kJ mol^{-1} of the calculated global energy minimum structure.

Chapter 4: A novel method is presented that simultaneously samples the molecular conformational space and the unit cell parameters when generating crystal structures where the former employs molecular principal displacements. This method was used for one test molecule that possesses 3 known polymorphs where only 1 is found when a standard CSP approach is used; this method now locates all 3 polymorphs.

Chapter 5: Presents a scientifically robust and computationally efficient method for finding a set of principal displacements and their corresponding contributions for converting one molecular conformation into another.

Chapter 6: The molecular strain energy induced by crystal packing forces was calculated for 224 molecules. It was found that a maximum number of principal displacements up to a force constant value of $0.084 \text{ mDyne \AA}^{-1}$ were required to accurately reproduce the in-crystal conformation from its gas phase conformer for approximately 95% of cases.

Chapters 7 & 8: Present two CSP case studies. The first was a sixth blind test molecule that failed to be predicted when using a rigid molecule search procedure coupled with a flexible molecule lattice energy minimisation and hence was used as motivation to implement the methodology presented in Chapter 4 on the second case study of a novel herbicide molecule.

Contents

Declaration of Authorship	ii
Quotation	iii
Acknowledgements	iv
Abstract	v
Contents	vi
List of Abbreviations	xi
List of Figures	xi
List of Tables	xix
1 Introduction	1
1.1 Introduction to Crystal Structure Prediction	1
1.2 The Intra-Intermolecular Energy Problem: the Motivation for Incorporating Molecular Flexibility into CSP	3
1.3 Thesis Overview	5
1.4 A Note on Software	6
1.5 A Note on Nomenclature	7
1.6 A Note on Computational Resources	7
2 Theory	9
2.1 The CSP Process	9
2.2 Molecular Conformational Searches	9
2.2.1 Molecular Flexibility	9
2.2.1.1 Internal Degrees of Freedom	9
2.2.1.2 Molecular Principal Displacements	13
2.2.1.3 Linear versus Curvilinear Space	14
2.2.1.4 Propagating Along Molecular Principal Displacements	15
2.2.2 Method for Searching Molecular Conformational Space	17
2.3 Gas Phase Geometry Optimisation	18
2.3.1 Density Functional Theory	18
2.3.2 Basis Sets	20
2.3.3 Minimisation of the Molecular Potential Energy	21
2.3.4 Molecular Electrostatic Potentials	23

2.3.5	Rigid Molecule Approximation	25
2.4	Crystal Structure Generation	26
2.4.1	From Molecule to Crystal	26
2.4.2	Defining the Search Space	27
2.4.3	Molecular Transformations	28
2.4.3.1	Molecular Translation	28
2.4.3.2	Molecular Rotation	29
2.4.3.3	Comparison of Molecular Geometries	30
2.4.4	Sampling the Search Space	31
2.4.5	Flexible Molecule Crystal Structure Generation	33
2.4.5.1	Surface Fitting	34
2.5	Crystal Structure Energy Minimisation	35
2.5.1	Intermolecular Interactions	35
2.5.1.1	Short-Range Interactions	37
2.5.1.2	Long-Range Interactions	37
2.5.2	Force-Field Models	39
2.5.3	Force-Field Parametrisation	40
2.5.4	Flexible Molecule Energy Minimisation	42
2.5.4.1	Molecular Flexibility in Crystal Structures	42
2.5.4.2	Molecular Strain Energy	44
2.5.5	Methods for Flexible Molecule Energy Minimisation	45
2.6	Clustering of Non-Unique Crystal Structures	49
2.7	Crystal Structure Ranking	50
2.8	The Blind Tests	51
2.9	Summary	61
3	Molecular Flexibility & Williams99 Force-Field Testing	63
3.1	Introduction	63
3.2	Obtaining a Test Set of Molecules	63
3.3	Defining Flexibility	64
3.4	Computational Method	66
3.4.1	A Note to the Reader	67
3.5	Results	68
3.6	Discussion	74
3.6.1	ΔE Values	74
3.6.2	N_{Lower} Values	78
3.6.3	$RMSD_{30}$ Values	78
3.6.4	Specific Cases	80
3.6.5	Conclusions	83
4	Principal Displacement Conformational Searches	85
4.1	Introduction	85
4.2	Motivation for a Flexible Molecule Search Procedure	86
4.3	Degrees of Freedom for FUQLIM	89
4.4	Molecular Conformational Space	90
4.4.1	Defining the Conformational Spaces	90
4.4.2	Potential Energy Surface Fitting	93

4.4.3	Sampling the Conformational Spaces	94
4.5	Generalised Methodology	95
4.6	Crystal Structure Prediction	97
4.6.1	How Much is Molecular Flexibility Needed?	98
4.6.2	(3,3) Molecular Flexibility	98
4.6.2.1	Structural Comparisons	105
4.6.3	1-DOF and 2-DOF Flexibility	108
4.7	Evaluation of the Sampling Procedure	109
4.8	Conclusions	111
5	Molecular Geometry Interconversion Using Principal Displacements	113
5.1	Introduction	113
5.2	The Test Set of Molecules	114
5.3	Pure Cartesian Approach	117
5.3.1	Theory	117
5.4	RMSD Minimisation	119
5.4.1	Theory	120
5.4.2	Hydrogen Peroxide: a Preliminary Example	120
5.4.3	An Alternative to RMSD	123
5.5	Direct Solution	125
5.6	Results	126
5.6.1	ODNPDS	127
5.6.2	FIBKUW	131
5.7	Conclusions	135
6	Decomposition of Molecular Strain in Crystals	137
6.1	Introduction	137
6.2	Defining Non-Bonded Intramolecular Interactions	138
6.3	Methodology	140
6.3.1	Obtaining a Test Set of Molecules	140
6.3.2	Analysis of the Test Set	142
6.3.3	Calculating ΔE_{strain} and Molecular Principal Displacements	145
6.3.4	Reproducing the Observed Crystal Structure	145
6.4	Results	147
6.4.1	ΔE_{strain} Comparison: CrystalOptimizer versus CRYSTAL09	147
6.4.2	ΔE_{strain} Analysis	151
6.4.3	Molecular Properties versus ΔE_{strain}	154
6.4.4	RMSD versus ΔE_{strain}	156
6.4.5	Reproducing the Observed Crystal Structure	159
6.4.6	Force Constant Analysis	164
6.4.7	Displacement Analysis	170
6.5	Conclusions	171
7	Case Study I: Blind Test Molecule XXVI	175
7.1	Introduction to the Sixth Blind Test	175
7.2	Author's Contribution to the Sixth Blind Test	177
7.3	Results for Molecule XXVI	180

7.4	Conclusion	184
8	Case Study II: Flufenacet	187
8.1	Introduction to Flufenacet	187
8.2	Crystal Structure Prediction	188
8.2.1	Rigid Molecule CSP Methodology	190
8.2.2	Flexible Molecule CSP and Analysis	190
8.2.2.1	Selecting the Number of Principal Displacements	192
8.2.2.2	Calculating the Principal Displacement Bounds	193
8.2.2.3	Fitting the Energy Model	195
8.2.2.4	Flexible Molecule Crystal Structure Generation	197
8.2.2.5	Flexible Molecule Lattice Energy Minimisation	197
8.3	Rigid Molecule CSP Results	198
8.3.1	Intermolecular Lattice Energy versus Molecular Surface Area	200
8.4	Flexible Molecule CSP Results	202
8.4.1	Evaluation of the Sampling Procedure	204
8.5	Conclusions	207
9	Conclusions & Future Work	211
9.1	Molecular Flexibility & Williams99 Force-Field Testing	212
9.2	Flexible Molecule Structure Generation	213
9.3	CSP Case Studies	214
A	54 Small, Organic Molecules	217
B	Molecular Geometry Interconversion Using Principal Displacements	223
C	Decomposition of Molecular Strain in Crystals	231
D	Case Study II: Flufenacet	237
	Bibliography	241
	Final Thought	256

List of Abbreviations

CSP	Crystal structure prediction
DFT	Density functional theory
CSD	Cambridge Structural Database
D	Dimensional
CCDC	Cambridge Crystallographic Data Centre
W99	Williams 1999
DOF	Degree of freedom
PES	Potential energy surface
DMA	Distributed multipole analysis
LMCS	Low-mode conformational search
OPLS	Optimized Potential for Liquid Simulations
LAM	Local approximate model
RMSD	Root mean squared deviation
BFGS	Broyden-Fletcher-Goldfarb-Shanno
MEP	Molecular electrostatic potential
SKD	Smoothing kernel density
PCM	Polarisable continuum model
CoM	Centre of mass
CPU	Central processing unit

List of Figures

1.1	(a) Ritonavir A possesses 6 intermolecular hydrogen bonds and has a larger surface area than (b) Ritonavir B that forms a dimer complex and only possesses two intermolecular hydrogen bonds (atoms that are in bold font are those that participate in hydrogen bonding).	3
1.2	(a) and (b) show the 2 molecular conformations of ROY.	4
2.1	CSP flowchart that shows the fundamental steps of the rigid molecule CSP process. Note the intramolecular geometry is fixed after the gas phase geometry optimisation stage.	10
2.2	(a) Gas phase geometry of hydrogen peroxide with ϕ at approximately $\pm \frac{2\pi}{3}$ radians. The dashed, red lines are included to aid in defining the 3D geometry of the molecule in a 2D space. (b) shows the internal energy, U_{dihedral} in kJ mol^{-1} of hydrogen peroxide as a function of ϕ between $-\pi$ and π radians.	11
2.3	The displacement of hydrogen peroxide along its lowest energy principal displacement. The red and blue geometry refer to the equilibrium and displaced geometries of hydrogen peroxide, respectively. The orange and purple arrows show the path of the hydrogen atom as it is displaced along this principal displacement in linear and curvilinear space, respectively. .	14
2.4	A Slater and a series of Gaussian functions, showing the approximation of using 1, 2 and 3 Gaussian functions in red, green and black, respectively, to approximate a single Slater function in blue.	21
2.5	An example of a simple potential energy surface plotted as the molecular energy as a function of molecular displacements. If the black point is the energy of the starting point, the green point exists as the post-minimisation energy. The blue point is the closest geometry minimum to the starting point and the red point is the global potential energy minimum. .	22
2.6	Showing, from top to bottom the monopole, dipole, quadrupole and octupole moments in a multipole expansion.	24
2.7	shows grid-based (a), purely-random (b) and quasi-randomly (Sobol sequence) (c) generated set of points between 0 and 1 in two dimensions. .	32
2.8	(a) shows a molecular conformation coloured by element and (b) that shows the same molecular conformation coloured by rigid fragments (red) and soft torsional angles (blue).	33
2.9	(a) the in-crystal geometry of a polymorph of penicillin. (b) compares the molecular conformers of the gas phase (blue) and the in-crystal (red) geometries, where the latter was calculated using the B3LYP-GD3BJ/6-311G** level of theory.	43

2.10	Two hypothetical potential energy surfaces showing (a) the energy difference between two molecular conformers, ΔE_{conf} , and (b) the energy difference between two molecular conformations that lie in the same potential energy well, ΔE_{strain}	44
2.11	The CrystalOptimizer algorithm. The DMACRYS and GAUSSIAN09 software packages are used during stages 1a, 3 and 1b, 2a, 2b, 2c respectively. U_{latt} and θ^r represent the lattice energy and the set of ‘rigid’ molecular degrees of freedom.	47
2.12	shows two potential energy surfaces plotting the potential energy as a function of the lattice parameters, structural coordinate. (a) illustrates a hypothetical potential energy surface where every red point represents a crystal structure that possesses a unique set of unit cell parameters and hence lies at a unique position. (b) shows the position of these unique crystal structures post-lattice energy minimisation.	49
3.1	(a) and (b) shows the comparison of the ΔE values for the RevFlex vs OrigFlex and RevFlex vs RevRigid methods, respectively. The black ‘+’, blue ‘ Δ ’ and red ‘O’ refer to molecules that are non-polar and do not participate in hydrogen bonding, molecules that are polar but do not participate in hydrogen bonding and molecules that are polar and do participate in hydrogen bonding, respectively. The black lines are present to aid in comparison of the two methodologies.	75
3.2	(a) and (b) shows smoothing kernel density plots for the ΔE values for the RevFlex versus RevRigid and OrigFlex versus RevFlex methodologies, respectively.	77
3.3	(a) and (b) shows the comparison of the N_{Lower} values for the OrigFlex versus RevFlex and RevFlex versus RevRigid, respectively. Bin ‘0’ represents a ‘perfect’ CSP result. The colour scheme follows the method that is expected to perform better is in blue.	79
3.4	(a) and (b) shows comparison of the $RMSD_{30}$ values for the RevFlex versus RevRigid and OrigFlex versus RevFlex, respectively. The black ‘+’, blue ‘ Δ ’ and red ‘O’ refer to molecules that are non-polar and do not participate in hydrogen bonding, molecules that are polar but do not participate in hydrogen bonding and molecules that are polar and do participate in hydrogen bonding, respectively. The black lines are present to aid in comparison of the two methodologies.	81
3.5	(a) and (b) shows the geometric comparison of the observed DUNVEN molecular geometry (red) and the RevRigid and RevFlex intramolecular geometry (coloured by element), respectively.	82
4.1	The in-crystal geometry of the FUQLIM polymorph (coloured by element). Note the 3 soft torsion angles (<i>Torsion 1</i> , <i>Torsion 2</i> and <i>Torsion 3</i>) formed by the disulfide bridge that connects the two aromatic systems. 86	
4.2	Illustration of the gas phase FUQLIM molecular conformation (red) against the in-crystal conformation (blue) for the 3 known polymorphs A, B and C in figures (a) , (b) and (c) respectively.	87

4.3	The principal displacements for FUQLIM with the first, second and third lowest force constants in (a) , (b) and (c) respectively. The curved red and blue lines represent the paths of the atoms in the positive and negative directions, respectively, of the principal displacement in curvilinear space about the gas phase geometry.	91
4.4	Principal and torsional displacement spaces for 22.5 kJ mol ⁻¹ and 5.0 kJ mol ⁻¹ energy limits above the energy of the equilibrium geometry for the FUQLIM molecule highlighted in blue and purple respectively.	92
4.5	The fitting errors in the intramolecular energy when using varying numbers of training points. The dashed line is intended to easily identify to the reader when the maximum absolute error reduces below 1 kJ mol ⁻¹	93
4.6	(a) shows the variation in calculated atomic charges on 4 selected atoms, (b) , as the molecular geometry is displaced along the first principal displacement. The MULFIT and CHELPG charges are presented as solid and dashed lines, respectively, where the latter indicates greater stability and smoother changes with the molecular geometry.	94
4.7	100
4.8	101
4.9	103
4.10	104
4.11	Crystal structures yielded using the torsion angle method (carbons in red) for FUQLIM polymorphs A (a) , B (b) and C (c) . Hydrogen atoms have been removed for clarity.	106
4.12	Crystal structures yielded using the principal displacement methodologies (carbons in red) for FUQLIM polymorphs A (a) , B (b) and C (c) . Hydrogen atoms have been removed for clarity.	107
4.13	All 4 plots show the number of unique crystal structures being generated for a given total energy against the number valid minimisations.	110
5.1	The 13 molecules included in this study. The molecules are referred to by their CSD reference codes and the 3 numbers in parentheses following each molecule refer to the number of known polymorphs, the number of independent molecular geometries (summed over all known polymorphs) and the number of unique conformers found in all known polymorphs, respectively.	115
5.2	Description of RMSD minimisation procedure work flow. The RMSD convergence criteria is a user-defined value.	121
5.3	122
5.4	Overlay of the target and base (coloured by element) geometries for the ODNPDS02 (blue) and ODNPDS11 (red) molecules.	128
5.5	(a) and (c) show the reduction in RMSD as more numbers of principal displacements are added into the procedure for ODNPDS02 and ODNPDS11, respectively. (b) and (d) show the displacements for each of the principal displacements with the 10 lowest valued force constants labelled for ODNPDS02 and ODNPDS11, respectively.	129
5.6	Overlay of the target (coloured) and base (coloured by element) geometries for the FIBKUW01 (blue) and FIBKUW02 (red) molecules.	132

5.7	(a) and (c) show the reduction in RMSD as more numbers of principal displacements are added into the procedure for FIBKUW01 and FIBKUW02, respectively. (b) and (d) show the displacements for each of the principal displacements with the 10 lowest valued force constants labelled for FIBKUW01 and FIBKUW02, respectively.	134
6.1	(a), (b) and (c) showing examples of non-covalent intramolecular hydrogen bonding, polar interactions and non-polar interactions, respectively, highlighted by the red, dashed lines.	139
6.2	(a) shows the distribution of the number of atoms with the ranges of molecular Sets 1, 2 and 3 and (b), the number of rotational bonds in the 224 molecule test set.	143
6.3	The work flow for measuring the maximum RMSD tolerance ($RMSD_{tol}$) required to approximate the in-crystal geometry to accurately reproduce (B). Crystal structures and molecular geometries are labelled with letters and numbers, respectively.	148
6.4	Comparison of ΔE_{strain} values, kJ mol^{-1} , calculated from the CRYSTAL09, $\Delta E_{strain}^{CRYSTAL}$, and CrystalOptimizer, $\Delta E_{strain}^{CrysOpt}$, software packages.	150
6.5	(a) a histogram of the ΔE_{strain} values separated by 1 kJ mol^{-1} bin widths. (b) distribution of ΔE_{strain} values, in kJ mol^{-1} , in descending order for the 224 molecule test set (molecules that form a non-covalent intramolecular interaction are highlighted in differing colours).	152
6.6	(a) and (b) shows the in-crystal and gas phase geometries, respectively, of the KIMSOO molecule. Note the twist of the amine and alcohol groups that allows 2 intramolecular hydrogen bonds to be formed.	153
6.7	155
6.8	RMSD values between the gas and in-crystal geometries for the 224 molecule test set partitioned by the number of rotatable bonds in the molecule.	157
6.9	The ΔE_{strain} values as a function of the RMSD between the in-crystal and gas phase geometries for the 224 molecule test set. SOKREQ and KIMSOO are labelled as they possess the largest RMSD and ΔE_{strain} values, respectively.	158
6.10	The in-crystal geometry (coloured by element) overlaid with that of the gas phase (red) geometry ($RMSD = 3.136 \text{ \AA}$), for the SOKREQ molecule.	158
6.11	The number of approximated crystal structures that gave matches to their original counterparts at a given RMSD tolerance. The number at the top of each column shows the percentage change, rounded to the nearest whole number, as the RMSD tolerance is reduced to the columns value from the RMSD tolerance of the immediate right column.	159
6.12	The distribution of the changes in the $RMSD_{30}$ values between the crystal structures containing the exact and approximate in-crystal geometries for each $RMSD_{tol}$ value. The average values for each distribution are highlighted in red.	161
6.13	$RMSD_1$ values (between the exact and approximated in-crystal geometries) as a function of the $RMSD_{30}$ values (between the exact and approximated crystal structures).	162
6.14	Distributions of the intermolecular component of the total lattice energy (relative to the exact lattice energy) for each $RMSD_{tol}$. The red points represent the average lattice energy for each distribution.	163

6.15	The distribution of force constants by molecule (dark grey) partitioned by the number of rotatable bonds. The larger, coloured points show the maximum force constant required to strain the molecule away from its gas phase geometry such that a match is yielded against the observed crystal structure. The blue, green, light grey and red points correspond to molecules that do not form, form polar, non-polar or hydrogen bonding non-covalent intramolecular interactions respectively.	165
6.16	The force constants required to bring the RMSD between the in-crystal and approximated in-crystal geometry to within 0.2 Å against the ΔE_{strain} (a) and RMSD values (b). Both of these graphs are cropped at 1 mDyne Å ⁻¹ to show the spread of the 99% of the force constant values.	167
6.17	169
6.18	The extent of displacement for each principal displacement that possesses a given force constant value for each molecule in Sets 2 and 3.	170
7.1	The target molecules for the sixth blind test referred to as XXII through XXVI.	176
7.2	Molecule XXVI annotated by the dihedral (blue) and bond (red) angles that are modelled as flexible by CrystalOptimizer.	179
7.3	XXVI final crystal structures for both sets using a polarisable continuum model with a permittivity of 3.0 ϵ_0 and 7.0 ϵ_0 for (a) and (b) respectively. The location of the observed crystal structure is labelled and was not found in either set.	181
7.4	(a) shows the molecular overlay of the conformer of XXVI within the observed structure (coloured by element) and the conformer of the best matched structure submitted by the Author (carbons in red). (b) shows the intermolecular hydrogen bonds formed between 2 XXVI molecules in the observed crystal structure. Hydrogen atoms have been removed for clarity.	182
7.5	(a) shows the molecular overlays between the in-crystal geometry (coloured by element) and the closest conformer match (carbons in red) from a conformational search of molecule XXVI. (b) shows the gas phase geometry of molecule XXVI. Hydrogen atoms have been removed for clarity.	182
7.6	The observed crystal structure of molecule XXVI with atoms coloured by element and a depth cue to aid in the visualisation of the 3D crystal structure. Hydrogen atoms have been removed for clarity.	183
8.1	The flufenacet molecule annotated with blue, curved arrows that represent the soft torsion angles.	187
8.2	(a), (b) and (c) showing conformers 1, 2 and 3 of the flufenacet molecule that possess the lowest $E_{\text{conf,bias}}$, the largest surface area and the lowest molecular energy respectively.	191
8.3	Shows the mean unsigned error (red), standard deviation (blue) and maximum absolute error (green) for each number of training points used for the flufenacet conformers 1, 2 and 3. The dashed lines are plotted at 1 kJ mol ⁻¹ to aid in the reading of the figures.	196
8.4	CSP landscape for the rigid molecule CSP of flufenacet.	198
8.5	Global lattice energy minimum crystal structure for the flufenacet molecule in the $P\bar{1}$ space group. Hydrogen atoms have been removed for clarity.	199

8.6	CSP landscape for the rigid molecule search and lattice energy minimisation CSP procedure of flufenacet conformers 1 (blue), 2 (red) and 3 (green).	199
8.7	The correlation between the molecular surface area, A_{Connolly} , and the intermolecular lattice energy, $E_{\text{latt,inter}}$, from the lowest energy crystal structures for each of the 22 conformers of flufenacet. Conformers 1, 2 and 3 are highlighted. The line represents a line of best fit to the data which possesses the equation $E_{\text{latt,inter}} = -0.27(\text{\AA}^{-2}) \cdot A_{\text{Connolly}}(\text{\AA}^2) - 45.89(\text{ kJ mol}^{-1})$	201
8.8	CSP landscape for flufenacet conformers 1 (blue), 2 (green) and 3 (red). .	203
8.9	Shows the overlay of the molecular geometries (RMSD=0.375 \AA) of the global energy minimum crystal structures for flufenacet from the rigid (carbons in red) and flexible (coloured by element) CSP processes. Hydrogen atoms have been removed for clarity.	203
8.10	The global energy minimum crystal structure yielded from the flexible molecule CSP process. Hydrogen atoms have been removed for clarity. .	204
8.11	CSP landscapes for conformers 1, 2 and 3 in (a), (c) and (e), respectively. (b), (d) and (f) shows the in-crystal molecular geometry that afforded the lowest total energy minimum crystal structure (coloured by element) against the original geometry of that conformer (carbons in red) for conformers 1, 2 and 3 respectively. Hydrogen atoms have been removed for clarity.	205
8.12	All 6 plots show the number of unique crystal structures being generated for conformer 1 of flufenacet for a given total energy against the number of valid minimisations. The figure captions indicate the space group. . .	206
8.13	CSP landscape for flufenacet conformers 1 (blue circles), 2 (green triangles) and 3 (red squares) under the rigid molecule approximation. An 'X' represents the location of crystal structures that were present in the rigid search set but not the flexible search set whereas the other shapes show the location of crystal structures that were present in both sets.	207
B.1	The two molecules, HIBGUV and HAJYUN, from the test set of Thompson that were excluded from the study in Chapter 5 due to the presence of halogen atoms.	223
D.1	All 6 plots show the number of unique crystal structures being generated for conformer 2 of flufenacet for a given total energy against the number valid minimisations. The figure captions indicate the space group. . . .	238
D.2	All 6 plots show the number of unique crystal structures being generated for conformer 3 of flufenacet for a given total energy against the number valid minimisations. The figure captions indicate the space group. . . .	239

List of Tables

2.1	Summary of the commonly referred to intermolecular interactions.	38
2.2	All W99 atom type definitions and their environments.	41
2.3	Molecules included in all of the six blind tests.	54
3.1	Illustration of all possible hydrogen bonding combinations defined by the W99 atom typing. Yellow and red cells are interactions that are included and not included, respectively, in the set of 50 molecules. Blue cells are combinations that are not included in the original set of 50 molecules but are included in the revised set of 54 molecules.	64
3.2	54 Small Organic Molecules and Results of the CSP for the RevRigid, OrigFlex and RevFlex methods. All ΔE and $RMSD_{30}$ values are quoted in kJ mol^{-1} and \AA , respectively. Results that are accompanied by an asterisk, *, will be the subject of discussion in Section 3.6.4. The '+', ' Δ ' and 'O' symbols refer to molecules that are non-polar and do not participate in hydrogen bonding, molecules that are polar but do not participate in hydrogen bonding and molecules that are polar and do participate in hydrogen bonding, respectively. The ΔE values are derived from taking the energy differences between the lowest energy unobserved structure and the lowest energy observed structure. The N_{Lower} values refer to the number of crystal structures between the lowest energy unobserved structure and the lowest energy observed structure.	69
3.3	collection of ΔE data points in each quadrant of Figure 3.1.	76
3.4	The integration values within the bounds of -10 kJ mol^{-1} to 1 kJ mol^{-1} of the smoothing kernel density functions displayed in Figure 3.2.	78
4.1	The 6 lowest force constants, in mDyne \AA^{-1} , for the gas phase conformer of FUQLIM calculated by B3LYP-GD3BJ/6-311G** level of theory. . . .	89
4.2	Number of valid crystal structures that were successfully optimised from a set of 10,000 trial crystal structures for the FUQLIM molecule.	97
4.3	A summary of the success of predictions for polymorphs A, B and C using the principal displacements and torsion flexibility. Results are broken down using different settings whereby (S,O) combinations represent the molecular flexibility dimensions in the search (S) and optimisation (O) procedures. Upon a successful prediction, the total lattice energy above the global minimum, ΔE , (kJ mol^{-1}), $RMSD_{30}$ (\AA), and the number of structure hits, 'Hits', that were found during the search are listed. A hyphen represents 0 Hits and hence an unsuccessful attempt to predict the polymorph.	98

4.4	Comparison of the properties of the observed versus the predicted crystal structures yielded from the torsion angles and principal displacement methodologies for the 3 known polymorphs of FUQLIM.	105
5.1	The number of atoms and rotatable bonds of the molecules in the test set ordered by the extent of molecular flexibility.	116
6.1	ΔE_{strain} values, in kJ mol^{-1} , for CRYSTAL09, $\Delta E_{\text{strain}}^{\text{CRYSTAL}}$, and CrystalOptimizer, $\Delta E_{\text{strain}}^{\text{CrysOpt}}$, DFT approaches for the 26 molecules. In addition, the number of atoms and rotatable bonds per molecule are also presented.	149
8.1	The set of 22 unique conformers of flufenacet within 30 kJ mol^{-1} of the global energy minimum ordered by their relative $E_{\text{conf,bias}}$ values.	189
8.2	Maximum number of principal displacements, and their corresponding force constants, mDyne \AA^{-1} , required to reduce the $RMSD_1$ value between the displaced gas phase and in-crystal geometries to below 0.2 \AA for all 10 molecules possessing 5 rotatable bonds. The molecule in bold text represents the largest minimum force constant value in this set of molecules.	193
8.3	Upper and lower bounds for each principal displacement for the flufenacet conformers 1, 2, and 3.	194
8.4	The relative intramolecular energy, $E_{\text{rel,intra}}$, intermolecular energy, E_{inter} , and the relative total energy, $E_{\text{rel,latt}}$, for the lowest energy crystal structures from flufenacet conformers 1, 2 and 3.	200
8.5	The number of valid crystal structure minimisations for flufenacet conformers 1, 2 and 3.	202
C.1	The 175 molecules listed in Chapter 6, along with the number of atoms and the number of rotatable bonds for each molecule. The molecules highlighted in red, green and grey are those that form non-covalent intramolecular hydrogen, polar and non-polar bonding interactions upon the geometry optimisation of the in-crystal geometry.	231

Chapter 1

Introduction

1.1 Introduction to Crystal Structure Prediction

Any organic molecule can crystallise into a solid form. Owing to the innumerable combinations of the number of atoms, bonding arrangements and atom types that these molecules can comprise of, their corresponding crystal structures yield a wealth of different properties. In addition to this, the same molecule can be energetically stable in multiple crystal structures with differing packing arrangements, a phenomenon called polymorphism. As a result, it is therefore perhaps unsurprising that organic molecular solids are the subject of intensive research that span the sciences [1–5].

It is clear that the wealth of different properties that can be exhibited by crystals that make them so valuable. Currently, no analytical techniques exist that can predict these properties before the crystal has been empirically formed. Therefore, if these properties can be known before the growth of the crystal, then a more rigorous selection process for candidate crystals can be implemented. This will save an immense amount of time, effort and money for both the industrial and consumer parties. To put it simply, if it can be known whether a candidate crystal structure will fail or fulfil its purpose before anyone has even stepped inside a laboratory, this would significantly contribute to reducing these three factors.

Fortunately, the field of crystal structure prediction (CSP) attempts to achieve the goal of predicting the properties of crystals by purely theoretical means. The field itself is rapidly developing and is also now globally recognised by industrial partners. Therefore, it is becoming a necessity to complement experimental findings when designing new products that will need to be delivered to the consumer in crystalline form.

The ideology for CSP is that any scientist (computational chemist or not) can take the structural formula of any molecule and the crystallisation conditions to correctly predict all of the attainable crystal structures and derive the properties of each of them. Of course, the methods employed by CSP are not yet perfect and hence the purpose of this research project is to improve these methods to obtain more accurate and reliable results, whilst keeping the computational costs to a minimum.

Although this research area may seem poignant and of high importance to many industries that make use of crystals, as late as 1988 Maddox [6] described the lack of research in the area of CSP as a ‘scandal’. Whether this article coincided with dawning of the boom in computer power in the 90’s and early 00’s or that simply Maddox was the first to explain this is not clear, however research began in the area of CSP that led to Gavezzotti publishing an article in 1994 entitled “Are Crystal Structures Predictable?” [7].

At this time Gavezzotti simply answered “No”. The reason for this concise response was that he highlighted issues for the, then, current state of CSP. These issues included the predictions of the unit cell parameters, the number of molecules in the asymmetric unit and the calculation of the lattice energy. Whilst each of these issues encompassed its own unique set of problems, a central theme of this review was that the theory was not sufficiently developed; not that the issues were unsolvable. Therefore although the conciseness of the “No” *was* justified in 1994, the critical word in this sentence however is “was”.

CSP has continued its development[8] and attempted to tackle these major obstacles posed by Gavezzotti in his 1994 review. The following years saw CSP research efforts poured into solving these issues that allowed larger, more complex molecules to be studied that pertained to pharmaceuticals [9], salts [10, 11], co-crystals [12, 13] and solvates [14, 15].

As more researchers became active within the field of CSP, so did the number of differing methodologies to tackle this problem [16] (which will be expanded on in detail in Chapter 2). Although these methods differed in theoretical detail, they share the same goal of predicting empirically accurate crystal structures. The success of these methods relies on how accurately the crystal packing environment is replicated; that is the arrangement to which the molecules pack around each other. This phenomenon is determined by the conformation of the molecule in that different molecular ‘shapes’ will pack different with the objective forming the most energetically stable crystal structure possible; not that this is always achieved. Delving further into this discussion, the modelling of the intermolecular and intramolecular forces is tantamount to achieving an accurate molecular conformation through to the replication of the crystal packing environment.

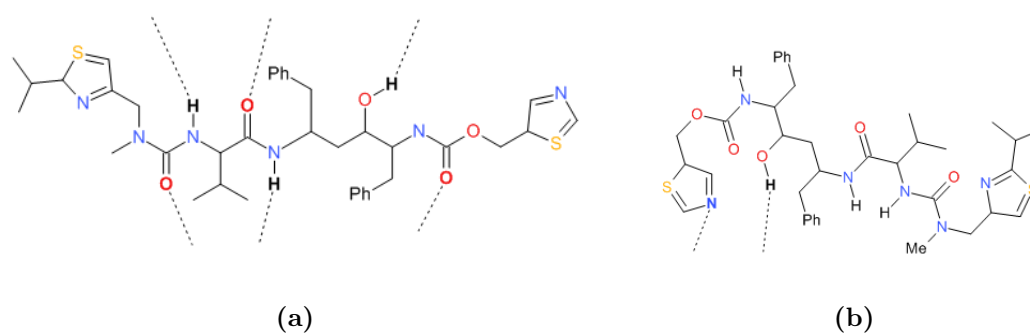


Figure 1.1: (a) Ritonavir A possesses 6 intermolecular hydrogen bonds and has a larger surface area than (b) Ritonavir B that forms a dimer complex and only possesses two intermolecular hydrogen bonds (atoms that are in bold font are those that participate in hydrogen bonding).

The responsiveness of the molecular geometry to flex to perturbations in the external crystalline environment and vice-versa represents a mutually beneficial, symbiotic relationship that seeks to yield the most geometrically accurate crystal structures to their observed counterparts. Since these factors will rely heavily on the energy model that is employed, the Intra-Intermolecular Energy Problem has now been defined to which the research contained in this thesis will be dedicated to working towards a solution to this problem.

1.2 The Intra-Intermolecular Energy Problem: the Motivation for Incorporating Molecular Flexibility into CSP

An important aspect of CSP is the description of molecular flexibility and conformational variability. Therefore, it is clear that this is a major obstacle for CSP especially for molecules that are flexible. In 2002, Gavezzotti outlined several major issues that CSP needs to overcome [17, 18] and, amongst others, the intra-intermolecular energy problem was presented. This issue is defined by how the intramolecular energy affects and is affected by the intermolecular environment, and vice versa.

Prior to the commencement of this research project, no large scale studies that specifically focus on flexible molecules had been conducted. However, there are a number of specific studies that were successful including glycol and glycerol [19], a study of aspirin [20] and also piracetam [21] where each study modelled the molecular as flexible and produced a crystal structure that provided a promising match, albeit not an exact match, to its empirical counterpart. Details of the specific methods are not mentioned here as to do so would only serve to confuse matters. The theoretical details will be

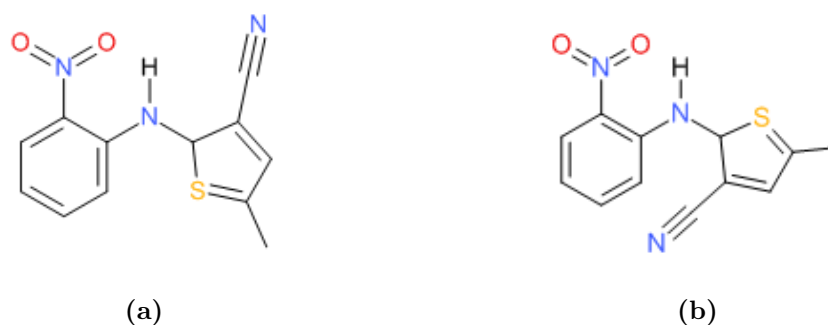


Figure 1.2: (a) and (b) show the 2 molecular conformations of ROY.

outlined in Chapter 2. A more thorough discussion on specific case studies will now commence. The first of these studies focusses on Ritonavir [22], Figures 1.1a and 1.1b, which is a well-recognised example of when molecular flexibility can lead to issues and begins to justify the need for more advanced methods to treat molecular flexibility in CSP. Ritonavir A is a more thermodynamically stable polymorph than Ritonavir B as there is more opportunity for intermolecular bonds to form, in this specific case these exist as hydrogen bonds. This reduces the solubility of Ritonavir A, causing it to have a lower bioavailability than Ritonavir B.

Owing to an inadequate polymorph screening procedure, Ritonavir A was not a known polymorph at the time of taking the drug to market. During the manufacturing process, Ritonavir A began to crystallise instead of Ritonavir B and eventually led to the temporary withdrawal of the product from the market. Two issues arise here in that the patient does not experience the benefits from the product and the pharmaceutical company loses money and resources. However, it could be argued that this was the first materialistic caveat that demonstrated the importance of including flexibility for molecules of this size and level of complexity.

The second case study is another well known example of ROY [23, 24], a flexible molecule that possesses 2 molecular conformations, Figure 1.2, that together yield a total of 9 different polymorphs, with 2 of these polymorphs being discovered 5 years after the original synthesis of ROY had occurred [25]. Therefore the omission of either conformation would ‘miss’ polymorphs and render the set incomplete. CSP calculations have been conducted on ROY [26] that found 5 of the 7 selected polymorphs involved in the study in the final list of crystal structures. This study then highlights the need for improvements to the models that treat molecular flexibility in crystals as 2 of polymorphs yield inaccurate approximations to their observed counterparts. Molecular flexibility therefore plays a crucial role in allowing the ROY molecule to become exceptionally polymorphic. Therefore these CSP calculations would need to include both molecular conformations and also allow the molecules to flex in response to the crystalline environment.

Although it is easy to say, with the benefit of hindsight, that a CSP study with the treatment of flexibility would have foreseen the catastrophe of Ritonavir or the polymorphism of ROY, but it would have certainly reduced the time taken to uncover these phenomena. Therefore molecular flexibility must be considered at all stages of the CSP process for the reliable and efficient study of the molecular packing arrangement within a crystal structure. This therefore will provide the highest probability of the molecular geometry within the crystal structure being correct and hence yield a complete and scientifically accurate set of crystal structures.

The accuracy of modelling the molecular geometry is a pre-requisite for a successful CSP but it is not known in advance what this molecular geometry will be. The potential differences in gas phase and in-crystal molecular geometries have been discussed in the literature [27] which highlights the importance of molecular flexibility within crystals to all parties involved within solid-state chemistry or physics; not merely just expressing it as a challenge within CSP. Quantifying these differences in molecular geometries is a well recognised issue that will be more stringently outlined in Section 2.5.4.2. In addition, methodologies to solve this issue have been developed whose merits will also be discussed in detail.

The inclusion of molecular flexibility within CSP indeed gains more accurate representations of crystal structures but adds a more detailed level of complexity and computational expense to the calculations. This effectively highlights the centrepiece of computational chemistry, where the three-way marriage of accuracy, complexity and computational cost constitutes the ultimate goal of maximising the former and minimising the latter two. However, all three have a direct relationship and the variation in one of these components will affect the other two. The art of computational chemistry is to obtain the optimal balance of these three factors. So whilst incorporating molecular flexibility is important, keeping the methods simple and minimising any increases in computational cost is also paramount.

1.3 Thesis Overview

The main focus of this PhD research project is developing methods to further incorporate molecular flexibility into CSP. Therefore, it is logical to start with Chapter 3 that attempts to quantify the inclusion of molecular flexibility by performing CSP calculations in a set of relatively small molecules that possess limited flexibility. These are simple organic molecules that the current CSP process can easily handle and allows the flexibility aspect to be more isolated. In addition, this chapter also focusses on benchmark testing a revised energy model using the same set of molecules.

Chapter 4 begins to develop a new technique to treat molecular flexibility by taking a molecule where a previous CSP attempt had failed to find 2 of the 3 known polymorphs. This chapter also analyses why the current standard of treating molecular flexibility in CSP is sometimes inadequate and provides a proof of concept discussion to suggest using molecular principal displacements to sample the molecular geometry, whilst simultaneously using conventional methods to search the crystal packing environment.

Chapter 5 attempts to generalise this method to larger molecules and discusses several novel approaches to converting the gas phase geometry into the in-crystal geometry by a selection of the principal displacements of the former. This shows the progression of ideas to finally settle on the method that is the most scientifically robust, but also the most computationally efficient.

Chapter 6 calculates the energy differences between the gas phase and in-crystal geometries for a 224 molecule test set and then implements the chosen method from Chapter 5 to obtain a set of principal displacements that performs the geometry interconversion. The chapter then identifies trends in which principal displacements are commonly occurring and attempts to relate this to the properties of the molecule and the force constants of the principal displacements.

Chapter 7 presents a concise CSP case study that is the Author's contribution to the sixth blind test and discusses its merits and shortcomings.

Finally, Chapter 8 then performs another CSP case study that implements the work from Chapters 4, 5 and 6 to perform a 'pseudo-blind test' on a molecule from experimental collaborators to which the observed crystal structure(s) were not known at the time this research commenced. This provides a unique opportunity to make predictions as to which principal displacements would be required to feed into the method outlined in Chapter 4. The chapter then draws comparison as to which level of methodology (a simple rigid molecule CSP or a the state-of-the-art flexible molecule CSP utilising the work from this thesis) was adequate enough to predict the crystal structures of the molecule, if at all.

1.4 A Note on Software

All data analysis and figures in this thesis were both conducted and produced using the Mathematica¹⁰ [28] software package. All skeletal formulae and molecular diagrams were produced using Symyx Draw 3.3. All searches of the Cambridge Structural Database (CSD) [29, 30] were conducted using Conquest 1.16 [31] and Mercury 3.7 [32] was used for

crystal comparisons and illustrating 3-dimensional (D) images of molecules and crystal structures.

1.5 A Note on Nomenclature

The 3D molecular formulas through this thesis use different colours to represent different atom types. The atomic colour scheme remains constant (unless explicitly stated otherwise) throughout the report: carbon (grey), nitrogen (mauve), oxygen (red), hydrogen (white), chlorine (green), fluorine (lime green) and sulfur (yellow).

There are many other software packages that have been implemented which are specific to this research and these will be referenced appropriately in the text.

1.6 A Note on Computational Resources

All calculations presented in this thesis were performed either on the Day Group's single workstation equipped with an Intel core i7-3770k 3.5 GHz quad-core processor or on the IRIDIS4 High Performance Computing Facility supercomputer at the University of Southampton. The latter is a Red Hat Enterprise GNU/Linux Beowulf cluster with 750 compute nodes, each with two 8-core Intel Xeon E5-2670 2.6 GHz processors. The nodes share a common global IBM GPFS file system over an FDR InfiniBand interconnect.

Chapter 2

Theory

2.1 The CSP Process

Though many different variations of the CSP process exist depending on the type of crystals that are being studied [16], a general process remains constant. The following sections describe in detail the CSP process when firstly using a rigid molecule approximation before discussing the techniques that currently exist to treat molecular flexibility within that particular stage of the CSP process. An overview of this process is illustrated in Figure 2.1 and each stage will now be discussed in detail.

2.2 Molecular Conformational Searches

A CSP study is usually initiated by a set of Cartesian coordinates for the target system that defines the atom positions and atomic types within a molecule. However, it is also not uncommon to initiate the process simply from a 2D molecular sketch of a molecule, or even 1D molecular notation such as a SMILES string or IUPAC name.

To understand how the methods that have evolved for the conformational search procedure, one must first appreciate the theoretical basis that describes molecular flexibility.

2.2.1 Molecular Flexibility

2.2.1.1 Internal Degrees of Freedom

Traditionally, molecular geometries are represented by a set of Cartesian coordinates for each atom. A distance matrix is created and compared to a dictionary of typical bond

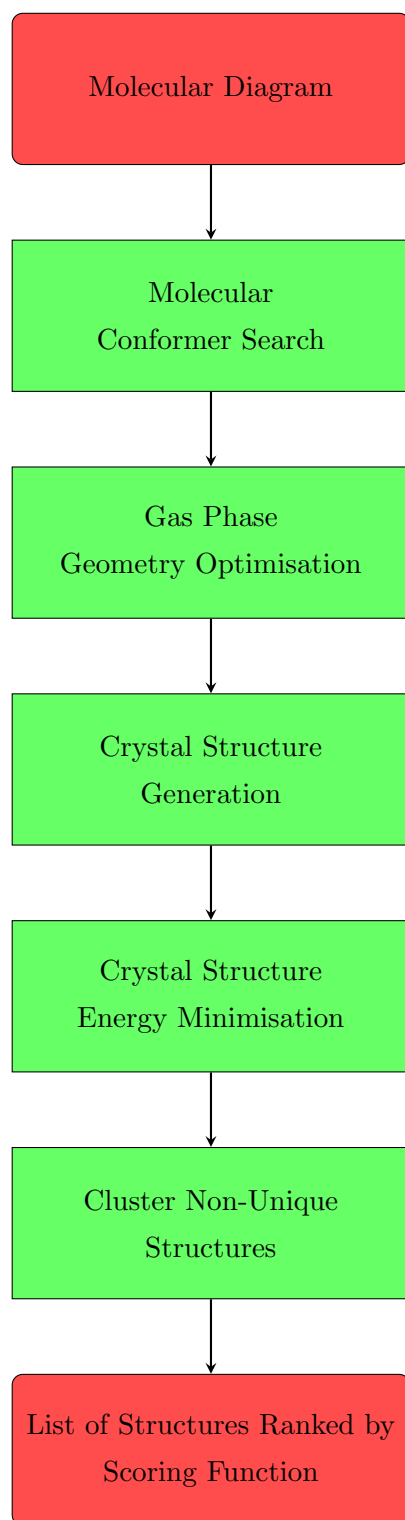


Figure 2.1: CSP flowchart that shows the fundamental steps of the rigid molecule CSP process. Note the intramolecular geometry is fixed after the gas phase geometry optimisation stage.

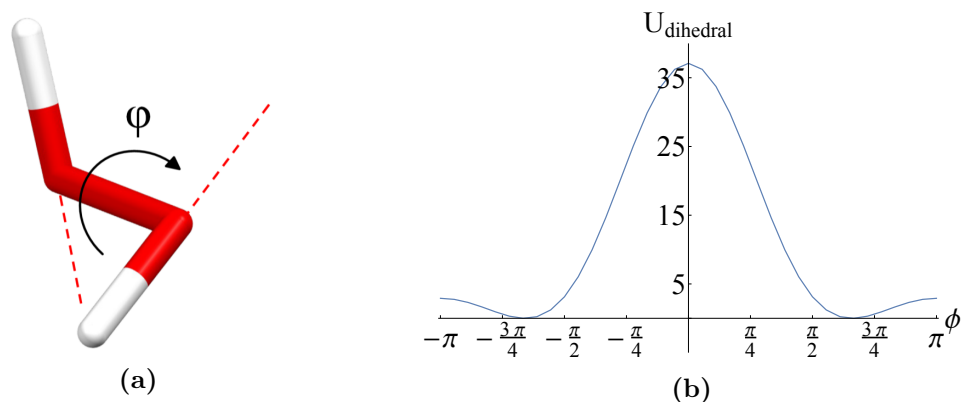


Figure 2.2: (a) Gas phase geometry of hydrogen peroxide with ϕ at approximately $\pm\frac{2\pi}{3}$ radians. The dashed, red lines are included to aid in defining the 3D geometry of the molecule in a 2D space. (b) shows the internal energy, U_{dihedral} in kJ mol^{-1} of hydrogen peroxide as a function of ϕ between $-\pi$ and π radians.

lengths to provide information on the connectivity of the molecule; information that is stored in a connectivity matrix. However, one can also define a molecule from its internal degree of freedom (DOF)s with only using the distance and connectivity matrices and hence excluding the Cartesian coordinates.

These DOFs are partitioned into 3 types: bond distances, bond angles and torsion angles. The former refers to the interatomic distance between two bonded atoms. A bond angle is defined as the angle formed between three adjacently bonded atoms. A torsion angle is formed from four adjacently bonded atoms where the first and last three atoms' coordinates define two planes to which the angle between these is the dihedral angle. Figure 2.2a illustrates this point for the trivial case of hydrogen peroxide with ϕ representing the dihedral angle. Every non-linear, isolated molecule can be defined with $3N - 6$ of these DOFs where N is the number of atoms in the system. This derives from every atom possessing 3 translational DOFs along each Cartesian axis ($3N$) and then the 3 molecular translational and 3 rotational DOFs are subtracted hence yielding the '-6'.

Using a molecular mechanics approach, the potential energy, U , is a function of the distance of a given DOF, x , from its equilibrium position and possesses a force constant, k , that describes the magnitude of the force exerted through this movement defined by:

$$U(x) = \frac{kx^2}{2} \quad (2.1)$$

where k is the force constant. Therefore, k can be expressed as the second derivative of the potential energy with respect to the displacement of the DOF:

$$k = \frac{\partial^2 U}{\partial^2 x}. \quad (2.2)$$

However, delving more specifically into the Equation 2.1, the potential energy can be calculated for each type of DOF. Bonds and bond angles are defined synonymously:

$$U_{\text{bond}}(r) = k(r - r_{\text{eq}})^2 \quad (2.3)$$

$$U_{\text{angle}}(\theta) = k(\theta - \theta_{\text{eq}})^2 \quad (2.4)$$

where r and θ represent the bond distance and bond angle, respectively, with ‘eq’ denoting the equilibrium position. The potential energy for the dihedral takes a different form [33]:

$$U_{\text{dihedral}}(\phi) = \sum_{n=1}^D \frac{V_n}{2} [1 + \cos(N\phi - \gamma)] \quad (2.5)$$

where V_n represents the height of the N^{th} cosine term (N^{th} energy barrier), D represents the number of cosine functions included in the torsion energy profile in one period of rotation, ϕ is the dihedral angle and γ is the angular offset.

In addition, a special case of torsion angle, an improper torsion, can be defined that describes the angle formed by an out-of-plane bend. A simple example of this would be the bend of a O atom out of the carboxyl plane. These types of torsion angles are more uncommon than ‘regular’ torsion angles but nonetheless need to be modelled by a variation of Equation 2.5:

$$U_{\text{improper}}(\phi) = V(1 - \cos 2\phi). \quad (2.6)$$

Figure 2.2 presents a trivial example of using a ‘regular’ torsion by visualising hydrogen peroxide to illustrate Equation 2.5. Figure 2.2b shows a maximum in U_{dihedral} when ϕ equals 0 as the two hydrogen atoms in the peroxide molecule are eclipsed. The two U_{dihedral} minima occur at approximately $\pm \frac{2\pi}{3}$ radians. These are the conformations in which the hydrogen atoms are positioned between the opposing hydrogen and the lone pairs of electrons on the oxygen atoms. This is before increasing slightly due to the H-O bonding electrons interacting with the lone pairs of electrons on the opposing O atom.

The profile of the torsional scan for more complex molecules can be complicated by intramolecular interactions and by steric effects about the bond formed between the two central atoms of the dihedral angle. This is the case for any bond possessing a bond order > 1 (a trivial case for this would be ethene, where the dihedral angles are rigid due to the π -bonding between the two carbon atoms).

2.2.1.2 Molecular Principal Displacements

Molecular principal displacements are derived from a $(3N \times 3N)$ Hessian matrix, \mathbf{H}_{ij} , where each element is the second derivative of the molecular energy, U , with respect to atomic displacements, x_i and x_j , in Cartesian coordinates about the equilibrium geometry, 0:

$$\mathbf{H}_{ij} = \left(\frac{\partial^2 U}{\partial x_i \partial x_j} \right)_0 \quad (2.7)$$

where each element, under a simple molecular mechanics approximation, is closely related to Equation 2.2 and therefore represents a force constant of the combination of atomic displacements, x_i and x_j .

The diagonalisation of \mathbf{H}_{ij} yields a set of orthogonal displacements that are referred to as principal displacements. Each principal displacement possesses a unique force constant that is a corresponding eigenvalue of \mathbf{H}_{ij} . This allows the set of principal displacements to be energy ordered where the lower and higher value force constants will perform distortions that will more probably affect the dihedral angles and bond lengths of the molecule, respectively.

Additionally, the normal modes of a molecule are more commonly referred to in the literature as they bear more spectroscopic significance. This is due to the fact that normal modes are derived from a dynamical mass-weighted Hessian matrix, $\mathbf{H}_{\text{mw}ij}$ that is completely analogous to Equation 2.7 and is defined as follows:

$$H_{\text{mw}ij} = \frac{H_{ij}}{\sqrt{m_i m_j}} = \left(\frac{\partial^2 U}{\partial x_i \partial x_j} \right)_0 \quad (2.8)$$

where the absolute atomic masses, m_i and m_j , are now included.

However, for the purposes of this research, the absolute atomic masses will be ignored and therefore only Equation 2.7 will be used instead of Equation 2.8. This is because the CSP procedure that is implemented, as we shall see in later sections, is required for energy minimisation techniques. This allows the calculations to isolate the pure electronic contribution of the principal displacements to the lattice energy. A detailed discussion pertaining to the inclusion of the absolute atomic masses can be found here [34].

The eigenvectors of each eigenvalue (force constant) from the non-mass-weighted \mathbf{H}_{ij} then affords $3N - 6$ sets of $3N$ normalised Cartesian atomic displacements. Any non-linear molecule will possess $3N - 6$ principal displacements (the ‘ -6 ’ derives from the subtraction of the translation and rotation DOFs as they do not affect the internal geometry of the molecule).

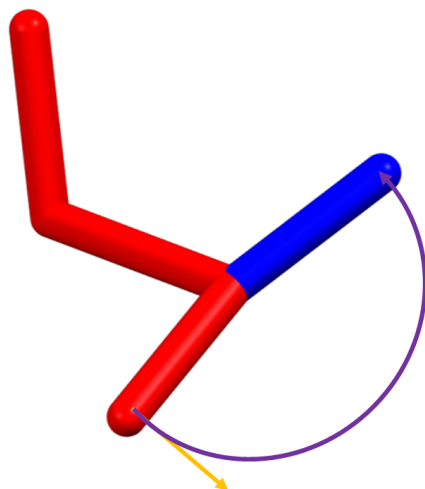


Figure 2.3: The displacement of hydrogen peroxide along its lowest energy principal displacement. The red and blue geometry refer to the equilibrium and displaced geometries of hydrogen peroxide, respectively. The orange and purple arrows show the path of the hydrogen atom as it is displaced along this principal displacement in linear and curvilinear space, respectively.

Each principal displacement of a molecule is often given as a set of vectors on each atom. These vectors are the directions of motion for each atom, but are only valid at infinitesimal displacement values. Large displacements along these linear vectors will cause large, positive changes in the molecular energy. However, if interpreted in curvilinear space, they will afford much smaller energy penalties than the linear equivalent.

2.2.1.3 Linear versus Curvilinear Space

Although the eigenvectors of the Hessian matrix are given in Cartesian coordinates, when performing molecular displacements along a set of principal displacements (see Section 2.2.1.4) the molecular geometry must be modified via changes in its internal DOFs; not along a linear path in Cartesian space. Therefore the starting set of Cartesian coordinates must be represented in terms of the $3N - 6$ internal DOF, the principal displacement vectors must be converted into changes in these internal DOFs and then once these two components have been combined appropriately, it is desirable for the final geometry to be converted back into Cartesian coordinates. These coordinate conversions may seem tedious and unnecessary as one can argue that, for small displacements, the start and end points could be, approximately, identical. However, this is certainly not the case for molecular distortions typically found in the simulations performed in this research. Figure 2.3 illustrates this point for the trivial example of hydrogen peroxide being displaced along its lowest energy principal displacement by an arbitrary amount in linear (orange arrow) and curvilinear (purple arrow) space. For the purposes of this example, we can approximate that the lowest energy principal displacement is a

pure rotation about the dihedral angle, ϕ . Figure 2.3 shows that the displacement of a molecule in linear space will force the geometry into an ‘un-physical’ state that will eventually result, at least in this example, in the breaking of the H-O bond.

2.2.1.4 Propagating Along Molecular Principal Displacements

Displacing a molecule along a (or a combination of) its principal displacement(s) in curvilinear space requires the prerequisite of a computationally efficient method of converting Cartesian forces or displacements into internal coordinates. This has been the subject of many articles in the literature [35–38]. This work is used but set in the context of motions along the principal displacements.

The principal displacement eigenvectors can be expressed by means of a $(3N \times 3N - 6)$ matrix, \mathbf{C} , where each column represents a principal displacement and each row of that column refers to the Cartesian displacement of an atom N in the x, y and z dimensions:

$$\mathbf{C} = \begin{pmatrix} pd_1x_1 & \dots & pd_{3N-6}x_1 \\ pd_1y_1 & \dots & pd_{3N-6}y_1 \\ pd_1z_1 & \dots & pd_{3N-6}z_1 \\ \vdots & \ddots & \vdots \\ pd_1x_N & \dots & pd_{3N-6}x_N \\ pd_1y_N & \dots & pd_{3N-6}y_N \\ pd_1z_N & \dots & pd_{3N-6}z_N \end{pmatrix}. \quad (2.9)$$

This matrix is not square because the translational and rotational principal displacements have been removed.

The subtleties of linear versus curvilinear space have already been highlighted (Section 2.2.1.3) yet a solution was not presented. Fortunately, ideas proposed by Wilson in 1955 [39] and then utilised by Baker [36] provide a solution to this problem using the infamous B-matrix, \mathbf{B} . More specifically, \mathbf{B} is comprised of all internal DOFs that can possibly be defined. This will therefore contain redundant DOFs.

Relating this back to the trivial example of hydrogen peroxide, an oxygen hydrogen bond will be defined by both O₁-H₁ and H₁-O₁ variations. These two DOFs will possess the same value and are therefore two different ways of labelling the same DOF. A bond angle will be defined by H₁-O₁-O₂ and O₂-O₁-H₁ and the one dihedral angle in hydrogen peroxide will be defined by both H₁-O₁-O₂-H₂ or H₂-O₂-O₁-H₁ variations. These duplicate definitions become in greater number the more atoms a molecule possesses. Nonetheless, this forms a redundant set of DOFs and provides a good starting point for defining the DOFs of a molecule.

The removal of these duplicate DOFs (yielding a non-redundant set of internal DOFs) can be achieved by taking the set of non-zero eigenvectors, \mathbf{K} , from the Wilson G-matrix defined as:

$$\mathbf{G} = \mathbf{B} \cdot \mathbf{B}^T \quad (2.10)$$

and then performing a dot product of \mathbf{K} and the original \mathbf{B} :

$$\mathbf{B}_{\text{nr}} = \mathbf{K} \cdot \mathbf{B} \quad (2.11)$$

that finally yields a $(3N-6) \times (3N-6)$ matrix, \mathbf{B}_{nr} , that now possesses a non-redundant, ‘nr’, set of internal DOFs. \mathbf{K} acts to combine DOFs in an efficient manner to remove these redundancies. \mathbf{B}_{nr} allows the conversion of the Cartesian forces and displacements, \mathbf{C} . More specifically, these displacement vectors within \mathbf{C} are then normalised and measured in units of Å. This leads to changes in internal coordinates:

$$\vec{Q} = \mathbf{B}_{\text{nr}} \cdot \mathbf{C} \quad (2.12)$$

that affords a \vec{Q} vector, of length $3N - 6$, that represents the change in internal DOF from a starting set of Cartesian displacements.

The multiplication of the columns in \mathbf{C} by an arbitrary amount before taking the dot product with \mathbf{B}_{nr} can yield larger distortions along chosen principal displacements and smaller, if any, distortions along other principal displacements. Using this logic, the $3N-6$ vector, \vec{s} , can be introduced to Equation 2.12, whose elements dictate the quantity of the corresponding principal displacement in \mathbf{C} to be used:

$$\vec{Q}(\vec{s}) = \mathbf{B}_{\text{nr}} \cdot \vec{s} \cdot \mathbf{C}. \quad (2.13)$$

The elements in \vec{s} are chosen by the user and now yields the $\vec{Q}(\vec{s})$ vector which is the quantity that each internal DOF will change for a given displacement along one, or a set of, principal displacement(s).

The mapping of $\vec{Q}(\vec{s})$ onto a current set of internal coordinates that define a starting (base, b) geometry of a molecule, q_b , is performed by a simple addition:

$$\vec{q}_t = \vec{q}_b + \vec{Q}(\vec{s}) = \vec{q}_b + \mathbf{B}_{\text{nr}} \cdot \vec{s} \cdot \mathbf{C} \quad (2.14)$$

where \vec{q}_t is the set of DOFs that define the final geometry (target, t) of the molecule. If just a pure set of internal coordinates is required, then the calculation is complete. However, it is often desirable to convert this set of internal coordinates back into a Cartesian set of coordinates.

Owing to the non-linear transformation from Cartesian to internal coordinates, the original \mathbf{B} cannot be used to simply convert the internal coordinates back to Cartesian coordinates as \mathbf{B} changes as the molecular geometry changes. Therefore an iterative procedure is implemented that updates \mathbf{B} with every cycle to circumvent this issue [36]:

$$\vec{X}_{\text{cart}}(k+1) = \vec{X}_{\text{cart}}(k) + ((\mathbf{B}^T)^{-1})^T(k) \cdot (\vec{q}_t - \vec{q}_t(k)) \quad (2.15)$$

where $\vec{q}_t(k)$ is the set of internal coordinates generated from the k^{th} iteration, $\vec{X}_{\text{cart}}(k)$ is the Cartesian set of coordinates generated from the k^{th} iteration and $\vec{X}_{\text{cart}}(k+1)$ is the Cartesian set of coordinates generated from the $k+1$ iteration. This iterative procedure terminates when $|\vec{q}_t - \vec{q}_t(k)|$ reduces below a chosen tolerance that is recommended by Baker [36] to be 10^{-10} . However, from the experience of the author, this tolerance was set to 10^{-6} and this procedure then typically converges within 5 cycles.

2.2.2 Method for Searching Molecular Conformational Space

Over a series of papers, Day *et al.* [40–43], developed a method that implemented a potential energy scan about the selected DOFs as a means of identifying stable conformers. This method commences by fixing the n user-selected DOFs before performing a density functional theory (DFT) calculation (see Section 2.3.1) such that the remaining DOFs are able to relax to their equilibrium positions. The n DOFs then form an n -dimensional space that is searched using a grid based method. The molecular energy is evaluated at each grid point and, assuming a fine enough grid has been used, the regions between points can be accurately interpolated to form a potential energy surface. Although this method has produced positive outcomes [40, 44], it is bound by the ‘curse of dimensionality’ [45] and the expensive DFT calculations that it must perform.

An equally valid approach begins by performing a global search for all local minima on the U_{intra} potential energy surface (PES). One implementation of this exists in the MacroModel [46] software package where a low-mode conformational search (LMCS) is implemented as several principal displacements are propagated along simultaneously [47, 48]. As this space is explored force-field geometry optimisations are performed (common force-fields include OPLS [49, 50] and AMBER [51] whose merits can be found elsewhere [52]). These force-field geometry optimisations are not as accurate as when using DFT but are significantly cheaper with respect to computational cost.

Similar logic can be applied to searching this space as was highlighted in Section 2.4 in that it is desirable that all molecular conformers are not just found, but found multiple times post-geometry optimisation. This ensures that the search space has been well sampled and all conformers have indeed been found.

2.3 Gas Phase Geometry Optimisation

All of these unique conformers require a more accurate molecular geometry than that was obtained from the conformational search. However, a scoring function can be applied at this stage, even when using a low level of accuracy, to remove any conformers that rank poorly and remove a significant proportion of the computational cost. This cropped set of molecular conformers now forms the foundation of modern rigid CSP methodologies.

However, the accuracy of the geometries pertaining to the conformers are insufficient as the electron density is extremely sensitive to minute distortions in the atomic positions [53]. Therefore a set of highly accurate atomic positions must be obtained to gain an accurate description of the electron density. To obtain this and to keep computational costs to a minimum, the well-established DFT methodology will be applied.

2.3.1 Density Functional Theory

Electronic structure theory describes the distribution and motion of electrons about atomic nuclei. The mandatory duty of electronic structure methods is to solve the time-independent Schrödinger equation (2.16) to obtain the molecular or atomic energy:

$$\hat{H}\Psi_{\text{total}} = E\Psi_{\text{total}} \quad (2.16)$$

where \hat{H} is the Hamiltonian operator, E is the energy and Ψ_{total} is the total wavefunction that contains the individual nuclear and electronic Ψ components. \hat{H} corresponds to the total energy of the system which is partitioned into kinetic and potential energy. Ψ_{total}^2 describes the probability of finding a particle within a given region of space.

The Ψ_{total} term can be decomposed into the nuclear and electronic Ψ 's by implementing the Born-Oppenheimer approximation [54]. This approximation states that because the nuclei are so much more massive than the electrons, the nuclear kinetic energy can be neglected in the first instance which allows each ψ to be solved individually:

$$\Psi_{\text{total}} = \psi_{\text{electronic}} \times \psi_{\text{nuclear}}. \quad (2.17)$$

Only $\psi_{\text{electronic}}$ is considered in this research as the molecules are considered as static entities. This is a many-body problem in that there are many interacting particles within the molecular system. This prevents the quantum mechanical calculations being solved exactly for systems containing more than one electron. In CSP, this is always the case so the methods required for calculating the energy must be approximated [55].

One method for calculating the electron density ($\psi_{\text{electronic}}^2$) of a system is DFT which uses functionals to ascertain an approximated total energy. More specifically, these functionals use the electron density, ρ . The Kohn-Sham method [56] creates a fictitious system which is identical to the given system but excludes the electron-electron repulsion interactions. This allows the total energy, E_{DFT} , to be partitioned into more manageable pieces and an explicit term is given to describe these electron-electron repulsion energies, $E_{\text{ee}}[\rho]$:

$$E_{\text{DFT}}[\rho] = E_{\text{nn}} + T[\rho] + E_{\text{ne}}[\rho] + E_{\text{ee}}[\rho] \quad (2.18)$$

where E_{nn} describes energy of the inter-nuclear interactions, $T[\rho]$ is the total kinetic energy and $E_{\text{ne}}[\rho]$ is the nuclear-electron interaction energy. $E_{\text{ee}}[\rho]$ is then further partitioned into:

$$E_{\text{ee}}[\rho] = J[\rho] + E_{\text{xc}}[\rho] \quad (2.19)$$

where $J[\rho]$ is the Coulomb operator that derives an average local potential at ρ and $E_{\text{xc}}[\rho]$ is the exchange-correlation energy. Now only the latter term is not exactly solvable but many functionals have been derived to tackle this issue [57].

Local-density approximation functionals are computationally cheap and widely used that evaluate the electron density at a given point in space [58]. However, this set of functionals assume that the electron density is constant throughout the system and hence overestimates the exchange-correlation energy. To correct for this, a generalised gradient approximation breeds another class of functionals [59] that now include the gradients (first derivative) of the electron density to yield more accurate results than the local-density approximation functionals [60]. Going one step further, meta-generalised gradient approximation functionals also take into account the second derivative of the electron density and are parametrised on from benchmark databases [61]. However, these functionals are considered to be over-fitted leading to inaccurate results, particularly with intermolecular interaction energies, when tested on systems that were not included in the training set [62].

Hybrid functionals are those that incorporate a portion of exact exchange-correlation obtained from *ab initio* or empirical sources. This set of functionals allow exact exchange energies to become incorporated into DFT and still keep the computational cost lower than when using *ab initio* methods. The most well established DFT hybrid functional is the B3LYP (Becke exchange, three parameter, Lee, Yang, Parr correlation) [63, 64] functional that delivers comparable accuracies to *ab initio* methodologies. Other hybrid functionals are used throughout CSP [65] but only B3LYP will be implemented in this work unless stated otherwise as this functional offers the most accuracy for the minimal computational cost.

2.3.2 Basis Sets

Despite using the ρ , a description of the molecular orbitals is still required which are represented in the form:

$$\psi_i(r) = \sum_{\alpha=1}^M B_{\alpha} \cdot C_{\alpha i} \quad (2.20)$$

where ψ describes the molecular orbital represented as a sum of the M atomic orbitals, B is the basis function that runs over orbital α to orbital M and C is the orbital coefficient which is determined by the DFT functional such that the electronic energy is minimised. B can take many different forms [66] but only Gaussian basis functions will be considered here.

The atomic orbitals can be separated into the radial and angular parts. The angular part of the atomic orbitals determine the orbital angular momentum, l , of the orbital and hence determines its shape (s , p , d or f).

A Slater function is an ideal description of the radial part of the atomic orbital, but proves computationally expensive to use in electronic structure calculations. Therefore a set of Gaussian functions allows a suitable approximation to be made. The generalised functional forms of the Slater and Gaussian functions are presented in Equations 2.21 and 2.22, respectively:

$$s(r) = e^{-ar} \quad (2.21)$$

$$g(r) = e^{-ar^2} \quad (2.22)$$

where r is the nuclear-electron distance and a is a constant. A sum of Gaussian functions are used to approximate a Slater function where more Gaussian functions provide a closer approximation. This is illustrated in Figure 2.4.

Split-valence or Pople basis sets are sets of Gaussian function that describe the atomic orbitals of a molecule that combine to form the molecular orbitals. They are denoted by names of the form $L-MNG$ where L , is the number of Gaussian functions, G , describing the core orbitals and M and N two basis functions describing the valence orbitals; one consisting of M and one consisting of N Gaussian functions.

Polarisation functions (denoted by asterisks, ‘*’) can be added to these basis functions which allows for shifts of electron density about atomic positions which aid in the description of polar bonds. To accomplish this, polarisation functions include functions of higher angular quantum numbers than those of the electrons residing about a given nucleus. This allows for the electron density to be more flexible and therefore can more accurately describe its asymmetry about their nuclei. Diffuse functions (denoted by plus

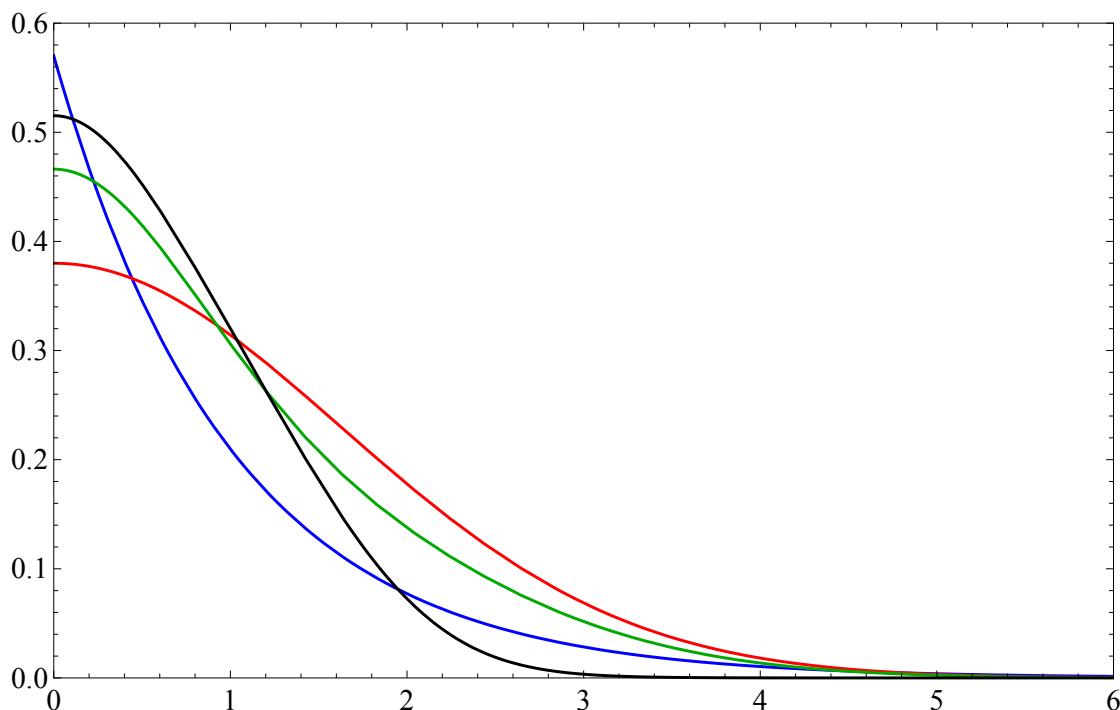


Figure 2.4: A Slater and a series of Gaussian functions, showing the approximation of using 1, 2 and 3 Gaussian functions in red, green and black, respectively, to approximate a single Slater function in blue.

symbols, ‘+’) can also be added to account for orbitals with a particularly large spatial extent.

An example of one of these basis sets is 6-31++G** that denotes one set of six Gaussian functions that describe the core molecular orbitals, 2 sets of 3 and 1 Gaussian functions that describe the valance molecular orbitals. The diffuse and polarisation functions are also included and are applied to all atoms in the molecule.

Due to the nature of these Gaussian type orbitals, all implementation of DFT within this work will be performed using Gaussian09 software package [67]. The advantage of using this software is that it allows the incorporation of both semi-empirical and post-Hartree-Fock methods into the DFT calculations. This in contrast to other well established software packages that implement a Gaussian type orbitals approach such as CRYSTAL09 [68] and MOLPRO [69] that do not allow for semi-empirical methodologies to be incorporated into the DFT calculations.

2.3.3 Minimisation of the Molecular Potential Energy

To obtain an accurate description of the molecular geometry and molecular potential energy, this DFT methodology is employed. This will seek to minimise the potential

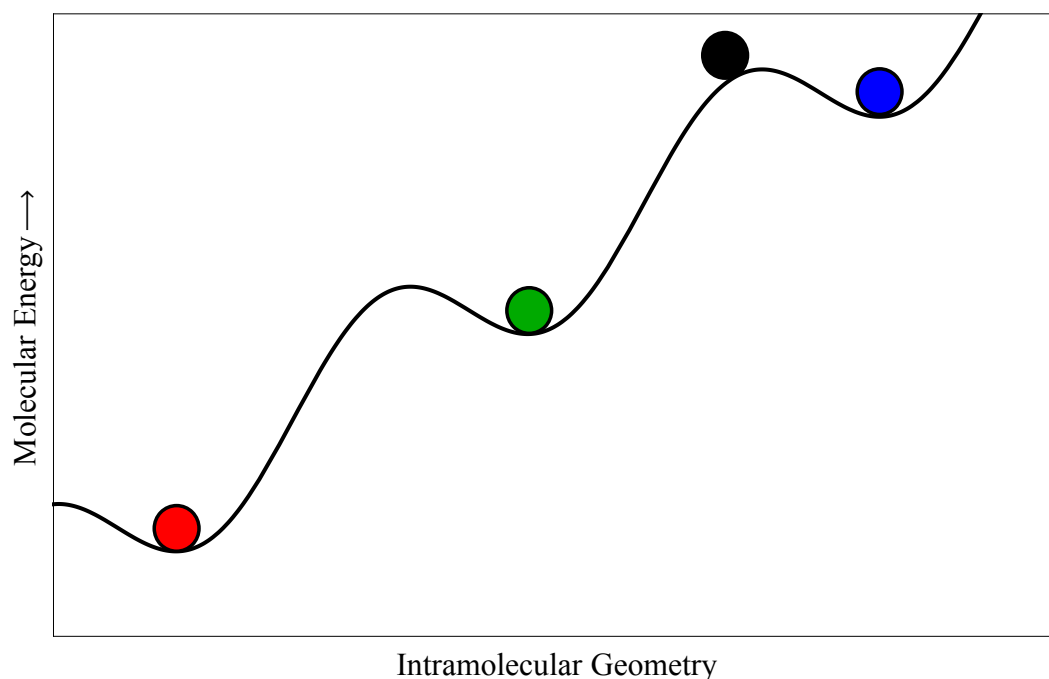


Figure 2.5: An example of a simple potential energy surface plotted as the molecular energy as a function of molecular displacements. If the black point is the energy of the starting point, the green point exists as the post-minimisation energy. The blue point is the closest geometry minimum to the starting point and the red point is the global potential energy minimum.

energy of the system with respect to the molecular geometry until a local minimum (or saddle point) is reached on the PES. A simple PES is illustrated in Figure 2.5 and whose features will now be described.

Firstly, if the black point represents the energy of the starting geometry, then the final geometry will be the green point as that is the local minimum. The blue point is the closest geometry to the starting geometry but, as a potential energy maximum lies between these two geometries, it will not be discovered (unless the initial search direction and step size of the minimisation algorithm steps over, thus ignoring, the separating maximum). Another observation is that the red point lies at the global minimum. However, this will not be found in this scenario if we assume a local energy minimisation approach. Therefore this method can ‘miss’ the global minimum.

Local minimisation missing global minima can be an issue for large molecules with many DOFs which will possess a high dimensional PES. This is discussed further in Section 2.2.

Many different methods exist for performing numerical minimisation. The conjugate gradient method [70] provide an accurate route for minimisation but proves computationally expensive due to the calculation of the gradients in order for it to determine the direction of the next step along the PES. Therefore, as computational efficiency is

paramount for this CSP process, these types of minimisation algorithms will be avoided. Algorithms where gradients are not required in the Powell method [71] that performs a simple bi-directional search along each dimension to be minimised in turn which aids for a efficient procedure. However, this method assumes complete independence between these dimensions and therefore does not allow previously optimised dimensions to respond to fluctuations in the another, hence making the Powell method very selective on the type of functions to be minimised. In relation to CSP, variations in a particular DOF will exhibit some affect on all of the other DOFs the molecules possesses. Hence the Powell method is not suitable for this type of problem. The simplex minimisation algorithm [72] does not require the gradient information but does allow for these revisions of the values for each dimension in response to these fluctuations. The essence of the simplex algorithm is to construct a simplex of $n + 1$ dimensions and trial the value of the objective function at each vertex which will then move to the point with the lowest value. This now takes into account of all dimensions simultaneously and hence provides a robust methodology to perform a numerical minimisation. Hence the simplex minimisation methodology will be used for all instances of minimisation of the energy throughout this thesis unless explicitly stated otherwise.

2.3.4 Molecular Electrostatic Potentials

Once an accurate description of the molecular geometry and energy have been obtained, an overlay of the molecular electrostatic potential (MEP) can be derived that will aid in the description of the intermolecular forces present between molecules; discussed in detail in Section 2.5.1 Therefore an accurate model of the MEP at a given point in time is paramount in CSP.

A simple point charge model offers a straightforward method for modelling the MEP. This model approximates the electron density by carving it into discrete point charges distributed about the molecular framework. The method gives a reasonable approximation of the electron distribution that is computationally inexpensive to evaluate.

However, the point charge model is extremely limited in its ability to accurately describe areas of the molecule that are either electron rich (lone pairs of electrons) or, more generally, when the electron distribution about each atom is not spherically symmetric (π -bonding arrangements) [73]. In addition, the MEP is rarely spherically symmetric about the valence electrons of each atom in a molecule. Therefore a point charge model is an unreliable method for calculating an accurate MEP.

A more accurate model that gives a better fit of the MEP utilises a multipole expansion that builds on the point charge model where the latter represents the zeroth order term

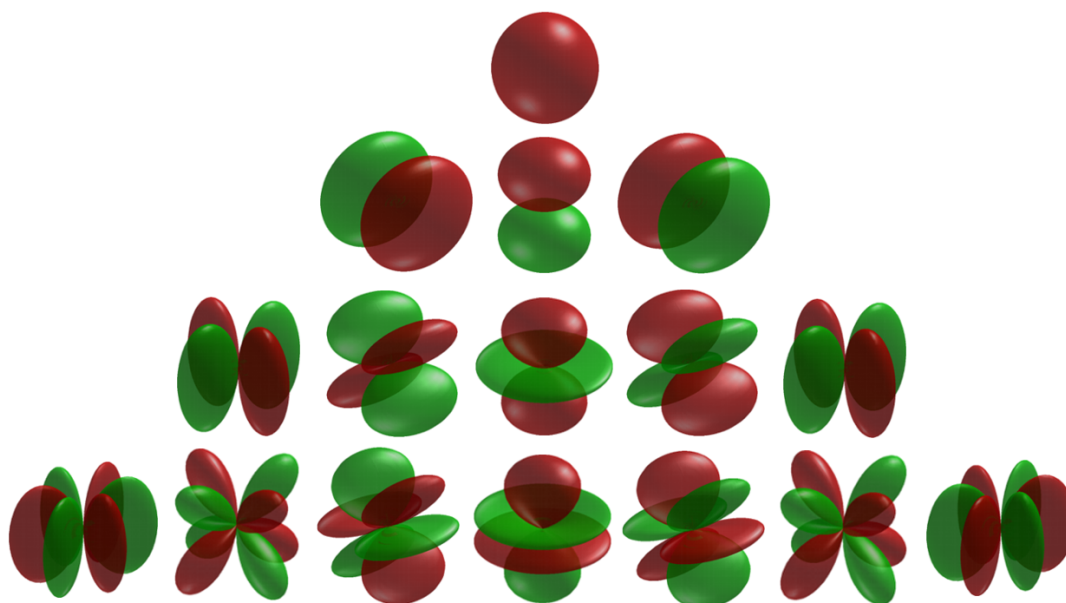


Figure 2.6: Showing, from top to bottom the monopole, dipole, quadrupole and octupole moments in a multipole expansion.

(monopole) in the expansion. More specifically, a distributed multipole analysis (DMA) implements a multipole expansion about a number of centres (typically the nuclei of atoms) within the molecule [74].

By using a Gaussian basis set, as described in Section 2.3.2, the total electron density is defined as a sum of a product of Gaussian functions:

$$\rho(r) = \sum_{i=1}^n \rho_i \phi_i^a(r) \phi_i^b(r) \quad (2.23)$$

where ϕ_i^a is a primitive Gaussian function at site a . The product of these two Gaussian functions is another Gaussian function that describes the electron density shared between sites a and b . The multipole series allows multiple electronic moments to be exactly calculated (dipole, quadrupole, octupole, etc...) about each centre where adding the higher order terms to the series gives a more accurate description of the MEP. Each higher order term becomes more computationally demanding to compute and makes a smaller contribution to total the expansion. These moments are visualised in figure 2.6 [75].

Although the multipole expansion is in principle an infinite series, and including all terms would describe the exact MEP, it was shown by Price *et al.* [76] that only terms up to and including the hexadecapole moment are needed for an accurate calculation where even the inclusion of the second-order dipole term can reduce the root-mean-squared deviation (RMSD) error against a highly accurate MEP by up to 50% [77].

A comparison of using atomic point charges against multipole expansions was investigated by Day *et al.* [78] where CSP calculations were performed on 50 small organic molecules. This quantitatively proved that the latter gave more accurate energy prediction than the former, albeit at an approximate ten-fold increase in computational cost. Using multipoles, 32 of the 64 crystal structures were predicted within 0.5 kJ mol^{-1} of the global minimum compared to 23 when using atomic point charges. Furthermore, the largest energy difference from the observed structure to the global energy minimum never exceeded 5.1 kJ mol^{-1} when using multipoles which compared to 7.3 kJ mol^{-1} for atomic point charges. In addition, the DMA expansion has also proved superior when predicting hydrogen bond geometries [79, 80] and mechanical properties [81] of smaller organic molecules.

A study by Nyman *et al.* [82], has also showed that these multipole methods can yield results that are as accurate as popular DFT-D methods but can incur errors that are up to 2-3 times larger than the best DFT-D methods. It was observed that the largest error is the underestimation of the lattice energy.

The strength of the above evidence for implementing a multipole expansion is vast and therefore this research will wholly make use of multipole series for describing an MEP unless indicated otherwise. Throughout this work, these multipoles are calculated using the GDMA code [83].

2.3.5 Rigid Molecule Approximation

The calculation of the molecular geometry and the generation of the MEP is applied to each and every conformer. For a ‘standard’, rigid molecule CSP process, the molecular geometry would be held rigid from this point onwards for the duration of the CSP procedure.

The definition of a rigid molecule is one whose geometry is not perturbed by external forces (regardless of their magnitude). This is equivalent to the molecular conformer represented by a single point at the bottom of a potential energy well where any deviation in the molecular geometry results in an infinite potential energy penalty. This is analogous with the hard sphere model in the kinetic theory of gases which dictates that the electron distribution of two atoms do not change when they collide. Furthermore, their electron densities cannot overlap as any occurrence will result in an infinite potential energy cost also.

The hard-sphere model is, of course, an approximation as molecular geometries will undergo some form of deformation when exposed to external forces. Nonetheless, this approximation has been proven effective for certain molecules.

The rigid approximation is valid for molecules that do not undergo sufficient geometrical distortions, when subject to crystal packing forces. A trivial example would be benzene where the aromaticity of the system enforces a planar and highly symmetrical geometry. Any deformation to the molecular geometry under the influence of crystal packing forces is negligible and therefore, to spare computational expense, minute perturbations to the molecular geometry are ignored.

2.4 Crystal Structure Generation

The final molecular geometries from the previous section is then used for crystal structure generation, in which a molecular conformer is packed into many different trial unit cells. Before delving into the procedures for this, the construction of a crystal from a molecule must commence.

2.4.1 From Molecule to Crystal

To form a crystal, molecules pack together in periodic arrangements held together by a variety of intermolecular forces that will be described in Section 2.5.1. Crystals are periodic in all 3 dimensions and are modelled as infinite lattices without surfaces. This is an approximation when only the crystal structures and properties of bulk crystals are of primary interest. Since an infinite lattice cannot be visualised, the concept of a unit cell is introduced that represents the minimum repeating unit in the crystal such that the entire crystal can be generated by simple translations.

The shape of the unit cell is determined by its 3 unit cell lengths (a , b , and c) and 3 unit cell angles (α , β and γ). There are seven different types of crystal systems (triclinic, monoclinic, orthorhombic, tetragonal, hexagonal and cubic) depending on the constraints on the six lattice parameters.

Every unit cell can be further simplified into an asymmetric unit which is the smallest symmetrically independent formula unit in a crystal. The number of formula units in the asymmetric unit is given by Z' and can possess both integer and non-integer values. Symmetry operations are then applied to the asymmetric unit to construct the unit cell where the number of formula units in the unit cell (Z) is defined by the number of symmetry operations used and the Z' value.

The number of symmetry operations is dependent on the space group of the crystal. Each space group possesses a different set of symmetry operations that defines the spatial and orientation relationships between pairs of molecules in the crystal.

There are 230 space groups but 80% of organic molecular crystals exist in one of the six most common space groups ($P2_1/c$, $P\bar{1}$, $C2/c$, $P2_12_12_1$, $P2_1$ and $P1$) [84]. Therefore it is not necessary to perform CSP on all 230 space groups. Different packing arrangements of molecules can reside in the same space group which therefore leads to different crystal structures that can possess different thermodynamic stabilities and exhibit different properties. Not all packing arrangements are thermodynamically stable and will form under experimental conditions. Nonetheless, it is known that multiple crystal structures using the same molecular component(s) are often observed, a phenomenon called polymorphism [85].

Polymorphs are more specifically classified by packing or conformational polymorphs [86]. The former is due to different packing arrangements of molecules in the crystal structure. The latter arises from the molecules adopting different molecular conformations in the crystal. This phenomenon then leads to different packing arrangements. A scenario occasionally occurs for which a molecule can adopt more than one conformation within the crystal. This leads to crystals that possess a $Z' > 1$.

These polymorphs can also exhibit different properties [87]. The most prominent example, amongst others [88], of the latter point is the anti-retroviral drug, Ritonavir [22] that has already been discussed in Section 1.2. This is just one example but it has the potential to arise in many organic molecular solids [89, 90].

2.4.2 Defining the Search Space

Each unit cell generated has 6 independent variables including 3 lattice vector lengths (a , b and c) and 3 lattice angles (α , β and γ). In addition, the unit cell possesses 3 translational and 3 rotational vectors per molecule in the unit cell. However, some translational dimensions can be omitted as the translation of all of the molecules in one unit cell is equivalent to translating the whole crystal lattice and therefore can be ignored. This represents a space of $6Z + 3$ dimensions to explore.

However, Section 2.4.1 has already shown that the molecules in the unit cell can be related by symmetry. Therefore the number of molecules in the unit cell can be reduced to Z' for single component crystals. The number of explorable dimensions can then be expressed as $6Z' + 6$. Hence the number of searchable dimensions is generally reduced, as $Z' \leq Z$.

Additionally, exploiting the symmetry imposed by a Bravais lattice type, the space groups potentially allow more of these dimensions to be omitted. For instance, a cubic space group requires $a = b = c$ and $\alpha = \beta = \gamma = 90^\circ$ hence reducing the number of dimensions to $6Z' + 1$. Therefore by searching each space group separately, the dimensionality of the problem can be further reduced. The space groups to be searched depend on user decision, typically discriminating based on the use of the statistical likelihood of a crystal structure residing in that space group. However very rarely are all 230 space groups searched, although this is possible in principle.

The decision of which space groups and Z' values to be included defines the search space and now leads to the problem of sampling this multi-dimensional space.

2.4.3 Molecular Transformations

For the overall CSP process, many different crystal structures will be generated by varying these dimensions just discussed in section 2.4.2. To generate a crystal structure, the molecule(s) in the asymmetric unit will be placed at a given position and orientation within the unit cell. This firstly starts with simple translational movement. These methods are important to convey as they represent the central theme of this research and will be extensively used and developed.

2.4.3.1 Molecular Translation

Molecular translation consists of identical variations in the Cartesian coordinates attributed to each atom in a molecule. Translation can occur solely in one, or a combination of, dimensions with the only pre-requisite being that each coordinate for each atom is altered by the same amount:

$$\begin{pmatrix} x_1' \\ y_1' \\ z_1' \\ \vdots \\ x_n' \\ y_n' \\ z_n' \end{pmatrix} = \begin{pmatrix} x_1 \\ y_1 \\ z_1 \\ \vdots \\ x_n \\ y_n \\ z_n \end{pmatrix} + \begin{pmatrix} D_x \\ D_y \\ D_z \\ \vdots \\ D_x \\ D_y \\ D_z \end{pmatrix}. \quad (2.24)$$

This condition ensures the internal geometry of the molecule does not change and a pure translational movement of the molecule is isolated.

2.4.3.2 Molecular Rotation

Rotating a molecule about the x, y or z axes in Cartesian space is performed by one of 3 rotation matrices:

$$R_x(\theta) = \begin{bmatrix} 1 & 0 & 0 \\ 0 & \cos(\theta) & -\sin(\theta) \\ 0 & \sin(\theta) & \cos(\theta) \end{bmatrix} \quad (2.25)$$

$$R_y(\theta) = \begin{bmatrix} \cos(\theta) & 0 & \sin(\theta) \\ 0 & 1 & 0 \\ -\sin(\theta) & 0 & \cos(\theta) \end{bmatrix} \quad (2.26)$$

$$R_z(\theta) = \begin{bmatrix} \cos(\theta) & -\sin(\theta) & 0 \\ \sin(\theta) & \cos(\theta) & 0 \\ 0 & 0 & 1 \end{bmatrix} \quad (2.27)$$

where θ is the angle of rotation about the given axis.

Taking the dot product of these matrices with a set of Cartesian coordinates, \vec{u} , rotates these points anti-clockwise by angle θ to yield a new, rotated set of Cartesian coordinates, \vec{u}' :

$$\vec{u}' = R \cdot \vec{u} \quad (2.28)$$

where R is the product of the three rotation matrices from Equations 2.25, 2.26 and 2.27:

$$R = R_x(\alpha) \cdot R_y(\beta) \cdot R_z(\gamma). \quad (2.29)$$

This expression allows the molecule to be rotated by a combination of angles simultaneously if required. In addition, the use of these matrices are independent of the internal geometry of the molecule which is left unchanged.

The rotation can occur about any point in space but generally it is performed about the centre of mass (CoM) of the molecular system, hence rotating the molecule about a local axis system. The CoM is a Cartesian coordinate where each coordinate is defined by the sum of the moments of each particle, i , divided by the total mass of the system, M :

$$x_{CoM} = \frac{\sum_{i=1}^N m_i x_i}{M} \quad (2.30)$$

$$y_{CoM} = \frac{\sum_{i=1}^N m_i y_i}{M} \quad (2.31)$$

$$z_{CoM} = \frac{\sum_{i=1}^N m_i z_i}{M} \quad (2.32)$$

where m is the mass of particle i . Applying an R centred on CoM leaves the internal coordinates of the molecule unchanged and therefore is denoted as a ‘pure’ rotation.

In CSP, it is often desirable to quantify the geometric difference between two molecules. This is performed by firstly aligning the CoM of both molecules by translational movements and then systematically applying these rotation matrices to one of the systems to minimise the RMSD.

2.4.3.3 Comparison of Molecular Geometries

The RMSD quantifies differences in two sets of Cartesian coordinates, A and B, and is defined as the square root of the arithmetically averaged sum of the squared differences between each corresponding coordinate:

$$RMSD = \sqrt{\frac{\sum_{i=1}^N (A_i - B_i)^2}{N}}. \quad (2.33)$$

These differences are commonly measured in Å. Chapter 5 will discuss an alternative method to quantifying differences in molecular geometries. To practically measure the RMSD, the first stage is to overlay the two CoMs of the two systems. This is achieved by setting the CoM of the first system, A, as the origin and subtracting the difference in CoMs, $\Delta\vec{CoM}$, from the coordinates of the second system, B:

$$\Delta\vec{CoM} = Co\vec{M}_B - Co\vec{M}_A = \begin{pmatrix} x_B \\ y_B \\ z_B \end{pmatrix} - \begin{pmatrix} x_A \\ y_A \\ z_A \end{pmatrix} = \begin{pmatrix} \Delta x_{CoM} \\ \Delta y_{CoM} \\ \Delta z_{CoM} \end{pmatrix} \quad (2.34)$$

$$B_{x,y,z}^{\vec{\text{align}}} = B_{x,y,z}^{\vec{}} - \Delta\vec{CoM} = \begin{pmatrix} x_1 \\ y_1 \\ z_1 \\ \vdots \\ x_n \\ y_n \\ z_n \end{pmatrix} - \begin{pmatrix} \Delta x_{CoM} \\ \Delta y_{CoM} \\ \Delta z_{CoM} \\ \vdots \\ \Delta x_{CoM} \\ \Delta y_{CoM} \\ \Delta z_{CoM} \end{pmatrix}. \quad (2.35)$$

Now the molecules are overlayed in space but their orientation remains askew.

To rectify this, the alignment of the rotation vectors now takes place. Generally this can be performed in one of two methods. The first and more simple, but more computationally expensive, is to minimise the RMSD between the two systems by varying the θ values for each of the rotation matrices (Equations 2.25, 2.26 and 2.27) in turn.

This method is relatively computationally expensive as one function evaluation requires multiple matrix-matrix dot products before computing the RMSD value. As this is a minimisation procedure, it is not possible to know in advance how many function evaluations will be required prior to commencing the procedure.

An alternative method is to implement the ideas proposed by Kabsch [91] that finds the optimal rotation matrix to most closely relate the two sets of Cartesian coordinates. This is still an iterative procedure but only requires one function evaluation and therefore is vastly superior in terms of computational expense.

The alignment of the CoMs and rotation matrices do not affect the internal molecular geometry which allows for a fair comparison when measuring the RMSD between two systems.

If the two molecular geometries are identical then the RMSD will be zero. Therefore after aligning the CoM and the rotation matrices, any non-zero value of the RMSD quantifies the difference in the internal DOFs of the molecule.

2.4.4 Sampling the Search Space

Using these methods just discussed, many crystal structures are able to be constructed. However, to determine which crystal structures are to be generated, a search procedure is initiated to fairly sample the n -dimensional space that exists. These search methods are generally classified into systematic or random searches. The former is implemented by sampling periodically along each dimension of the search space, for example grid-based methods. However, this does not prove to be an efficient method as the search must be completed to ensure the space has been sampled evenly, see Figure 2.7a.

Random sampling methods are more commonly employed as these ensure a more effective sampling of the search space. However purely random sampling¹, Figure 2.7b, also presents its own issues in that there is no guarantee that all space will be searched and therefore crystal structures can frequently be ‘missed’ during the search. This can be observed in Figure 2.7b as the large areas of white space that exist. In addition, the areas of white space occurs randomly and therefore it is not possible to know in advance where it will occur.

A more effective method is to actually combine these two approaches into generating a set of quasi-random points to sample the search space. To achieve this, a Sobol

¹More accurately, truly random sampling is only possible in principle; computational techniques can only produce a set of pseudo-random numbers (numbers that are designed to appear random). A more detailed discussion can be found elsewhere [92].

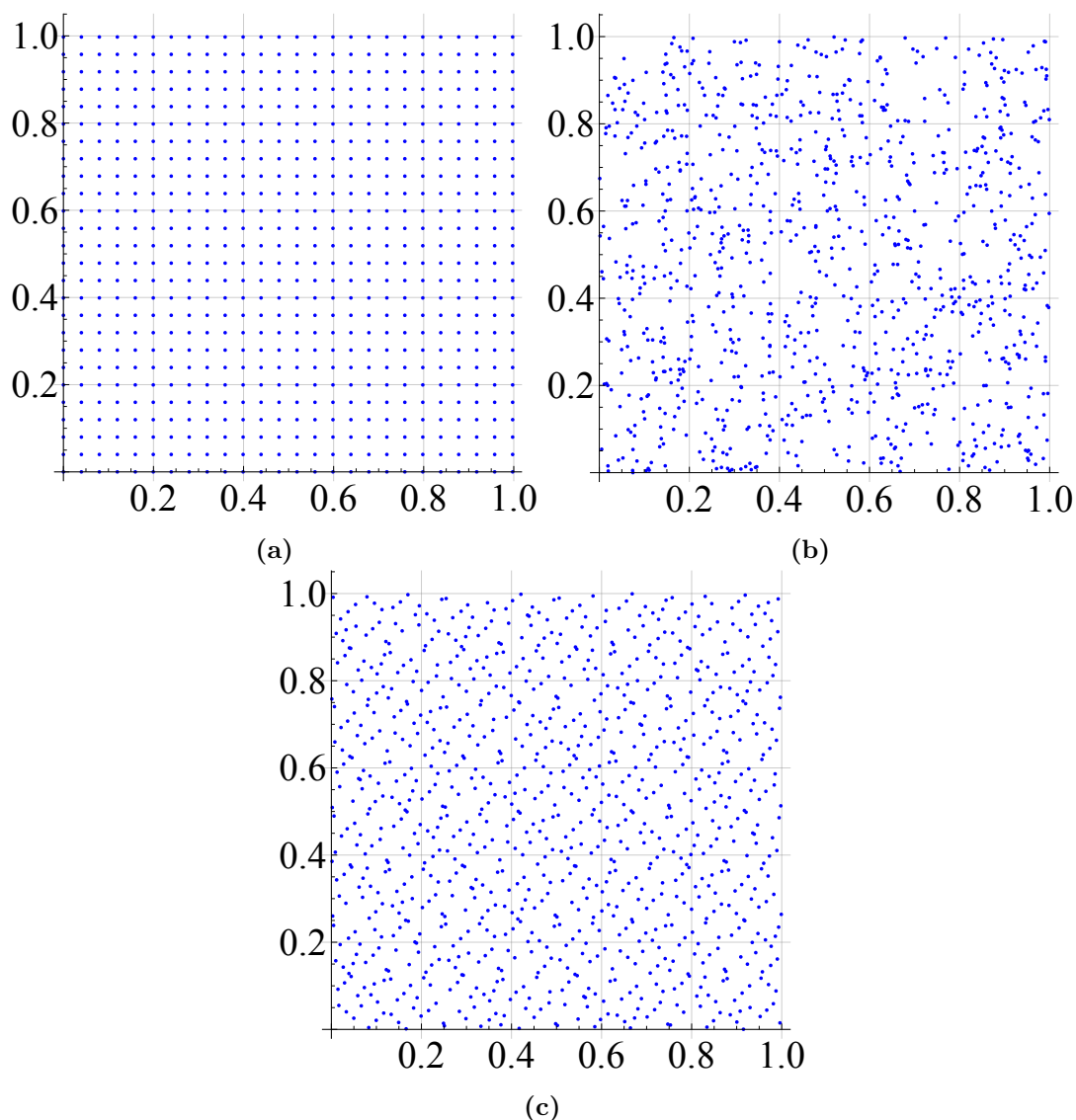


Figure 2.7: shows grid-based (a), purely-random (b) and quasi-randomly (Sobol sequence) (c) generated set of points between 0 and 1 in two dimensions.

sequence [93] is implemented which generates this set of numbers, Figure 2.7c. Figure 2.7 clearly illustrates that using a set of quasi-random points yields a far more evenly sampled space and hence is implemented in the CSP method [94]. Figure 2.12a illustrates an unsimplified, idealised situation of a well sampled PES where, under local energy minimisation procedures, every possible minima is found.

Another search method is simulated annealing that uses a Monte Carlo methodology to apply random changes to the crystal structure for which an energy penalty (ΔE) will be exhibited. The change, or move, is accepted if the Boltzmann factor ($e^{\frac{-\Delta E}{kT}}$) is greater than a random number between 0 and 1 [95]. The simulated annealing systematically increases the temperature as more steps are taken until all steps are accepted. This allows all energy barriers to be crossed and all of the PES can be explored (even though

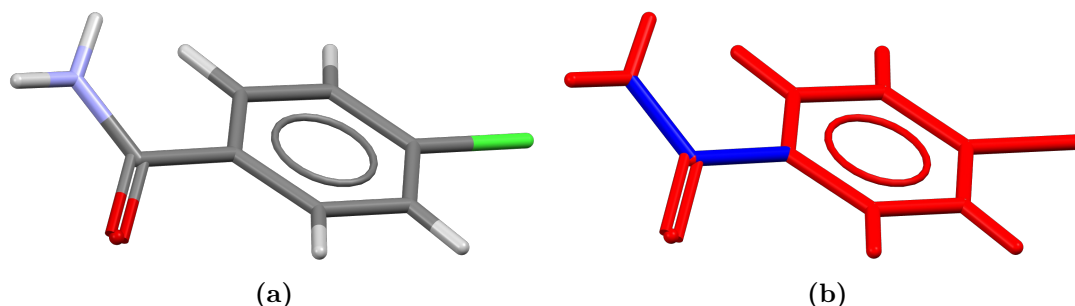


Figure 2.8: (a) shows a molecular conformation coloured by element and (b) that shows the same molecular conformation coloured by rigid fragments (red) and soft torsional angles (blue).

there is no guarantee that it will be). However, this methodology is computationally expensive which does not offset the increased accuracy of the calculations.

Basin hopping is another form of global optimisation technique [96]. This essentially flattens the PES by assigning a point in the search space to a local minimum such that any activation energy barriers to different regions are more likely to be overcome and therefore more of the PES is available for exploration. However due to the increased searching, more iterations are required for successful convergence of the minimiser which leads to a vast increase in computational cost. From a practical perspective, one could generate one crystal structure for every space group and employ a basin hopping algorithm. However, if the search methods outlined in Section 2.4 sample the PES well enough, local energy minimisation should be sufficient as all minima will be found.

2.4.5 Flexible Molecule Crystal Structure Generation

The context of the search and sampling procedures described thus far have all assumed that the molecule is defined by the rigid molecule approximation. However, it is also prudent to simultaneously sample the molecular conformational space as well as the dimensions of the crystal environment. Whilst performing the CSP calculations on separate conformers allows for more rigorous searching of the conformational space, the issue of inserting gas phase conformations into a crystal still remains unsolved.

CrystalPredictor [97, 98] attempts to address this by first partitioning the conformer into a set of rigid fragments separated by user-defined, flexible torsional angles, θ , where $-180^\circ < \theta < +180^\circ$. These rigid fragments are then rotated in space about these flexible torsions. An example of this is shown in Figure 2.8.

Single point energy calculations are conducted at regular intervals throughout the θ range for each flexible torsion angle. There are 2 flexible torsion angles in the molecule

shown in Figure 2.8 for which a 2D plot could be created that shows the intramolecular energy as a function the θ values.

The points are then interpolated by fitting a set of Hermite polynomials [99] to the existing points. These polynomials are truncated after the first derivatives to prevent spiralling computational costs. From visualising this function, stable minima can be observed with respect to θ which can give insight into different possible molecular conformations to perform structure generation on.

In addition to this, the atomic point charges, that will be passed to DMACRYS, are computed at only half the number of original points on the grid but fitted using the same procedure described in the previous paragraph. This coarser grid can yield errors of up to 20% [98] as the accuracy of the calculations are reduced but with the interest of minimising computational cost.

CrystalPredictor allows for the simultaneous search of the conformational and crystal environments and therefore could be considered as a ‘hybrid’ flexible search method as it couples Sections 2.4 and 2.2. This method has seen past success in successfully predicting the crystal structures of blind test molecules XX [100] and XXVI [65].

Another program is Polymorph Predictor [101] that implements a Monte Carlo search to simultaneously sample the molecular geometry and the unit cell parameters. The molecular geometry is searched, again, by demarcating the the molecule into rigid fragments connected by flexible DOFs. These DOFs and unit cell parameters are explored using simulated annealing. Whilst this methodology is computationally cheap, it does not yield a DFT level of accuracy.

2.4.5.1 Surface Fitting

One method was just briefly described to fit a surface to a set of points. The importance of surface fitting is that it allows near quantum mechanical accuracy of the intramolecular energies and atomic point charges of a flexible molecule without having to actually perform the quantum mechanical calculations as every point of the surface. Now a concise description of a variety of methods will be commenced as this will prove of importance in Chapter 4.

A simple approach to surface fitting can be performed by fitting a set of polynomial functions to a set of grid points. Interpolation of this function allows a smooth surface to be derived that can be searched for low function value regions and wells to identify minima. In the context of CSP, the function would describe the potential energy of a molecule as a function of molecular distortion.

Common methods for deriving functions to interpolate between grid points include the Le Gendre polynomials [102] and a Taylor series expansion [103] as well as other methods that will not be discussed further in this research [104].

Another method is one that employs Gaussian processing which will be used in Chapters 4 and 8. This uses Gaussian functions to predict a mean value, μ , which gives the best unbiased prediction of a point, X , on a PES:

$$\mu(X) = f(X)^T \cdot \beta + r(X)^T \cdot \gamma \quad (2.36)$$

where $f(X)$ is a vector that contains a second-order polynomial, β is the optimised co-efficient such that:

$$\mu(X) \mapsto \beta_0 + \sum_i X_i \beta_i + \sum_i \sum_j X_i X_j \beta_{ij} \quad (2.37)$$

where $r(X)^\alpha$ is a covariance function where the argument is the distance between X and a trial point, α , and γ is a parameter that depends on $r(X)$. The latter exists in the form:

$$r(|X - X'|, \theta) = \exp\left(\sum_i \theta_i (X_i - X'_i)^2\right) \quad (2.38)$$

where θ is a hyperparameter which is optimised and yielded, along with β , during the fitting procedure. The form of Equation 2.38 possesses a Gaussian framework, Equation 2.22, that always exists as a smooth function regardless of the order of differentiation performed on it.

2.5 Crystal Structure Energy Minimisation

Once these crystal structures have been generated, each one is subject to a lattice energy minimisation. Efficient modelling of the intermolecular interactions is paramount during this phase as this will determine where the minimised crystal structure resides on the PES [105]. The goal of this stage is to minimise the lattice energy with respect to the intermolecular geometry, whereby the crystal structure is only permitted to modify its $6Z' + 6$ dimensions. Since this is still a minimisation procedure, the same principles illustrated in Figure 2.5 and discussed in Section 2.3.3 are applicable.

2.5.1 Intermolecular Interactions

Chapter 2 has so far discussed methods to effectively model the energy of an individual molecule. This section will now focus on the interactions that exist between molecules.

To effectively model the energy of the intermolecular interactions between two molecules, all of the individual atom-atom interactions must be taken into account:

$$U_{AB} = \frac{1}{2} \sum_{a \in A} \sum_{b \in B} u^{ab}(r_{ab}, \Omega_{ab}) \quad (2.39)$$

where a and b are atoms within molecules A and B , respectively, u^{ab} is the interaction between a and b which is a function of the interatomic distance, r and the relative molecular orientation of the atoms, Ω . Multiplying the whole equation by $\frac{1}{2}$ rectifies the issue that arises from double-counting a single interaction.

The goal of modelling intermolecular interactions in the field of CSP is to accurately calculate the lattice energy of a molecular crystal. It is widely accepted that the accuracy and confidence of CSP depends on how well the intermolecular interactions have been modelled.

To obtain the exact lattice energy, one would need to sum all of the atom-atom pairwise interactions in the crystal, however pairwise interactions within the same molecule are ignored in this intermolecular energy model. Intramolecular interactions are accounted for using the DFT methodology whereas the electrostatic multipoles describe these intermolecular interactions (Sections 2.3.1 to 2.3.4) This would be to generate an infinite number of interactions to be computed (from Equation 2.39) for an infinite number of atoms in an infinite number of molecules which is, of course, an impossible calculation to perform.

The infinite summation is often cast into a finite form with either a finite cutoff radius in real space or by means of an Ewald summation in reciprocal space. This depends on how fast the interaction decays at long-range.

The former applies this radius about a molecule's CoM (typically 15 Å but this can also depend in the size of the molecule). This simplifies the problem greatly and vastly reduces the number of calculations to be performed. This is a valid approximation as the molecules with CoMs that lie outside of the cutoff range (assuming it is set to a sensible value) do not affect the subject molecule a significant amount.

The latter solution acknowledges that long-range electrostatic interactions (charge-charge, charge-dipole) can still have an effect, in which case an Ewald summation [106] is implemented which is computationally efficient as the energy of the system rapidly converges.

Fundamentally, these intermolecular interactions are governed by the distribution of molecular electron densities, Section 2.3.4.

2.5.1.1 Short-Range Interactions

Short-range interactions exist when the electron densities between molecules begin to overlap. This firstly induces a strong repulsive interaction due to the violation of the Pauli exclusion principle. The electron density then redistributes to reduce this overlap, but at a high energy cost.

Secondly, the exchange interaction arises from the electronic wavefunctions overlapping between non-identical molecules. This gives an increased electron separation and a partial cancellation of the aforementioned repulsion interaction.

2.5.1.2 Long-Range Interactions

The relevant long-range interactions are partitioned into three physical classifications electrostatic, dispersion and induced interactions. Other long-range interactions do exist [107] but are not significant for ground-state, closed-shell organic molecules.

The electrostatic interaction is a classical pairwise interaction (Equation 2.39) that can be attractive or repulsive. These depend simply on the charge distributions between two molecules and is mathematically described by Coulombs law if point charges are being used:

$$V = \frac{q_1 q_2}{4\pi\epsilon_0 r} \quad (2.40)$$

where V is the energy of interaction, q is the charge on the species, r is the inter-particle distance and ϵ_0 is the permittivity of the medium to conduct an electric field.

Induction interactions result from the electric field of neighbouring molecules acting on a central molecule. The charge density of the affected molecule is polarised and the charge is redistributed. This is an attractive interaction as the cost of charge redistribution is offset by the increased strength of the intermolecular bonds formed as a result.

Dispersion forces are weak intermolecular forces that occur due to the instantaneous fluctuations of movement of the electron density within molecules becoming correlated. These create instantaneous dipoles and higher order multipoles that cooperatively interact to contribute additional stabilisation to the crystal structure. This type of intermolecular force is not well modelled by DFT as the B3LYP functional only provides information about the local molecular environment. However, an explicit dispersion term, E_{Disp} , that corrects for this inadequacy can be added to the E_{DFT} to afford a dispersion corrected energy, $E_{\text{DFT-D}}$:

$$E_{\text{DFT-D}} = E_{\text{DFT}} + E_{\text{Disp}} \quad (2.41)$$

Interaction Type	Short (S)/Long (L) Range Partition
van der Waals	Dispersion(L) Exchange-Repulsion(S) Induction(L)
Hydrogen bond	Electrostatic(L) Induction(L) Dispersion(L) Exchange-Repulsion(S) Charge-Transfer(S)
$\pi - \pi$ interactions	Dispersion(L) Exchange-Repulsion(S) Induction(L)
Ionic bond	Electrostatic(L) Induction(L) Dispersion(L)

Table 2.1: Summary of the commonly referred to intermolecular interactions.

where the latter term possesses the functional form:

$$E_{\text{Disp}} = - \sum \frac{C}{r^6} \quad (2.42)$$

where r is the distance between the CoMs. The inclusion of this dispersion energy is accounted for in this thesis by the D3 version of Grimme’s dispersion with Becke-Johnson damping (GD3BJ) functional [108] in cooperation with B3LYP. This functional was found to produce reliable non-bonded distances when compared with other dispersion corrections [55, 109].

Additionally, charge-transfer interactions are those which involve the transference of electron density between two molecules in close proximity to one another. This is more commonly observed in hydrogen bonding interactions (Table 2.1) due to the large quantity of positive charge that is present on the hydrogen atom.

Combinations of these short- and long-range interactions form the more general intermolecular interactions (these combinations are summarised in Table 2.1):

- **van der Waals interactions:** this interaction is the net attractive or repulsive force between molecules. The interaction describes forces between permanent dipoles, permanent-induced dipoles and instantaneous dipoles. The force is comprised of the dispersion, induction and exchange-repulsion interactions.

- **Hydrogen bonds:** these interactions are formed from a highly electronegative atom (O, N or F) covalently bonded to a hydrogen atom which then interacts with another polar hydrogen bond acceptor atom. The electronegativity differences polarise the hydrogen atom that results in a highly directional, strong permanent dipole being formed. Therefore this interaction is comprised of the electrostatic, induction, dispersion, exchange-repulsion and charge transfer components.
- **Ionic bonds:** formal charges localised on atoms can be modelled as point charges and therefore the Coulomb equation (2.40) can be used to describe this interaction. Electrostatic, induction and dispersion forces are required to accurately model an ionic bond. Whilst this research is restricted to organic molecular crystals, and not ionic lattices, ionic interactions are present in zwitterionic compounds such as amino acids; a phenomenon that will occur in Chapter 3.

2.5.2 Force-Field Models

The Lennard-Jones force-field [110] is a standard potential that can be used to describe interacting molecules in all gas, liquid and solid phases and whose functional form is as follows:

$$V_{\text{LJ}} = 4\epsilon \left[\left(\frac{\sigma}{r} \right)^{12} - \left(\frac{\sigma}{r} \right)^6 \right] \quad (2.43)$$

where ϵ is the depth of the potential well, σ is the distance where the inter-particle interaction energy is 0 and r is the inter-particle distance. The r^{-12} and r^{-6} terms describe the repulsive, short-range, and attractive, long-range, interactions between the two species, respectively. An issue with this potential is that it only possesses two parameters and therefore is limited in its flexibility to describe other types of interactions (for instance, the rest of those described in Section 2.5.1). However, a larger issue is that the r^{-12} description of the repulsive interactions is not an accurate representation for molecules in a crystalline environment.

For this reason, the use of a Buckingham potential functional form [111] is used in this research. This possesses three parameters which allows more flexibility when describing intermolecular interactions:

$$V = Ae^{-Br} - \frac{C}{r^6}. \quad (2.44)$$

Within this functional form, the short range exchange-repulsion interactions are described by the e^{-Br} term whereas the long-range dispersion interactions are captured by the $1/r^6$ term.

The Buckingham potential also possesses a repulsive region at small values of r that corresponds to the short-ranged exchange-repulsion interaction. However, if r is further

reduced, a strong attractive interaction results where the Buckingham potential breaks down and fails. The implemented version of this force-field should provide a significantly high energy barrier to ensure this region is never reached during any calculation.

Taking Equation 2.44 further, the treatment of the electrostatic interactions can be added simply by appending the Coulomb equation, 2.40, to the Buckingham potential, Equation 2.44 thus:

$$V = Ae^{-Br} - \frac{C}{r^6} + \frac{q_1 q_2}{4\pi\epsilon_0 r}. \quad (2.45)$$

It is worth noting that the atomic charges, q_1 and q_2 , are conformation dependent and so are calculated in this research by electronic structure methods as the molecular conformation changes.

2.5.3 Force-Field Parametrisation

It is possible to tailor a force-field for every molecular system [112]. However, this is computationally expensive so a more practical approach is to use a transferable force-field where the parameters that describe the intermolecular forces can be transferred to different molecules.

CSP regularly employs a Williams ‘99 (W99) force-field [73, 113, 114] as the A , B and C parameters were yielded from empirical data from a set of organic crystals. This data consisted of the structure of the crystal determined from diffraction methods and its corresponding sublimation enthalpy. The A , B and C parameters were optimised such that this empirical data was as accurately reproduced as possible by the Buckingham potential. The W99 force-field only accounts for the four major atom types that exist in organic crystals: C, N, O, H. Each element is partitioned by the different chemical environments it can exist in within organic crystals, see Table 2.2. Homo- and hetero-terms refer to the A , B and C parameter values used when identical and different W99 atom types interact, respectively. The hetero terms are calculated from the geometric and arithmetic means of the homo-terms for the A/C and B parameters, respectively:

$$A^{ij} = (A^{ii} \cdot A^{jj})^{1/2} \quad (2.46)$$

$$B^{ij} = \frac{(B^{ii} + B^{jj})}{2} \quad (2.47)$$

$$C^{ij} = (C^{ii} \cdot C^{jj})^{1/2} \quad (2.48)$$

where i and j are the different W99 atom types. Therefore only 13 sets of parameter values need to be calculated (one set for each atom type).

Element(Type)	Description
H(1)	C- H atom
H(2)	alcohol O- H atom
H(3)	carboxyl COO H atom
H(4)	amino N- H atom
C(2)	sp ¹ hybridised C atom
C(3)	sp ² hybridised C atom
C(4)	sp ³ hybridised C atom
N(1)	sp ¹ hybridised N atom
N(2)	N atom with no adjacent H atoms
N(3)	N atom with 1 adjacent H atom
N(4)	N atom with 2 adjacent H atoms
O(1)	sp ² hybridised O atom
O(2)	sp ³ hybridised O atom

Table 2.2: All W99 atom type definitions and their environments.

Further modifications have been made to the W99 force-field [115, 116] but it still failed to describe the dispersion interaction of hydrogen-bonded hydrogen atoms so the C parameter was set to 0 for the H(2), H(3) and H(4)) atoms. Therefore hydrogen bonds are only described by the reduction in repulsion energy. This is a well known issue with pure electrostatic force-field methods [117]. This issue led to work by Pyzer-Knapp *et al.* [118] where this ‘original’ W99 potential (W99orig) was re-parametrised to correct this deficiency in the hydrogen-bond description. The C parameters remained at 0 as these values were negligible and increased the number of terms that needed calculating.

Instead, line searches were performed on the A parameters to obtain more accurate values. This was performed with the hetero-terms, A_{ij} where $i = \text{H}(2), \text{H}(3), \text{H}(4)$ and $j = \text{O}(1), \text{O}(2), \text{N}(1), \text{N}(2), \text{N}(3), \text{N}(4)$, rather than deriving them from the geometric means, Equation 2.46.

The electrostatic component used to model the total interaction energy was calculated using the triple-zeta Gaussian basis set, 6-311G**, as the analogous double-zeta Gaussian basis set, 6-31G**, is known to yield underestimations of the molecular dipole moments [119]. This is more prominent for certain multipole moments, in particular those involving amino groups. This work led to the ‘revised’ W99 potential (W99rev) whose main purpose was to determine a set of Buckingham parameters that can be used in conjunction with atomic multipole electrostatic models (Section 2.3.4). A product of this also incorporates an improved description of hydrogen-bonding into the force-field.

The testing of this W99rev will be performed in Chapter 3. The DMACRYS [76] software package is used for all crystal structure lattice energy minimisation performed in this work. This program uses the atomic multipoles of the molecule and an atom-atom intermolecular potential (Section 2.5.2) to calculate all of the inter-atomic interactions relevant to each atom in a molecular system.

2.5.4 Flexible Molecule Energy Minimisation

Again, throughout this section, a rigid molecule approximation has been assumed. However, there are methodologies that exist to simultaneously minimise the intermolecular lattice energy in conjunction with the intramolecular lattice energy. Before exploring these methods, an understanding of molecular flexibility within the crystal environment must be established.

2.5.4.1 Molecular Flexibility in Crystal Structures

The boundary where the rigid molecule approximation begins to break down is extremely vague. There is currently not a discrete set of rules that defines whether molecular flexibility should be included in the CSP process or not. There are of course some obvious cases when molecular flexibility is not needed, a point that was discussed in Section 2.3.5. However, the converse is not true as a molecule with a large number of atoms does not imply that it will be flexible.

Therefore, at least currently, chemical intuition and experience is the only measure that judges whether or not a molecule should be treated as flexible in the crystalline environment. This intuition is largely based on identifying ‘soft’ molecular motions that possess low force constants which have the potential to scan across all 3 types of DOF (bond lengths, bond angles and torsion angles). If it is decided that molecular flexibility should be included, it then requires further intuition to decide to what degree the molecule should be allowed to flex.

The magnitude of crystal packing forces is finite and will exert an effect on the DOFs with lower force constants (torsion angles) and a smaller, if any, effect on those DOFs that possess larger force constants (bond lengths). Therefore, to reduce computational effort, the molecule can be carved into DOFs that are considered flexible and those that are not; more specifically, those DOF that will be significantly perturbed by crystal packing forces, and those that will not. To whet the appetite, Chapter 3 attempts to define a general set of rules (therefore removing any chemical intuition) to partition these DOF into the flexible and non-flexible sets.

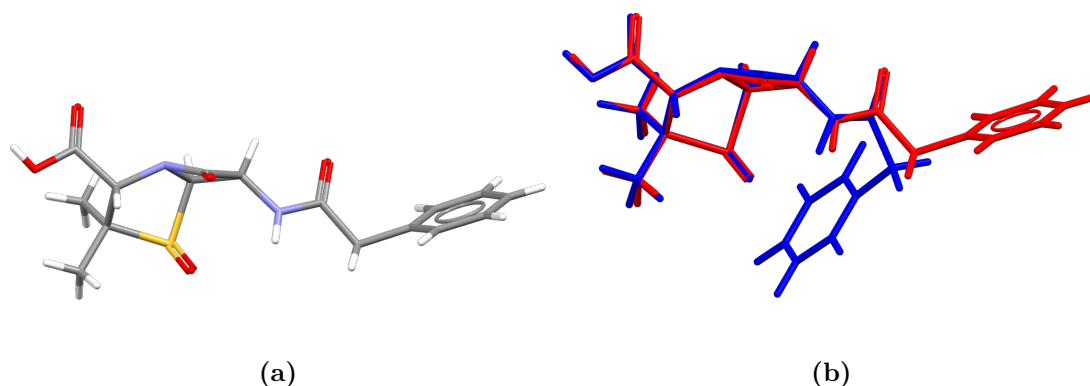


Figure 2.9: (a) the in-crystal geometry of a polymorph of penicillin. (b) compares the molecular conformers of the gas phase (blue) and the in-crystal (red) geometries, where the latter was calculated using the B3LYP-GD3BJ/6-311G** level of theory.

If crystal packing forces can significantly perturb certain DOFs in a target molecule, then the rigid approximation is no longer valid, hence it might not be energetically favourable for a molecule to remain in its gas phase geometry once inserted into the crystal. However, it is also possible that a flexible molecule could remain in its gas phase conformation, but at present it is not possible to predict whether this will be the case (Chapter 6 will endeavour to address this issue). An example has already been explored in Section 1.2 with Ritonavir where multiple molecular conformations exist therefore giving rise to conformational polymorphism. The difference in the intramolecular geometries for Ritonavir, Figures 1.1a and 1.1b, is the major issue for the rigid molecule CSP process outlined in Section 2.1 and illustrated in Figure 2.5.

Another example is the antibiotic, penicillin. Figure 2.9 highlights the differences in the gas phase and in-crystal molecular geometries where the latter was calculated using the B3LYP-GD3BJ/6-311G** level of theory. This is evidence that inserting the gas phase geometry into the CSP process under the rigid approximation may not yield the observed crystal structure. The reason for this is that once the molecule is inserted into the crystal, an activation energy barrier can exist between the gas phase and in-crystal geometries. This prevents the interconversion of these geometries and can distort the packing arrangement for that particular crystal structure. Hence leading to an unobserved crystal structure.

When performing CSP on more complex molecules that can be considered as ‘flexible’, geometry optimisations from Section 2.1 yields a geometry for the molecule that is stable in the gas phase which is not necessarily as stable when it exists within the crystalline environment. When the molecule is isolated, it can form intramolecular, non-covalent bonds which aids in stabilising the conformer. As a result of this behaviour, the molecule

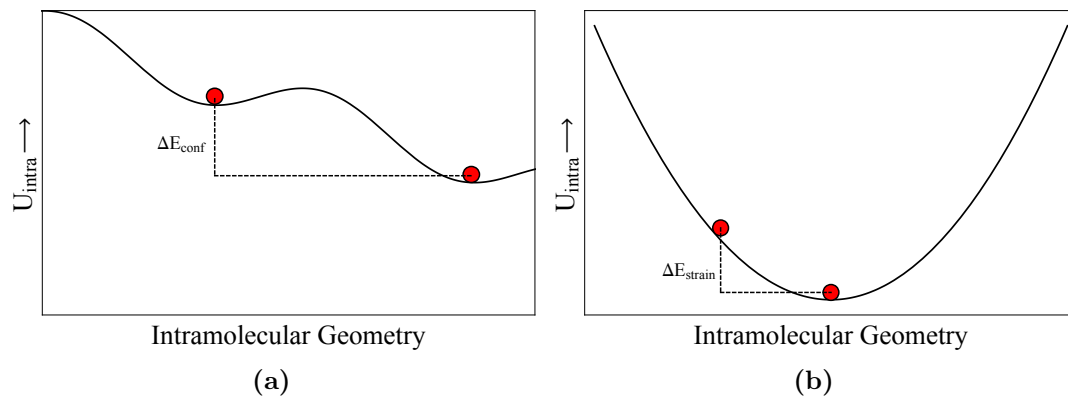


Figure 2.10: Two hypothetical potential energy surfaces showing (a) the energy difference between two molecular conformers, ΔE_{conf} , and (b) the energy difference between two molecular conformations that lie in the same potential energy well, ΔE_{strain} .

‘curls-up’ on itself. This can be observed in both of the aforementioned Figures 1.1a, 1.1b and current Figure 2.9.

2.5.4.2 Molecular Strain Energy

Thompson & Day [120] used the term ΔE_{conf} that measures the difference in energy from the gas phase optimised, in-crystal molecular geometry and the global energy minimum geometry found by a conformational search (these searches will be discussed in Section 2.2). The ΔE_{conf} is displayed by the schematic in Figure 2.10a. This hypothetical PES shows two minima and two conformers that reside at the bottom of these potential energy wells. The ΔE_{conf} value is hence the difference in U_{intra} between the two conformers. This research also proved a strong, linear correlation between the Connolly surface area [121] ($\Delta A_{\text{Connolly}}$) of a molecule and the intermolecular interaction energy in the crystal. The A_{Connolly} surface area was used to estimate the intermolecular interactions that a molecular conformation has the potential to form in the crystalline environment. This knowledge allows the ΔE_{conf} value to be adjusted to energetically describe these interactions. Molecular conformations that have a large surface area generally possess larger ΔE_{conf} values as their geometries are more ‘open’ and thus more strained from the gas phase geometry. Therefore a bias can be added to the ΔE_{conf} term:

$$\Delta E_{\text{conf,bias}} = \Delta E_{\text{conf}} + \Delta E_{\text{pseudo,inter}} \quad (2.49)$$

$$\Delta E_{\text{conf,bias}} = \Delta E_{\text{DFT-D}} + m_{\text{EvsSA}} \Delta A_{\text{Connolly}} \quad (2.50)$$

where $\Delta E_{\text{pseudo,inter}}$ is an energy correction that takes into account the relative molecular interaction energy from the crystal structure, $\Delta E_{\text{DFT-D}}$ is the ΔE_{conf} calculated by

DFT with a dispersion correction and m_{EvsSA} is the gradient of the linear regression of sublimation energy of the known crystal structure against A_{Connolly} .

The application of the correction in Equation 2.50 after a conformational search energetically reorders the structures by slightly skewing them to possess lower relative energies. However, the $\Delta E_{\text{pseudo,inter}}$ term in Equation 2.49 dramatically reorders the molecular conformations such that all of the conformations that occur in the observed structures occur within the first 7.1% of total structures (previously 39.5%).

As shall be explained further in Section 2.2, purely sampling the different conformers of a molecule is not a thorough enough method for sampling all of the possible molecular geometries available to the molecule.

Therefore the sampling of each molecular geometry about each potential energy well will lead to a more thorough distribution of all likely molecular conformations that a molecule could possess in the crystalline environment.

To allow for this, the same research [120] also derived the ΔE_{strain} term which measures the energy required to strain the molecule from its gas phase optimised, in-crystal conformer back to its in-crystal conformer, Figure 2.10b. In contrast to the ΔE_{conf} term, these two molecular conformations reside in the same potential energy well. It was found that up to 75% of molecules in the test set of molecular crystals resulted in $\Delta E_{\text{strain}} < 10.0 \text{ kJ mol}^{-1}$. These molecules possess approximately 4 to 6 soft torsion angles and a propensity to form intermolecular (however not intramolecular) hydrogen bonds, that gave a maximum ΔE_{strain} value of 22.5 kJ mol^{-1} . This increase in U_{intra} allows for larger offsets in U_{inter} hence overall reducing U_{latt} .

The structural formulas of molecules in this test set are illustrated in the later Figure 5.1, as we will return to these molecules in Chapters 5 and 6. Although this was a small test set of molecules, this ΔE_{strain} value indicates the magnitude of crystal packing forces for these flexible molecules. This research provides an insight into the properties of molecular deformation when residing in the crystalline environment which will therefore serve as a foundation for a large proportion of this thesis. However, a robust application and development of these findings firstly requires an understanding of the current methodologies available to treat molecular flexibility in CSP.

2.5.5 Methods for Flexible Molecule Energy Minimisation

UPACK [122] was the first software package to incorporate quantum mechanical calculations and force-field methods to describe the intra- and intermolecular energies, respectively. This was based on ideas presented by van Eijck *et al.* [123] (which were

also implemented in DMAFlex [124]) that attempted to incorporate molecular flexibility into the energy minimisation stage of CSP. This coupled DMAREL [35] (DMACRYS' predecessor), as this allowed the use of anisotropic atom-atom intermolecular potentials, with GAUSSIAN98 [125] that attempted to optimise the unit cell parameters in tandem with the molecular conformation with respect to its internal DOFs, respectively [124]:

$$\min_{\theta} [\Delta U_{\text{intra}}(\theta) + \min_x U_{\text{inter}}(X; \theta)] \quad (2.51)$$

where θ represents the intramolecular DOFs, U_{intra} is the intramolecular energy, U_{inter} is the intermolecular energy which is a function of X , the lattice angles and lengths, and θ . The latter can be partitioned into rigid and flexible DOFs, θ^r and θ^f , respectively. θ^f encompasses the user-defined DOFs that are explicitly varied by DMAflex. After a θ^f adjustment has been made, a geometry optimisation of the new molecular conformation is performed whilst holding all θ^f rigid and allowing all θ^r (non-user-selected DOFs) to become flexible.

This allows θ^r to react to the changes in θ^f and hence yields a more accurate representation of the molecular geometry. This modifies Equation 2.51 to [124]:

$$\min_{\theta^f} [\Delta E^{\text{intra}}(\theta^f) + \min_x U^{\text{inter}}[X; \theta(\theta^f)]] \quad (2.52)$$

This now shows the distinction between the two types of DOF. Partitioning the DOFs into two bins reduces the computational cost of the calculation whilst still allowing the molecule to be flexible.

CrystalOptimizer [126] (the successor to DMAflex) uses the same underlying theoretical approaches as described above (except for using GAUSSIAN09 [67] instead of GAUSSIAN98) but also provides a more sophisticated application. The CrystalOptimizer algorithm is illustrated in Figure 2.11.

The main difference is the use of local approximate models (LAMs) to approximate the intramolecular energy and electrostatics of the molecule, both of which are extracted from the DFT electron density calculation. The former uses a quadratic Taylor series expansion to estimate the intramolecular energy from a reference structure that is geometrically similar to the considered structure. The latter attempts to save performing expensive multipole calculations.

The local approximate model (LAM) allows the conversion of the multipole moments from the reference system into their Cartesian forms and rotates them about a local axis system so they overlap with the system of interest. This can be a bad approximation as the multipole moments of certain functional groups can drastically change with small

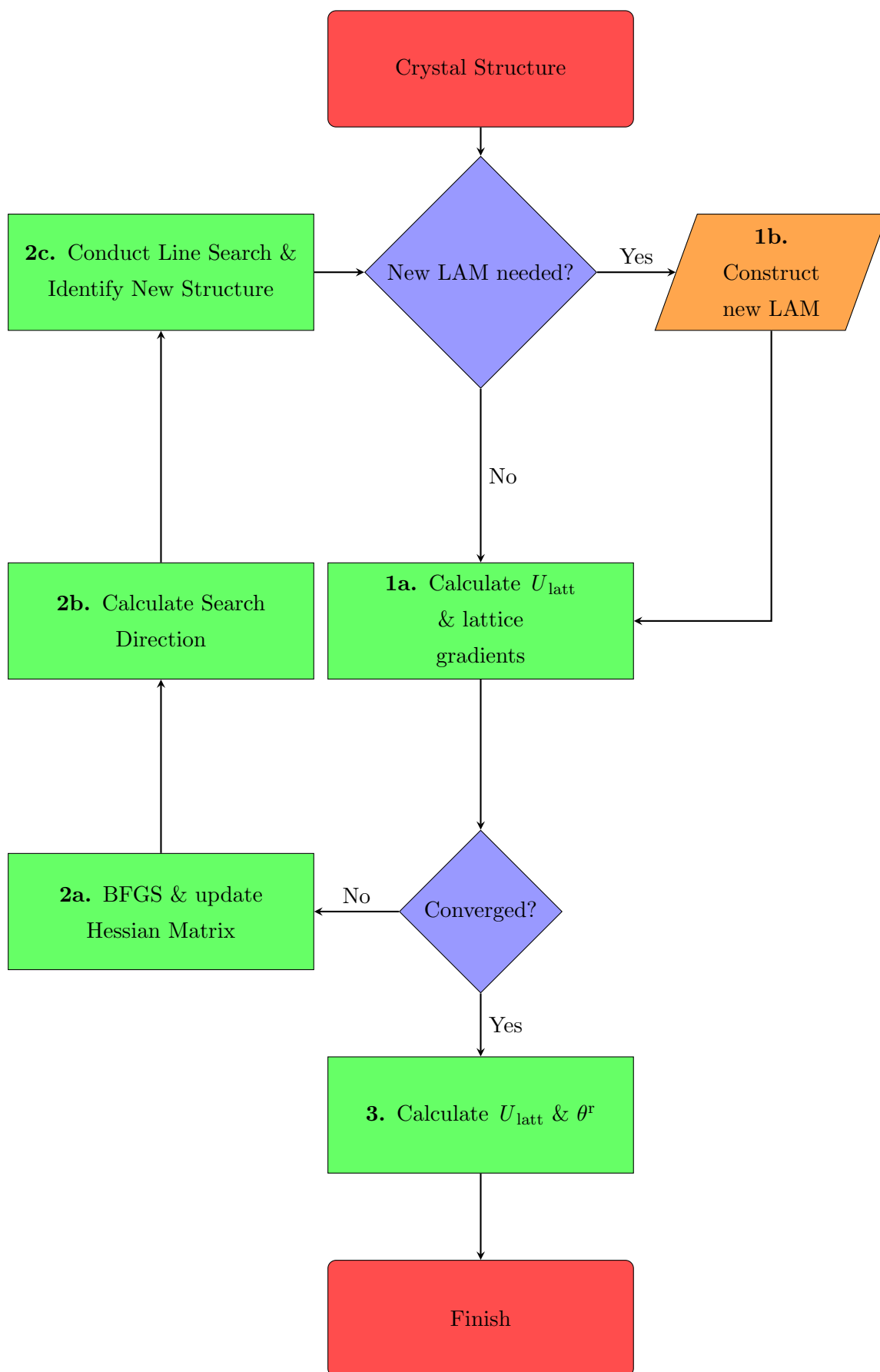


Figure 2.11: The CrystalOptimizer algorithm. The DMACRYS and GAUSSIAN09 software packages are used during stages 1a, 3 and 1b, 2a, 2b, 2c respectively. U_{latt} and θ^r represent the lattice energy and the set of 'rigid' molecular degrees of freedom.

changes in molecular geometry. Therefore an additional, linear correction can be applied to the Cartesian rotation tensor. Example functional groups that require this correction include amine and alcohol functional groups [126].

An interesting feature of CrystalOptimizer is that it utilises databases to further reduce computational cost. All LAMs, once computed, are stored in one of two databases; intramolecular energy and molecular electrostatics.

Before a DFT calculation is performed, both databases are checked and if a set of θ^f values are within a given tolerance, then those values are used for that molecular conformation. If more than one entry matches the given molecular conformation, then the entry with the lowest RMSD is chosen. The range that the LAMs possess varies with functional groups that the molecule contains.

However, it has been shown that a $\pm 5^\circ$ and $\pm 0.1 \text{ \AA}$ distortion in dihedral angles or bond angles and bond lengths, respectively, does not possess a significant change in either LAM. Therefore this allows less strict criteria to be applied when checking the databases for matches which further reduces the computational cost of this calculation. However, this data is wholly representative of the molecule in the gas phase; not the molecule in the crystalline environment. Therefore these databases can be reused for this molecule indefinitely and do not depend on the crystalline environment.

After the initial crystal structure has been loaded into CrystalOptimizer, the databases are checked for matching conformations. If the given conformation does not match any of the database entries, the molecular energy and multipoles are calculated using GAUSSIAN09 which are used to create the LAM. Stage 1a implements DMACRYS to calculate the lattice energy and lattice gradients. If these numbers have converged (the difference between the presently and previously calculated values are within a given tolerance), a final geometry optimisation is performed with respect to θ^r and then the lattice energy is calculated.

If the lattice energy and lattice gradients have not converged, the Hessian matrix is updated by means of the Broyden-Fletcher-Goldfarb-Shanno (BFGS) method [127]. This method forces the Hessian matrix to remain positive-definite which ensures changes to the flexible DOF yield a reduction in lattice energy, stage 2b. From a physical perspective, the Hessian matrix describes the forces acting on each atom and therefore allows the molecular conformation to distort itself based on that information. Stage 2c, implements a quasi-Newton line search method [127] which gives a less computationally demanding rate of convergence. The structure is then distorted accordingly and the iteration repeats.

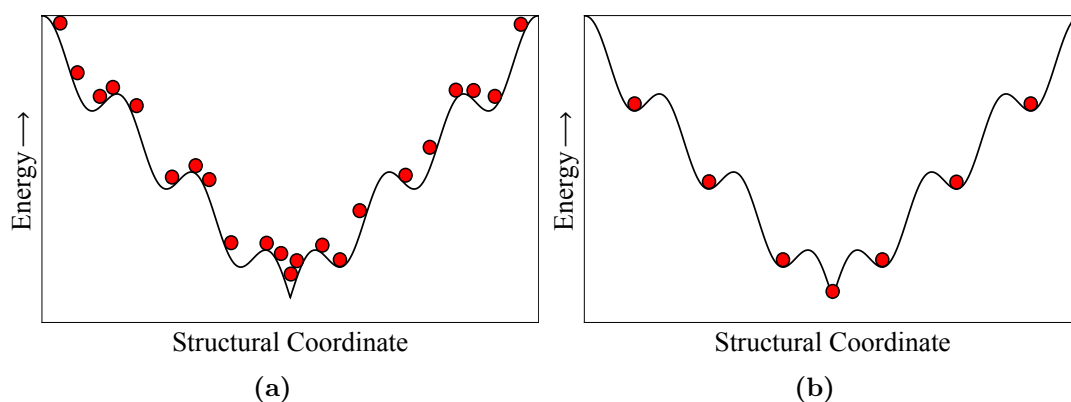


Figure 2.12: shows two potential energy surfaces plotting the potential energy as a function of the lattice parameters, structural coordinate. **(a)** illustrates a hypothetical potential energy surface where every red point represents a crystal structure that possesses a unique set of unit cell parameters and hence lies at a unique position. **(b)** shows the position of these unique crystal structures post-lattice energy minimisation.

Testing of the CrystalOptimizer algorithm has shown that including these additional steps into the CSP process gives a more accurate and reliable representation of the crystal [126]. While this is more computationally efficient, it does not distract from the fact that the whole process is computationally demanding. This vastly increases the time and resources required to calculate the final crystal structures although the use of the LAMs certainly aids in quelling this issue.

2.6 Clustering of Non-Unique Crystal Structures

Whether the crystal structures to this point have been subject to the rigid molecule approximation or the a flexible molecular energy minimisation has been performed, every single starting crystal structure resides at a different point on the PES. Assuming the search methodology was sufficient, multiple crystal structures should have been generated in and around the same potential energy well. However post-energy minimisation, these crystal structures will converge to the same local energy minimum. These structures possess near-identical lattice energies and crystalline geometries. Figure 2.12b illustrates this point when contrasted against Figure 2.12a. All structures that now exist in the same potential energy well undergo a process called ‘clustering’ in which the lowest energy structure is taken and the rest are discarded. One software package to carry out this procedure is COMPACK [128] which compares whether any two (or more) crystal structures are identical within given structural tolerances; if so, the highest energy structure(s) are removed from the list of crystal structure matches. COMPACK

avoids the use of space-group and unit cell information by comparing the relative molecular position and orientation as well as the interatomic distances which provides a more than adequate representation of the crystal structure.

Another method for comparing crystal structures is the Křivý-Gruber [129] methodology that produces the Niggli cell from an arbitrary primitive cell of a Bravais lattice. The mathematical details of this methodology are beyond the scope of this thesis will not be discussed further. Nonetheless, the advantage of this methodology is that it is computationally efficient as it calculates a direct solution for the rotation matrix required to find the optimal RMSD overlay between two crystal structures.

The clustering process simplifies the data handling of the results as there are less structures present to analyse as CSP regularly produces 10^2 - 10^6 crystal structures [16].

2.7 Crystal Structure Ranking

Once the final list of crystal structures is obtained, they are then ordered by a user selected scoring function.

The most common, when using the rigid molecule approximation, is simply the intermolecular component of the lattice energy, U_{inter} , as the assumption is made that the crystal structure will adopt the thermodynamic minimum. Ordering by the total lattice energy in this scenario, U_{latt} , yields the same result as the intramolecular contribution, U_{intra} , will be constant as the molecule is subject to the rigid molecule approximation.

The total lattice energy, U_{latt} , is more relevant and needed when either considering multiple conformers under the rigid molecular approximation or when including molecular flexibility within the CSP process:

$$U_{\text{latt}} = U_{\text{intra}} + U_{\text{inter}} \quad (2.53)$$

where U_{intra} is determined by the DFT methodology presented in Section 2.3.1. Unless otherwise noted, all structure rankings in this work are done by the U_{latt} .

In addition, a polarisable continuum model (PCM) can also be implemented to model the crystalline environment post clustering. Organic crystals possess a dielectric constant of approximately $3.0\epsilon_0$ [130] which a final set of calculations are performed before ranking the crystal structures by U_{latt} . In this calculation, the crystal structures are re-energy minimised to which their positions and orientations will be tweaked such that the crystal packing will be better modelled.

However, this equation assumes a perfect static crystal that exists at 0K and thus excludes free energy of a crystal at any realistic temperature. Nyman *et al.* [131] developed a methodology to determine the Helmholtz free energy of a crystal structure, A , by calculating the entropy contribution from phonon frequencies, S , present in a crystal at a given temperature, T , thus:

$$A = U_{\text{total}} - T \cdot S. \quad (2.54)$$

The subtraction of $T \cdot S$ gives a more physically accurate ranking of the crystal structures in the final list.

The ranking of the crystal structures is paramount as it is assumed that the crystal occurring at the bottom of the list is empirically observed. Assuming that an observed structure is available, to determine if the ranking is correct, the algorithms implemented within COMPACK [91, 132] to compute the lowest RMSD between the observed and every structure in the final list. An ideal CSP situation exists in that the crystal structure with the lowest U_{latt} in the final list is the observed structure where the latter does not completely match any other entries in the final list.

2.8 The Blind Tests

Now a basis of theory has been established, the story of the progression of CSP can be quantitatively measured and explained in technical detail. The blind tests are a series of six papers [18, 65, 100, 133–135] organised by CSP groups in collaboration with the Cambridge Crystallographic Data Centre (CCDC) where the first and sixth blind tests occurred in 1999 and 2015, respectively.

Each blind test is constructed in a similar way to its predecessors where the organiser(s) firstly select a set of molecules which are given to all participants. The crystal structures for these molecules are already known but are not released to the participants, hence the participants conduct their CSP studies ‘blind’. Once all of the participants have submitted their crystal structures, the actual crystal structures are released and the results are then analysed.

The purpose of these collaborative studies is to rigorously evaluate the theoretical basis of CSP and hence outline areas for improvement. The author was involved in the sixth blind test and whose exact contribution is detailed in Chapter 7. The molecules involved in each test are summarised in Table 2.3 and a chronological overview of all six blind tests will now be given.

The first blind test (molecules I, II and III in Table 2.3) consisted of three small organic molecules that could be classed as rigid bodies where the 11 participants were allowed to propose up to three crystal structures for each compound. 7 of these 11 participants proposed crystal structures that matched the observed crystal structures but overall no methodology stood out as superior.

The second blind test was conducted in an identical manner to the first and included 4 molecules (molecules IV, V, VI, VII). The molecules IV, V and VII could be modelled as rigid bodies and hence yielded successful predictions. Molecule VI gave no successful predictions as this molecule possessed conformational flexibility. The majority of participants chose to rank their crystal structures using lattice energies as a scoring function. Hence it is worth mentioning that molecule VI was found in the lists of crystal structures submitted by some participants but was never identified as the lowest energy crystal structure. The difficulties introduced by molecular flexibility were a known problem at the time of the publication of the second blind test but nonetheless highlighted the issue more prominently.

The third blind test (molecules VIII, IX, X and XI) yielded a more unsuccessful result than the second blind test. The only success was borne from molecule VIII that could be modeled as a rigid body and possessed three successful predictions. However molecules IX and XI were also rigid molecules and bore no successful predictions. The failures of these two molecules point to the energy models employed. The flexible molecule X yielded no successful predictions which only further highlighted the need for methods to effectively treat conformational flexibility.

The fourth blind test marked a significant improvement for the field of CSP where 14 participants yielded 13 successful predictions over the 4 target systems (molecules XII, XIII, XIV and XV). Each target possessed at least 2 successful predictions which led to the conclusion that CSP could be used to complement experimental results for simple organic molecules that can be modelled as a rigid body. The flexible molecule (XIV) was the least successful of the 4 that, once again, highlighted that improvements for the treatment of molecular flexibility in CSP were needed.

Additionally, Neumann, Kendrick and Leusen (implementing methods developed by Neumann *et al.* [112, 136, 137]) located all of the observed crystal structures correctly in the fourth blind test; albeit at an extremely high computational cost of 32 CPU years. These exact details of this method are shrouded in secrecy as it of commercial value, but high level DFT level of theory was coupled with a tailor-made force-field which are both well recognised as computationally expensive techniques.

The fifth blind test included 6 molecules (2 rigid molecules (XVI and XVII), 1 semi-flexible molecule (XVIII), a salt (XXI), a hydrate (XIX) and 1 larger, more flexible molecule (XX)). This was a significant increase in complexity of the targets that were suggested but nonetheless there was at least 1 successful prediction for each crystal. This includes 2 successful predictions of the flexible molecule which shows that the techniques for describing more complex systems are being developed with some success. Neumann's methodology again gave more positive results and successfully predicted 4 of the 6 crystal structures.

The sixth blind test, again, yielded more positive results and consisted of a rigid molecule (XXII), a molecule possessing many polymorphs (XXIII), a salt hydrate (XXIV), a co-crystal (XXV) and a flexible molecule (XXVI). All targets were predicted by at least 1 of the participants with the exception of one of the polymorphs for molecule XXIII. This blind test consisted of more participants than the previous tests and boasted a broader range of methodologies used. In particular, this test shows that the treatment of molecular flexibility within CSP is progressing well with 4 participants correctly predicting the crystal structures of molecule XXVI. The variety of systems in the sixth blind test shows that CSP is becoming ever more applicable to real systems and now can complement experimental studies.

Table 2.3: Molecules included in all of the six blind tests.


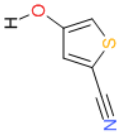
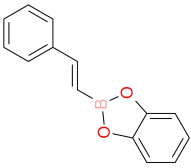
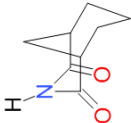
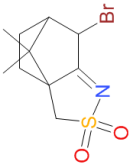
Molecular Index	Structural Formula	Number of Lists Submitted	Lists Containing Observed Structure	Lists Containing Observed Structure as 1st Prediction
I		11	Unknown	0
II		8	Unknown	0
III		11	Unknown	1
IV		16	10	1
V		15	11	3
Continued on the next page...				

Table 2.3 – Continued from the previous page

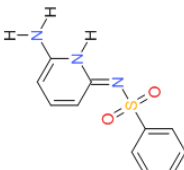

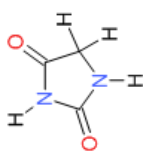
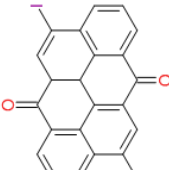
Molecular Index	Structural Formula	Number of Lists Submitted	Lists Containing Observed Structure	Lists Containing Observed Structure as 1st Prediction
VI		11	4	0
VII		6	Unknown	1
VIII		14	11	3
IX		16	9	1
Continued on the next page...				

Table 2.3 – Continued from the previous page

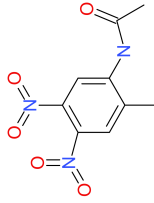


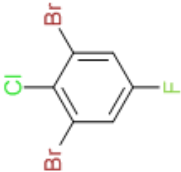
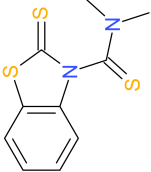
Molecular Index	Structural Formula	Number of Lists Submitted	Lists Containing Observed Structure	Lists Containing Observed Structure as 1st Prediction
X		15	7	0
XI		18	4	0
XII		13	10	2
XIII		14	9	4
XIV		12	9	3
Continued on the next page...				

Table 2.3 – Continued from the previous page

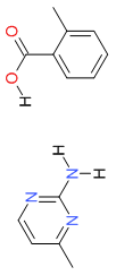
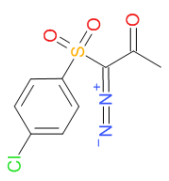
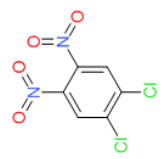
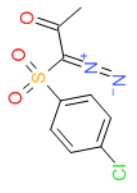
Molecular Index	Structural Formula	Number of Lists Submitted	Lists Containing Observed Structure	Lists Containing Observed Structure as 1st Prediction
XV		12	5	1
XVI		15	7	1
XVII		13	6	1
XVIII		12	4	1
Continued on the next page...				

Table 2.3 – Continued from the previous page

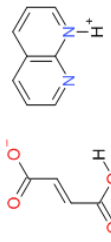
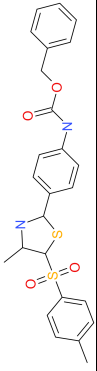
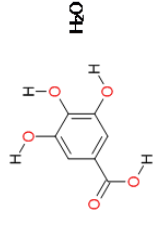
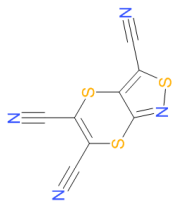
Molecular Index	Structural Formula	Number of Lists Submitted	Lists Containing Observed Structure	Lists Containing Observed Structure as 1st Prediction
XIX		11	4	0
XX		10	3	1
XXI		10	5	0
XXII		35	18	4
Continued on the next page...				

Table 2.3 – Continued from the previous page

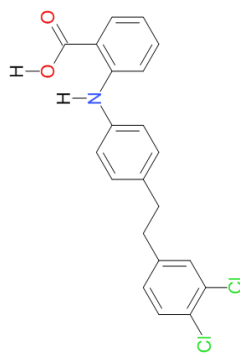
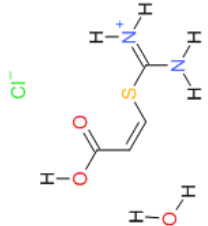
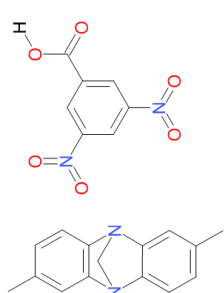
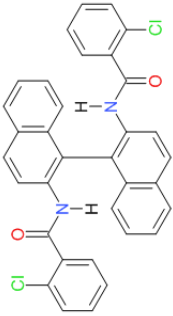
Molecular Index	Structural Formula	Number of Lists Submitted	Lists Containing Observed Structure	Lists Containing Observed Structure as 1st Prediction
XXIII		82	28	2
XXIV		16	1	0
XXV		21	8	5
Continued on the next page...				

Table 2.3 – Continued from the previous page

Molecular Index	Structural Formula	Number of Lists Submitted	Lists Containing Observed Structure	Lists Containing Observed Structure as 1st Prediction
XXVI		19	6	4

2.9 Summary

There is no doubt that the vast increase in computational power has contributed to the ability to the treatment molecular flexibility within CSP. There any many methods that have been developed to tackle this problem but an ideal solution is far from being accomplished.

However, a major factor that remains is the computational expense of these calculations. Although significant progress has been made to reduce this problem, it can still take many months of CPU time to generate a list of crystal structures that have been subject to the full treatment of flexibility. This is an issue that also scales with the number of atoms in the molecular system. Nonetheless, more and more articles are published with successful CSP predictions and progress of the CSP field as a whole is benchmarked by the blind tests. Each blind test possesses more participants, yields more positive results and yet contains more challenging molecular systems than its predecessor.

Nonetheless, this challenge of molecular flexibility is far from complete and requires more stringent theoretical techniques to model more complex molecular systems. Therefore we arrive at the turn where the current state of CSP has been summarised and can now move towards new methods to treat molecular flexibility in CSP as is the purpose of this research. We will begin by quantifying the effect of using flexibility for a set of small organic molecules.

Chapter 3

Molecular Flexibility & Williams99 Force-Field Testing

3.1 Introduction

The original W99 potential was reparametrised to include an accurate description of hydrogen bonding, Section 2.5.3. This chapter will quantify the differences between the original and this revised W99 potential (W99orig and W99rev, respectively) by the medium of a test set of small, organic molecules. This test set will also be used to quantify the change in accuracy of the CSP calculations when molecular flexibility is included and when it is omitted where the former will implement the CrystalOptimizer algorithm.

3.2 Obtaining a Test Set of Molecules

54 small organic molecules, Appendix A, were initially selected for this study. 50 of these had already performed as a solid test case in comparing the use of multipoles against atomic point charges for describing the electrostatics of a molecule, Section 2.3.1 [78], and also acting as a benchmark for other CSP studies [138].

The W99rev was reparametrised to provide a better description of hydrogen bonding in the crystalline environment. Therefore a rigorous analysis of these original 50 molecules was undertaken in order to identify which types of hydrogen bond were absent from this set. This information is summarised in, Table 3.1. Out of a total of 18 combinations, only 9 were included in the set of 50 molecules. Upon investigation, and searches of the CSD, it was discovered that H(3) is more energetically stable when dimerised as opposed

Hydrogen Bonding Interaction Combinations	Donors		
Acceptors	H(2)	H(3)	H(4)
N(1)	Blue	Red	Blue
N(2)	Yellow	Yellow	Yellow
N(3)	Red	Red	Red
N(4)	Blue	Red	Yellow
O(1)	Yellow	Yellow	Yellow
O(2)	Yellow	Red	Yellow

Table 3.1: Illustration of all possible hydrogen bonding combinations defined by the W99 atom typing. Yellow and red cells are interactions that are included and not included, respectively, in the set of 50 molecules. Blue cells are combinations that are not included in the original set of 50 molecules but are included in the revised set of 54 molecules.

to the N(1)-H(3) and O(2)-H(3) combinations. This is purely due the hydrogen bond acceptor strength not being stronger than the O(1) acceptor.

N(3) represents the nitrogen of the **N-H** functional group that is consistently a better hydrogen bond donor than acceptor therefore does not empirically form. The N(4)H(3) combination represents a nitrogen **NH₂** acting as an acceptor. Although this is not obvious, amino groups therefore are poor acceptors in organic crystals.

Combinations that were found to exist include N(1) atoms which refers to a nitrogen atom existing in a cyano-group that can form a hydrogen bond with an alcoholic hydrogen or amino hydrogen atom. In addition, N(4) acceptor H(2) donor combinations were found that correspond to a **NH₂-HO** hydrogen bonding combination.

Four molecules were added to the original 50 molecule test set and whose molecular formulas can also be observed in Appendix A. These additions are highlighted as blue cells in Table 3.1. This selection of molecules now incorporates all types of hydrogen bonding arrangements that are empirically capable of forming. This now acts as a more rigorous test for the W99rev potential.

3.3 Defining Flexibility

Setting all DOF to ‘flexible’ for a given molecule within a CrystalOptimizer calculation can lead to an extreme computationally expensive. For molecules with less than 10 atoms, this is a realistic possibility but 34 of the 54 molecules possess more than 10 atoms so molecular flexibility must be approximated.

Therefore a list of rules for defining molecular flexibility for this research was generated and exists as follows:

1. All bond lengths are classified as rigid.
2. Aromatic systems are rigid.
3. Non-aromatic, cyclic systems are flexible (including those that are adjacent to aromatic systems).
4. The exception to rule 2 is that if the aromatic ring contains a heteroatom, X (where $X = N$ or O), all bond angles and dihedrals containing this atom X are classed as flexible.
5. Exclude any dihedral or angle that starts or ends with a HC- or -CH, respectively, unless rule 4 can be applied or the molecule possesses dihedral angles that propagate along one, and only one, bond axis.
6. Exclude any dihedral or angle where 3 or 2 of the atoms, respectively, are included in an aromatic system unless rule 4 can be applied.
7. Include all dihedrals and angles starting or ending with HX- or -XH, respectively, (where $X = N$ or O) regardless of if it lies in an aromatic or non-aromatic system.

In addition, aromaticity is defined by four rules:

1. The system must be planar.
2. The system must be cyclic.
3. Each atom must be capable of delocalising the electron density by the medium of a vacant orbital(s).
4. Hückel's rule [139] must apply where the number of delocalised p-electrons equals $4n + 2$ where $n \in \mathbb{Z}^+$.

These rules allow the interatomic distances to stay constant but allow the molecule to distort through its bond angles and dihedral angles. Modelling aromatic systems as rigid is due to the π -electrons enforcing planarity. This rule allows the aromatic moiety to rotate and translate as a whole molecular segment.

The exception to this rule is if a heteroatom exists within the aromatic system. This can, occasionally, pucker the ring system and affect the outer crystalline environment therefore affecting molecular packing.

Excluding DOFs that start or end with HC- or -CH, respectively, dramatically reduces the number of DOFs to be included. The argument for this approximation is not dissimilar to the bond length justification where all of the bond lengths have been geometrically optimised with respect to the energy and therefore are unlikely to flex enough to affect the outer crystalline environment. This vastly reduces the computational cost particularly for the larger, polycyclic hydrocarbon molecules.

3.4 Computational Method

The initial list of crystal structures for the original 50 molecules was obtained from previous studies [78, 138]. The gas phase molecular conformers were obtained by implementing the Dmol3 [140] program with a PW91 DFT functional [141] and the double numerical polarised basis set [142].

Structure generation was performed using the Accelrys Polymorph Predictor code from the Cerius2 software package [101] for every molecule over the nine space groups the 50 molecules resided in ($P2_1/c$, $P\bar{1}$, $P2_12_12_1$, $P2_1$, $C2c$, $Pbca$, $Pnma$, $Pna2_1$ and $Pbcn$). These were generated using a number of molecules in the asymmetric unit (Z') value of 1.

All crystal structures were then checked that they were energetically stable within the space group in which they were generated. If they were not stable, the symmetry was relaxed by removing the symmetry constraints of the unit cell. This increases the Z' and shifts the crystal structure to a lower symmetry space group (that is still included in the nine space groups just listed). This process was repeated until a stable structure was obtained. Hence the structures used for the original 50 molecules possessed Z' values of 1, 2, 4 and 8.

The molecular geometries for the additional 4 molecules were geometry optimised and multipoles derived using the B3LYP/6-311G** level of theory with GD3BJ correction [108] within the GAUSSIAN09 software package. As a result, the molecular geometry from each structure file for all of the original 50 molecules were re-energy minimised using this updated DFT-D(B3LYP-GD3BJ)/6-311G** level of theory within GAUSSIAN09 in the interest of consistency.

The relative positions and orientations of the molecules in the unit cell were not modified in any way using this procedure; merely the intramolecular geometry was re-optimised with respect to the energy in the gas phase.

Structure generation was performed on the additional 4 molecules using the Global Lattice Energy Explorer [94] software over the nine space groups listed above. Approximately 2,000 crystal structures per space group were generated that produced a total of approximately 18,000 structures per molecule. This preparation now placed all of the crystal structures for each of the 54 molecules at the same level of theory.

All 54 sets of crystal structures were then energy minimised with respect to their unit cell geometries using DMACRYS, clustered and structurally compared to their corresponding observed structures; the latter two processes both using COMPACK where a 30/30 molecule match to within $RMSD_1 < 0.5\text{\AA}$ at a distance tolerance of 20% and angle tolerance of 20° was required to provide a satisfactorily reproduced crystal structure. This procedure was performed using W99rev potential only.

The set of 54 molecules were also subject to energy minimisation using CrystalOptimizer which employed a B3LYP-GD3BJ/6-311G level of theory for both the intermolecular electrostatic interactions and the intramolecular energy model. This round of calculations allowed molecular flexibility to be incorporated into the calculations. The results were, again, clustered and structurally compared to their corresponding observed structures; latter two processes, again, both used COMPACK where a 30/30 molecule match to within $RMSD_1 < 0.5\text{\AA}$ at a distance tolerance of 20% and angle tolerance of 20° was required to provide a satisfactorily reproduced crystal structure. This procedure was performed using both W99rev and W99orig potentials.

This procedure yielded three sets of results: rigid molecules with W99rev (RevRigid), flexible molecules with W99rev (RevFlex) and flexible molecules with W99orig (OrigFlex). Comparing RevRigid against RevFlex and OrigFlex against RevFlex quantifies the difference in accuracy when using the W99rev potential and the incorporation of molecular flexibility, respectively, into the CSP calculations.

3.4.1 A Note to the Reader

The interpretation of these results must be proceeded with caution. Unfortunately, after these calculations were performed and analysed, it was discovered that a bug in the Day Group's in-house software may have affected the results for the CrystalOptimizer calculations (OrigFlex and RevFlex sets). This issue arises during the multipole calculations within CrystalOptimizer that are calculated by the GDMA program. The in-house software gave the atomic positions in units of \AA but GDMA takes in these distances in units of Bohr radii. Since there are approximately 1.88 Bohr to 1.0\AA these calculations still proceeded but with elongated bond lengths.

It was perhaps unfortunate that these calculations did proceed and not fail as the failure would have highlighted the issue early in the development of the software. For instance, were these distances in metres, for example, the program would have failed.

Nonetheless this is an issue to be aware of but information and conclusions can still be gleaned from the data.

3.5 Results

The results for this section are presented in Table 3.2. The three methods are RevRigid, OrigFlex and RevFlex. Each method possesses a set of N_{Lower} and ΔE values. Both are extracted from the final list of crystal structures that is derived from the CSP calculations.

The final list of crystal structures is then ordered in terms of the total energy (intramolecular energy + lattice energy) with the lowest energy structure occurring as the first entry. When changing methods the absolute energies of the crystal structures change but the ΔE values are only concerned with the relative energies. Using the relative energies lead to values that are directly comparable.

The ΔE values are derived from taking the energy differences between the lowest energy unobserved structure and the lowest energy observed structure. If this value is negative then this also refers to a perfect CSP prediction as the observed structure lies energetically lower than the lowest energy unobserved structure.

The N_{Lower} values refer to the number of crystal structures between the lowest energy unobserved structure and the lowest energy observed structure. This will always be an integer value and is set to 0 if the calculation of this result returns a negative quantity.

If N_{Lower} is 0 or negative then a perfect CSP result is yielded as the observed structure is or lies lower in energy than the lowest energy, calculated structure.

The CSD reference codes, in the left most column of Table 3.2, are accompanied by a shape (Δ , + or O) that refers to the polarity and hydrogen bonding capabilities of that molecule when in the crystalline environment. A '+' refers to molecules that are non-polar and do not participate in hydrogen bonding, a ' Δ ' refers to molecules that are polar but do not participate in hydrogen bonding and an 'O' refers to molecules that are polar and do participate in hydrogen bonding.

Therefore results yielded from molecules that are labelled with a '+' or a ' Δ ' should not change when the potential is changed from the W99orig to the W99rev.

Table 3.2: 54 Small Organic Molecules and Results of the CSP for the RevRigid, OrigFlex and RevFlex methods. All ΔE and $RMSD_{30}$ values are quoted in kJ mol^{-1} and \AA , respectively. Results that are accompanied by an asterisk, *, will be the subject of discussion in Section 3.6.4. The ‘+’, ‘ Δ ’ and ‘O’ symbols refer to molecules that are non-polar and do not participate in hydrogen bonding, molecules that are polar but do not participate in hydrogen bonding and molecules that are polar and do participate in hydrogen bonding, respectively. The ΔE values are derived from taking the energy differences between the lowest energy unobserved structure and the lowest energy observed structure. The N_{Lower} values refer to the number of crystal structures between the lowest energy unobserved structure and the lowest energy observed structure.

CSD Refcode	Space Group	RevRigid			OrigFlex			RevFlex		
		N_{Lower}	ΔE	$RMSD_{30}$	N_{Lower}	ΔE	$RMSD_{30}$	N_{Lower}	ΔE	$RMSD_{30}$
ETHLEN(+)	$P2_1/c$	0	-0.2439	0.268	0	-0.2439	0.268	0	-0.2439	0.268
QQQCIV(Δ)	$P2_1/c$	1	0.2246	0.221	1	0.2246	0.261	1	0.2246	0.261
	$Cmcm$	8	1.7563	0.152	7	1.7563	0.259	7	1.7563	0.259
METAMI(O)	$Pbca$	0	-0.1938	0.518	0	-0.0089	0.445	1	0.0779	0.253
NTROMA(Δ)	$P2_12_12_1$	2	0.6463	0.159	0	-0.1708	0.109	0	-0.1791	0.106
FORMAM(O)	$P2_1/c$	0	-0.5020	0.296	0	-0.5459	0.519	0	-0.0307	0.329
ACETAC(O)	$Pna2_1$	2	0.2528	0.228	8	1.4789	0.249	2	0.6953	0.236
TETROL(O)	$P\bar{1}$	28	5.4894	0.894	15	3.9590	0.634	21	4.5471	0.798
TETROL1(O)	$P2_1$	4	2.3398	0.450	7	2.4824	0.378	1	0.3219	0.471
GLYCIN(O)*	$P2_1/c$	174	19.3548	0.626	17	2.3307	0.310	9	2.0035	0.244
GLYCIN01(O)*	$P2_1$	108	14.5780	0.386	95	7.0585	0.405	50	5.3755	0.057
GLYCIN02(O)*	$P32$	141	16.5937	0.731	83	5.8814	0.278	20	3.4924	0.349
GLYCIN35(O)*	Pn	130	15.9563	2.389	85	5.9210	1.660	121	8.3306	1.739
GLYCIN67(O)*	$P2_1/a$	10	4.7094	0.632	7	1.6598	0.324	11	2.4239	0.280
GLYCIN68(O)*	Pn	-	-	-	91	6.1139	0.676	111	7.8707	0.494

Continued on next page...

Table 3.2 – continued from previous page

CSD Refcode	Space Group	RevRigid			OrigFlex			RevFlex		
		N_{Lower}	ΔE	$RMSD_{30}$	N_{Lower}	ΔE	$RMSD_{30}$	N_{Lower}	ΔE	$RMSD_{30}$
EDAWP(O)	$P2_1/c$	6	1.6929	0.328	6	1.4086	0.278	5	1.1070	0.293
FEPNAP(O)	$P\bar{1}$	1	0.3506	0.255	1	0.0547	0.249	2	0.1293	0.253
TRAZOL(O)	$Pbca$	4	2.8552	0.224	2	1.3922	0.215	2	1.5967	0.212
NEZMUA(O)	$P2_1/c$	0	-0.9480	0.279	0	-4.7774	0.186	0	-5.2498	0.304
SUCANH(Δ)	$P2_12_12_1$	0	-2.1006	0.102	0	-2.5635	0.125	0	-2.5376	0.125
SUCCIN(O)	$Pbca$	0	-1.8009	0.151	0	-1.0521	0.200	0	-0.9461	0.122
OXAZDO(O)	$C2c$	0	-3.4402	0.125	1	0.6751	0.138	0	-0.7284	0.090
VINYLC(Δ)	$P2_1/c$	0	-0.6765	0.592	0	-0.3438	0.607	0	-0.0372	0.607
OXAZIL(O)	$P2_1/c$	0	-1.4134	0.152	0	-1.6581	0.272	0	-1.9674	0.137
YOBQAH(O)	$P2_1/c$	14	1.7385	0.177	2	0.4998	0.146	0	-0.3038	0.147
DUNVEN(O)	$P2_1/c$	65*	8.6990*	0.330	14	2.5289	0.867	13	2.8558	0.697
KOXRIY(O)	$P2_1/c$	1	0.5789	0.445	31	4.8567	0.385	19	2.5714	0.375
PARBAC(O)	$P2_1/c$	0	-0.6805	0.243	0	-1.2689	0.335	0	-0.7047	0.303
JIZREP(+)	$P2_1/c$	9	1.1370	0.221	3	0.4421	0.124	3	0.4421	0.124
XULDUD(Δ)	$Pbca$	5	1.3369	0.589	7	1.3163	0.426	7	1.3166	0.426
XULDUD01(Δ)	$P2_1/c$	18	2.7965	0.470	31	4.0107	0.906	31	4.0068	0.905
BENZEN(+)	$Pbca$	0	-0.0624	1.078	0	-0.0624	1.078	0	-0.0624	1.078
BENZEN03(+)	$P2_1/c$	1	0.4408	0.363	1	0.4408	0.363	1	0.4408	0.363
PRMDIN(Δ)	$Pna2_1$	4	0.5466	0.224	2	0.4645	0.229	2	0.4645	0.638

Continued on next page...

Table 3.2 – continued from previous page

CSD Refcode	Space Group	RevRigid			OrigFlex			RevFlex		
		N_{Lower}	ΔE	$RMSD_{30}$	N_{Lower}	ΔE	$RMSD_{30}$	N_{Lower}	ΔE	$RMSD_{30}$
HXACAN01(O)	$P2_1/c$	1	0.2962	0.227	1	0.3207	0.224	5	1.9765	0.229
MALEHY01(O)	$P2_1/c$	27	6.5968	0.688	38	10.4026	0.155	20	2.8111	0.300
MALEHY10(O)	$P\bar{1}$	10	3.8289	0.414	40	10.7511	0.195	10	1.1021	0.286
MALEHY12(O)	$P2_1/c$	11	4.2950	0.278	45	11.5711	0.338	16	2.0297	0.296
HEZQUY(Δ)	$P2_1/c$	0	-2.0815	0.205	0	-1.7225	0.200	0	-1.7221	0.200
FADMIG(O)	$P2_1/c$	5	5.3056	0.471	3	0.5023	0.633	4	0.8413	0.346
NIVBUP(Δ)	$P\bar{1}$	0	-0.6203	0.212	0	-0.7204	0.227	0	-0.7204	0.227
FULKUS(O)	$P2_1/c$	1	0.0989	0.169	2	0.4520	0.232	1	0.4120	0.216
QAJYIJ(O)	$Pna2_1$	7	1.8068	0.168	9	4.5824	0.245	1	0.2327	0.208
GEYWIQ(Δ)	$Pbca$	2	0.1201	0.155	3	0.9600	0.159	3	0.9600	0.160
NAPTYR(Δ)	$P2_1/c$	0	-0.6347	0.587	0	-0.6346	0.587	0	-0.6346	0.587
QUINCB10	$P2_1/c$	3	0.7166	0.327	10	2.7512	0.158	7	2.2912	0.130
DIHIXL10(Δ)	$P2_1/c$	0	-1.6098	0.155	0	-0.9036	0.155	0	-0.8840	0.156
HOBBOP(O)	$P2_1/c$	1	0.2331	0.114	1	0.2347	0.096	1	0.2347	0.096
HEBBEV(Δ)	$C2c$	0	-4.2629	0.099	0	-3.9630	0.114	0	-3.9630	0.114
BOQQUT(O)	$P2_1/c$	4	1.1332	0.121	1	0.3061	0.205	2	0.4846	0.083
DOGTIC(O)	$P2_12_12_1$	24	10.1156	0.173	10	2.1617	0.214	15	4.3659	0.203
ABEGEU(O)	$P\bar{1}$	1	0.2690	0.322	2	0.0055	0.538	1	0.5438	0.277
AMPHOL(O)	$Pna2_1$	1	1.7480	1.014	1	1.4355	0.395	1	0.7540	0.394

Continued on next page...

3.6 Discussion

3.6.1 ΔE Values

Figure 3.1 compares the differences in ΔE values for the RevRigid, OrigFlex and RevFlex methods. The diagonal lines are separated by $\sqrt{2}$ kJ mol⁻¹ and encloses a zone that represents an insignificant change in ΔE between the two methods. The spacing between the lines is arbitrary and is also designed to allow for easier comparison between results.

Data points that lie in the lower-left quadrant possess negative ΔE values for both methods and therefore are ‘perfect’ CSP results. Data points that lie in the upper-left quadrant represent bad data as these are points that yield negative and positive ΔE values for the ‘original’ and ‘improved’ methods, respectively. This would suggest that the ‘improvements’ are not profitable for these molecules.

Data points that lie the upper-right quadrant and above the upper diagonal line show a more significant decrease in accuracy when using the ‘improved’ method. And points that lie within the lower-right quadrant and below the lower diagonal line in the upper-right quadrant represent significant improvements to the CSP ΔE values when using the ‘improved’ method. The exact number of data points in each quadrant are summarised in Table 3.3.

Figure 3.1a compares the ΔE values for the RevFlex and RevRigid methods. The comparison quantifies the effect of including molecular flexibility into the CSP calculations. Using the information from the previous paragraph, it is obvious that the majority of points lie below the lower diagonal line. From Table 3.3 this number is quantified to 24. This is an extremely successful result as almost half of the CSP results are improved by incorporating molecular flexibility. This is a clear justification that flexibility vastly increases the accuracy of the CSP calculations even for these small organic molecules where flexibility is limited.

METAMI and XAYCIJ molecules possess a change in sign of the ΔE values, albeit these are very similar, when the level of theory is changed. The reason for these results are unclear although referring the reader back to Section 3.4.1, A Note to the Reader, may provide the explanation in that the units required for the multipole calculation were input as Å instead of Bohr radii.

Figure 3.1b compares the differences in ΔE values for the OrigFlex and RevFlex methods. This comparison quantifies the effect of using the W99rev against the W99orig potential. The differences when using this change do not affect molecules that do not participate in hydrogen bonding. This can be observed in 3.1b in that the blue Δ ’s

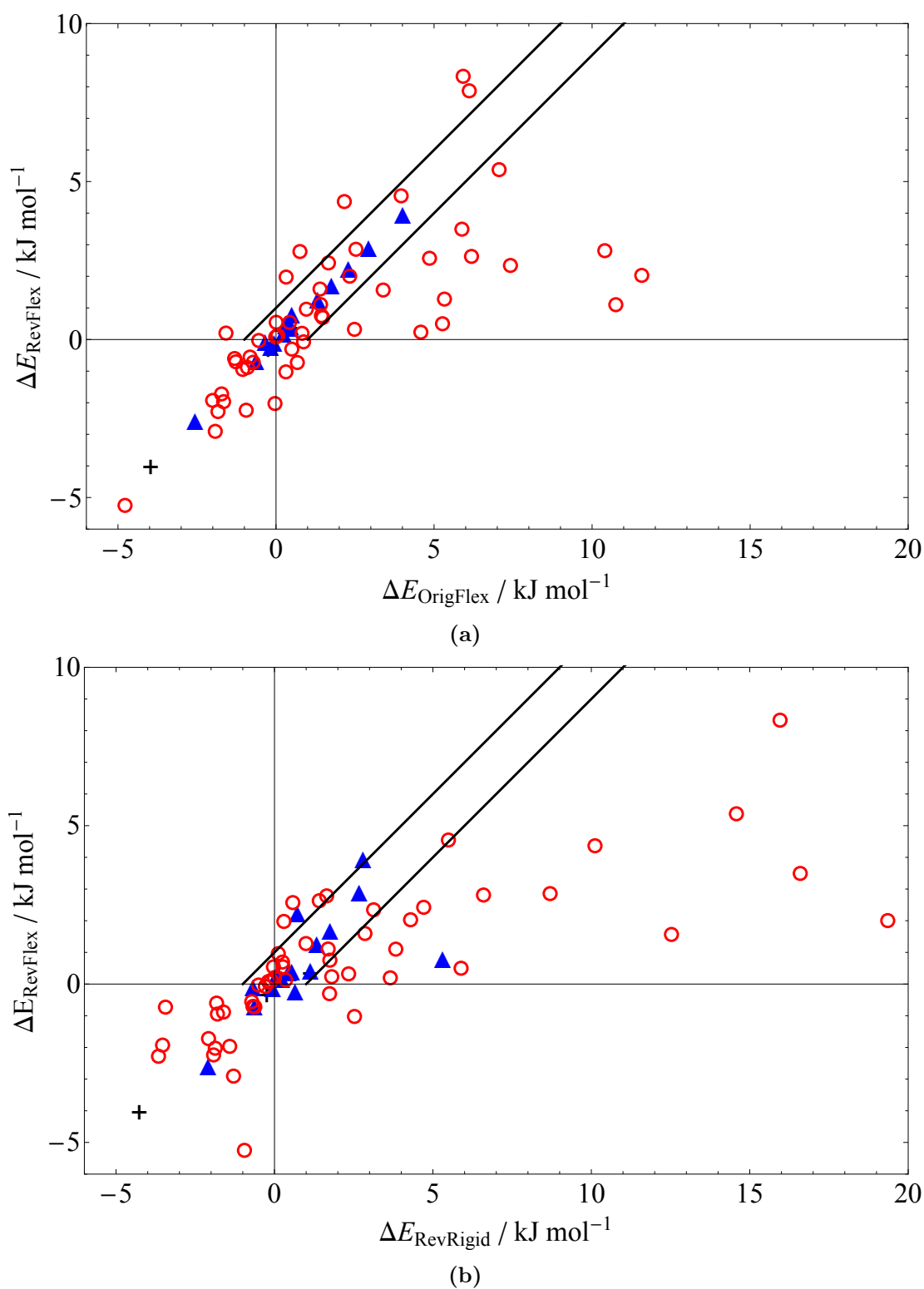


Figure 3.1: (a) and (b) shows the comparison of the ΔE values for the RevFlex vs OrigFlex and RevFlex vs RevRigid methods, respectively. The black '+', blue ' Δ ' and red 'O' refer to molecules that are non-polar and do not participate in hydrogen bonding, molecules that are polar but do not participate in hydrogen bonding and molecules that are polar and do participate in hydrogen bonding, respectively. The black lines are present to aid in comparison of the two methodologies.

Quadrant	RevFlex vs RevRigid	RevFlex vs OrigFlex
Upper Left	2	1
Lower Left	19	22
Upper Right (above line)	7	4
Upper(below line)/Lower Right	24	17

Table 3.3: collection of ΔE data points in each quadrant of Figure 3.1.

and “+”’s lie on an $y = x$ line. Although this result is not surprising, it reinforces the reliability of the methods used in that sensible results are yielded from these calculations.

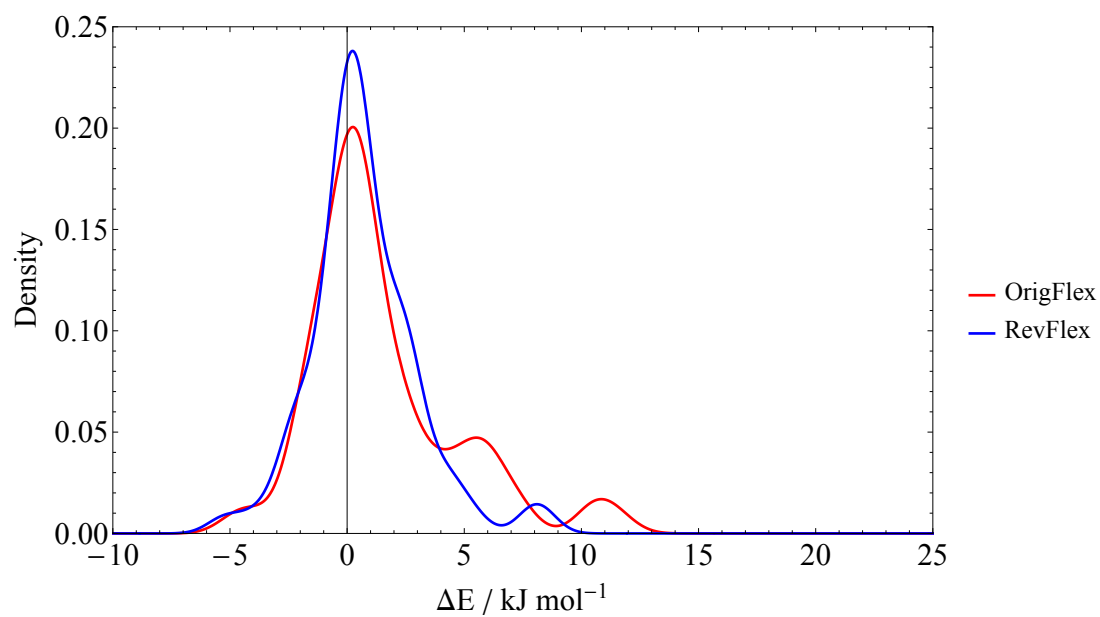
Observing the values in the right most column of Table 3.3, the number of improved results (Upper(below line)/Lower Right) lies at a slightly lower value than the flexibility comparison (17 and 24 respectively). However, using the W99rev potential still vastly increases the accuracy of the ΔE values. In addition, this method modification yields less bad results (those that lie in the upper left and upper right (above line) regions). Nonetheless, the incorporation of both of these methods increase the accuracy of the CSP ΔE values and bring the energy of the observed crystal structures closer to the calculated global energy minimum.

The addition of the W99rev adds no computational cost to the calculations. The flexibility does increase the computational cost (with respect to time) of the calculations. The calculation time for a typical ‘rigid’ CSP is on the order of magnitude of CPU hours. When flexibility is included, this calculation time can increase by an order of magnitude to tens of CPU hours. Therefore care must be taken when deciding to include, and to what extent, flexibility into the CSP calculations.

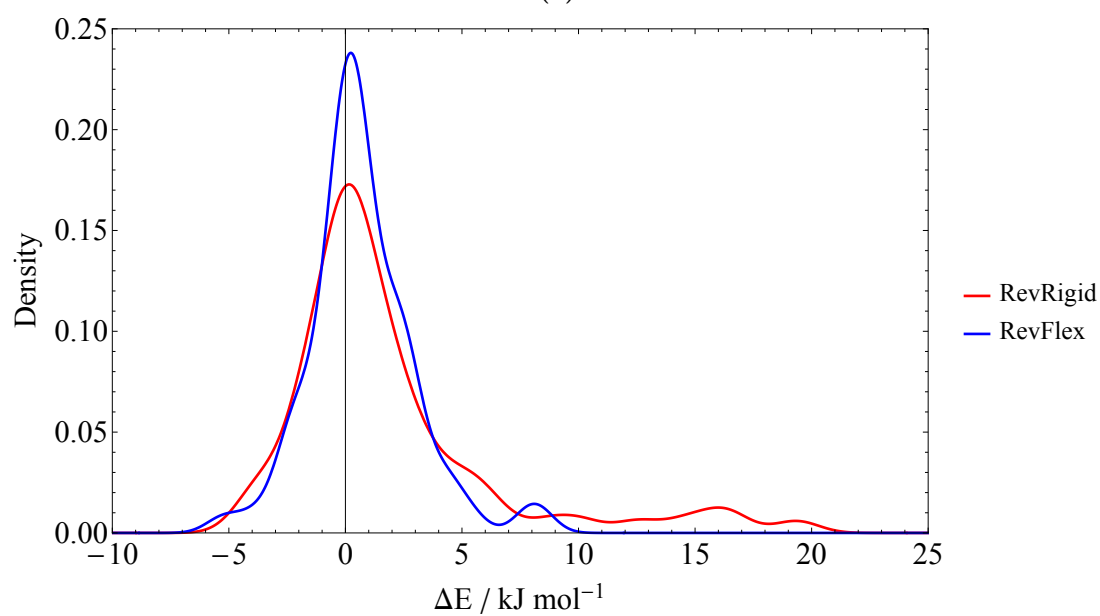
Another useful method for analysing these ΔE results is to use a smoothing kernel density (SKD) estimation function. The ΔE statistic can be modelled as a continuous random variable and a function can be fitted to the data. Hence a plot of the normalised probability distribution density as a function of the ΔE values is displayed in Figure 3.2. The convention of these figures is that the method that would be expected to give more accurate results possesses a blue trace. These plots allow an easy visual comparison of the ΔE values and demonstrates a clear increase in performance of CSP results when using the W99rev potential or including flexibility. This is shown by a shift in density from the right to left upon applying these modifications.

The integration values of the SKD functions within the bounds of -10 kJ mol^{-1} to 1 kJ mol^{-1} are displayed in Table 3.4. This allows a quantification of the shift in density.

The ΔI values show a 5.5% and 6.3% increase in ΔE values that are within the bounds when including flexibility and W99rev, respectively. These percentages quantify the



(a)



(b)

Figure 3.2: (a) and (b) shows smoothing kernel density plots for the ΔE values for the RevFlex versus RevRigid and OrigFlex versus RevFlex methodologies, respectively.

Method	I	I(RevFlex)	ΔI / %
RevRigid	0.553	0.608	+5.500
OrigFlex	0.545	0.608	+6.300

Table 3.4: The integration values within the bounds of -10 kJ mol^{-1} to 1 kJ mol^{-1} of the smoothing kernel density functions displayed in Figure 3.2.

effect of these two modifications when performing CSP on small organic molecules that participate in hydrogen bonding.

3.6.2 N_{Lower} Values

Figure 3.3a compares the N_{Lower} values when changing the potential from W99orig to W99rev where it is clear that there is a shift in density from right to left. This is a slight shift but nonetheless it is consistent with previous results from Section 3.6.1.

Bin ‘0’ shows an additional 3 molecules now yield perfect CSP results upon the inclusion of flexibility. Bin ‘21+’ also shows a reduction in N_{Lower} values that give a more consistent set of results. From Table 3.2, GLYCIN possesses 4 and 3 values in this bin for the OrigFlex and RevFlex methods, respectively. The details for why this molecule appears untameable will be discussed further in Section 3.6.4.

Figure 3.3b compares the N_{Lower} values when including flexibility into the CSP calculations. There is not an obvious shift in density that would have been expected. However, density shifts from bin ‘21+’ to the lower valued N_{Lower} bins which suggests flexibility tames some of the molecules that give erratic results. GLYCIN has a strong presence in the bin ‘21+’ with 4 and 3 structures for the RevRigid and RevFlex methods, respectively.

3.6.3 $RMSD_{30}$ Values

The $RMSD_{30}$ matches are analysed using a similar technique from Section 3.6.1 by expressing the $RMSD_{30}$ of one method as a function of the other. The method that would be expected to give more accurate values is on the y-axis and therefore any improved results will appear underneath the lower black diagonal line.

Figure 3.4a shows 5 improved and 2 worsened results. This distribution is trivial and could suggest that using the W99rev potential yields slightly more geometrically accurate crystal structures but does not possess any strong correlation between the two methods.

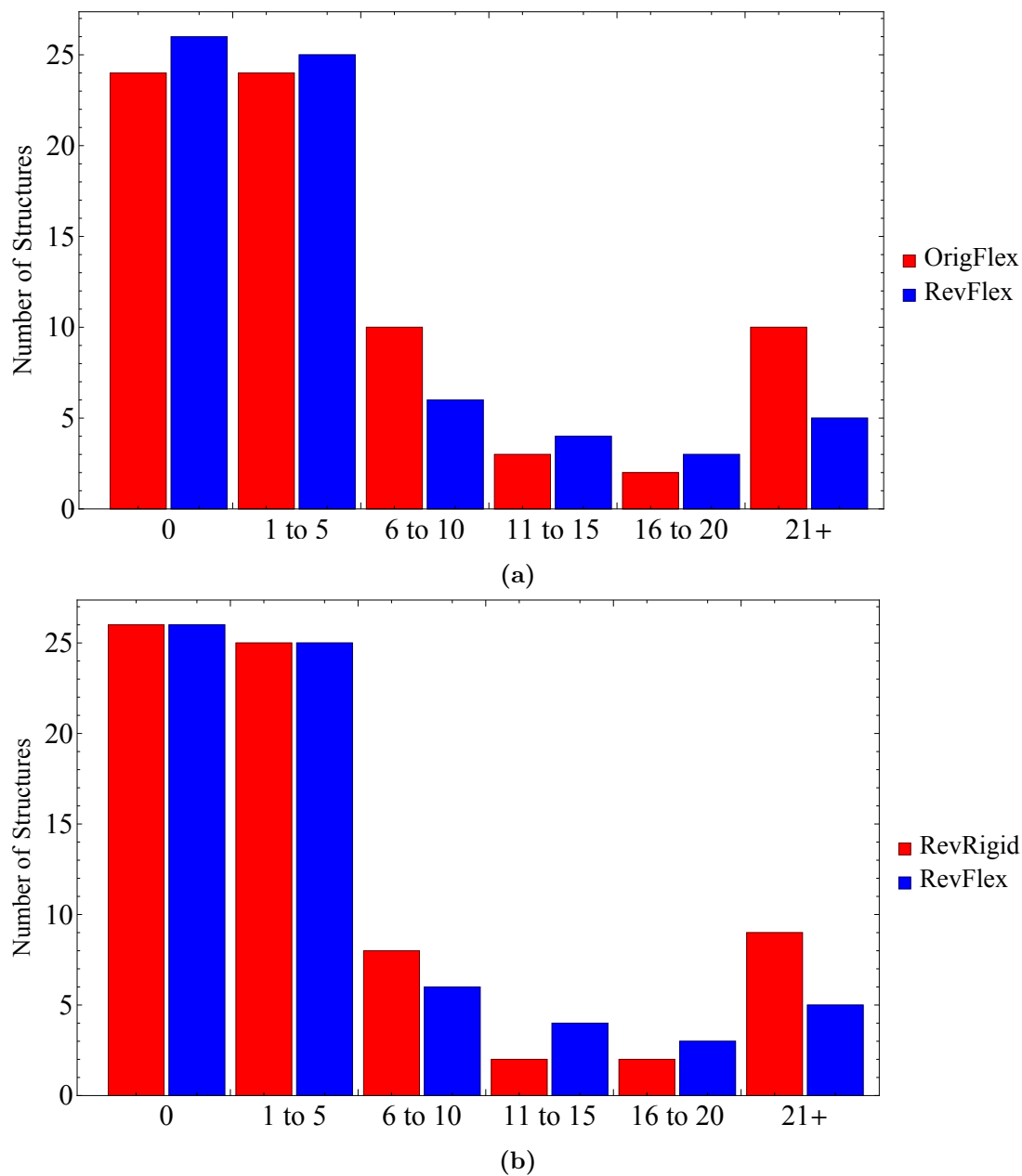


Figure 3.3: (a) and (b) shows the comparison of the N_{Lower} values for the OrigFlex versus RevFlex and RevFlex versus RevRigid, respectively. Bin '0' represents a 'perfect' CSP result. The colour scheme follows the method that is expected to perform better is in blue.

An unexpected result within figure 3.4a is that 3 of the polar-non-hydrogen bonding molecules (blue Δ 's) yield different $RMSD_{30}$ values. This is counter intuitive as these two methods should yield the same structure. The exact reason for this remains unclear but a possible explanation is to attribute this to COMPACK where small changes in molecular geometry cause larger deviations in $RMSD_{30}$ values.

Another explanation could be that because the molecules are polar there is a slight polarisability experienced by the neighbouring molecules in the crystal. This can affect the packing of the molecules and therefore lead to different $RMSD_{30}$ values; although the magnitude of this intermolecular interaction is not significant enough to classify it as a hydrogen bond. It is also worth mentioning that this could also be attributed to the issues raised in Section 3.4.1 where small deviations in the multipole moments about a molecule can lead to these more inaccurate crystal structures.

Figure 3.4b shows a broader distribution in $RMSD_{30}$ results. There are 15 and 9 data points that lie below and above diagonal lines, respectively. This is not unexpected as the RevFlex method allows the intramolecular geometry to also vary and therefore the crystal structure can become significantly more geometrically similar to the observed crystal structure.

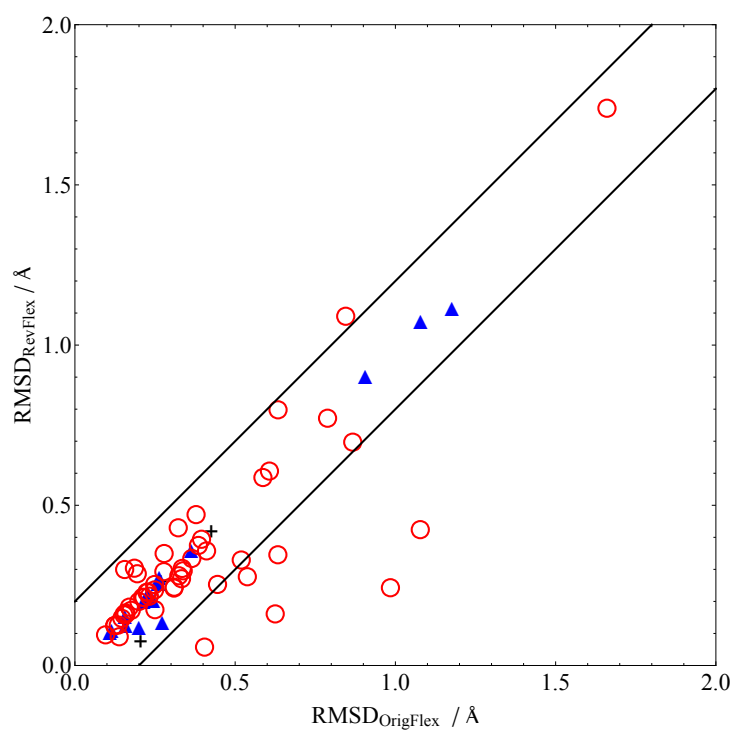
However, the counter argument can also justify the reason why some $RMSD_{30}$ values become worse. The intramolecular geometry can relax into a stable potential energy minimum but this can be geometrically further away from the starting intramolecular geometry and yield higher $RMSD_{30}$ values.

3.6.4 Specific Cases

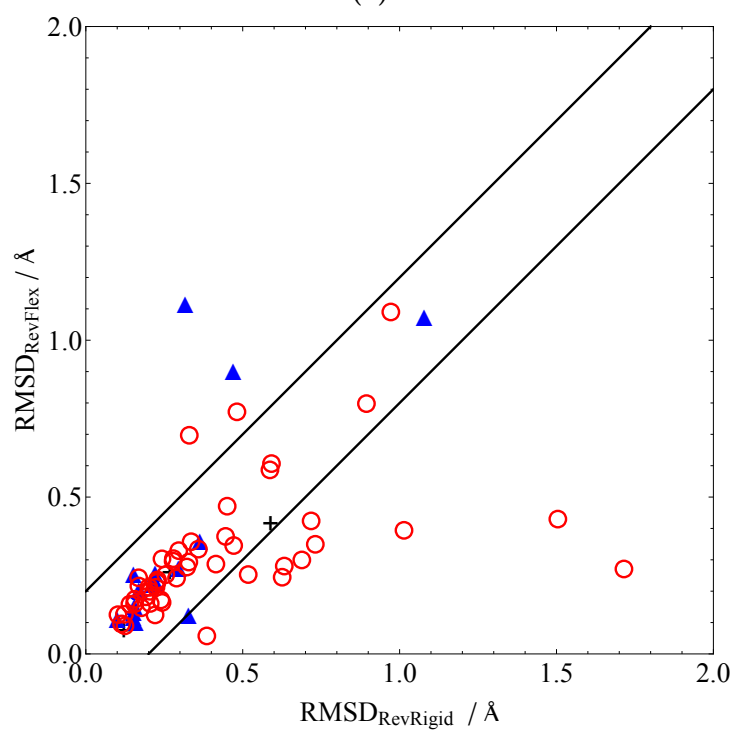
- **GLYCIN:** the N_{Lower} and ΔE values are extremely variable across all three methods. The energetic ordering of the polymorphs also change and it is not clear which polymorph is the most energetically stable. An explanation for the high values for the RevRigid results can be partially attributed to the differences in gas phase and in-crystal molecular geometry (a more complete explanation of this reason is discussed below in the DUNVEN bullet point).

Since GLYCIN is an amino acid, the gas phase molecule exists in a neutral form whereas the in-crystal molecule is a zwitterion. This issue was overcome by fixing the N-H bond lengths during the gas phase geometry optimisation to prevent the neutral molecular conformer being formed.

Zwitterions possess formal charges on certain regions of the molecule. Therefore this charge will propagate throughout space and affect the electrostatic distribution



(a)



(b)

Figure 3.4: (a) and (b) shows comparison of the RMSD_{30} values for the RevFlex versus RevRigid and OrigFlex versus RevFlex, respectively. The black '+', blue ' Δ ' and red 'O' refer to molecules that are non-polar and do not participate in hydrogen bonding, molecules that are polar but do not participate in hydrogen bonding and molecules that are polar and do participate in hydrogen bonding, respectively. The black lines are present to aid in comparison of the two methodologies.

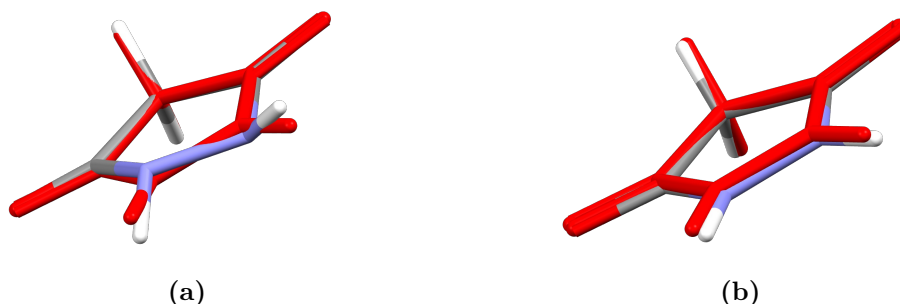


Figure 3.5: (a) and (b) shows the geometric comparison of the observed DUNVEN molecular geometry (red) and the RevRigid and RevFlex intramolecular geometry (coloured by element), respectively.

of the neighbouring molecules. The magnitude of these interactions is not fully described by the W99rev potential.

Both of these reasons are accounted for, to some extent, by the RevFlex method. To fully account for these electrostatic interactions a PCM model would need to be used. This allows a molecule to be energy minimised with respect to the geometry in a potential of the neighbouring molecules (instead of a vacuum).

In addition since GLYCIN is arguably the most polar molecule in the set of 54 molecules, the issues raised in Section 3.4.1 can cause vast deviations in the electrostatics about the molecule and therefore led to inaccurate crystal structures. This is a highly probable reason as to why GLYCIN is consistently ‘misbehaved’ within these calculations.

- **DUNVEN:** this molecule possesses exceptionally high ΔE and N_{Lower} values for the RevRigid method before these values return to more realistic quantities when using the OrigFlex and RevFlex methods. The difference between these values occurs between the rigid and flexible calculation; not between the W99orig and W99rev.

Figure 3.5 shows the difference in intramolecular geometry between the calculated observed structure for the RevRigid and RevFlex methods and the empirical observed structure. This is an ideal example of why flexibility is important in CSP.

Since the intramolecular geometry in the gas phase differs from the in-crystal geometry, the starting geometry is in the wrong conformation and holding the molecular geometry rigid prevents the correct molecular geometry being obtained and large N_{Lower} and ΔE values result.

- **BZAMID06:** this polymorph of BZAMID is not found for any of the methodologies. The observed crystal structure of BZAMID06 was extracted from the CSD and fed through all three methods and none of them yielded a structure that matched the observed structure.

The starting and final intramolecular geometries are not dissimilar but the packing differs. The reason for this anomaly is unclear but the issues explained in Section 3.4.1 may provide the explanation.

- **FOYBOL:** the CSP calculations using the RevRigid method fails to locate the observed crystal structure in the final list of structures. The reason for this is believed to be due to difference in molecular geometry between the gas phase between in-crystal conformations (the same reason described in DUNVEN). However, the final crystal structures were taken from the RevFlex method and re-optimised using the RevRigid method the observed structure was now found in the final list. The values in Table 3.2 are the values derived from this methodology.

3.6.5 Conclusions

The replacement of the W99orig potential with the W99rev potential offers a vast increase in accuracy of CSP calculations for molecules that are capable of hydrogen bonding. This modification adds no computational cost to the calculations and places an additional 6.3% of observed structures within 1 kJ mol^{-1} of the lowest calculated unobserved structure. The N_{Lower} values are also lowered which brings the molecules that possessed erratic results to more realistic values. In addition, the $RMSD_{30}$ values are reduced for some molecules which suggests that the crystal structures are also being predicted with more geometric accuracy.

The inclusion of flexibility into the CSP calculations also lowers the ΔE values and encompasses a 5.5% increase in the number of observed crystal structures existing within 1.0 kJ mol^{-1} of the lowest calculated unobserved structure. This modification does incur a computational cost of roughly 1 order of magnitude and therefore should only be used on molecular DOFs that possess low force constants and are likely to be distorted by crystal packing forces.

The N_{Lower} values appear not to improve for molecules that already perform well in rigid CSP calculations. However, the inclusion of flexibility plays strong role in ‘calming’ molecules that exhibit high N_{Lower} values and lower them to more comfortable values. This is shown in the specific examples of DUNVEN and GLYCIN where the gas phase and in-crystal intramolecular geometry differs but the inclusion of flexibility allows the potential interconversion between the two conformers. Rigid-body CSP does not allow this change which can lead to small changes in intramolecular geometry significantly affecting the crystal packing and therefore produce structures which are geometrically far away from the observed crystal structure.

A large quantity of $RMSD_{30}$ values improve upon the inclusion of flexibility. This is unsurprising as both the inter- and intramolecular geometry can be modified to find the most stable potential energy minimum. However, a smaller proportion of molecules also experience an increase in $RMSD_{30}$ values. This can be justified using the same argument that since intra- and intermolecular geometries can be modified. This changes the PES to possess different potential energy minima which the crystal structures can become trapped in. These are close enough to be classified as the observed structure but lie marginally geometrically further away and so the $RMSD_{30}$ values increase.

Although some of these anomalous results can be attributed to the issue outlined in Section 3.4.1, the overall trend of the results are generally encouraging. This chapter has proved that for even small organic molecules with limited flexibility, the inclusion of this flexibility is paramount for obtaining more accurate crystal structures.

The methods outlined in this chapter to include molecular flexibility within CSP remain as the industrial benchmark. However, the next chapter will now push this benchmark beyond its current limit and attempt to uncover an alternative, and potentially more powerful, method to treat molecular flexibility in CSP.

Chapter 4

Principal Displacement Conformational Searches

The work in this chapter was performed in collaboration with Dr Peter J Bygrave (PJB) and Dr David H Case (DHC); both post-doctoral researchers within the Graeme Day research group. The author led stages 1 through 6 listed in Section 4.5, stages 7 and 8 and stages 9 through 11 were led by DHC and PJB respectively. Nonetheless, the author was heavily involved in all stages of this research.

4.1 Introduction

This chapter will present a principal displacement conformational search algorithm that will be implemented within the CSP process. This method shares the same goal as the CrystalPredictor algorithm, Section 2.2, but attempts to exploit the properties of molecular principal displacements as opposed to torsion angles. Nonetheless, the molecular principal displacements and torsion angle procedures will be compared throughout this chapter so that a consistent understanding of this novel method relative to current procedures can be ascertained.

It is important to appreciate that the principal displacement methodology is presented as a comparison to the current torsion procedure, not as a competitor. This chapter attempts to provide an alternative methodology for performing flexible molecular search procedures and certainly does not attempt to define itself as *the* methodology.

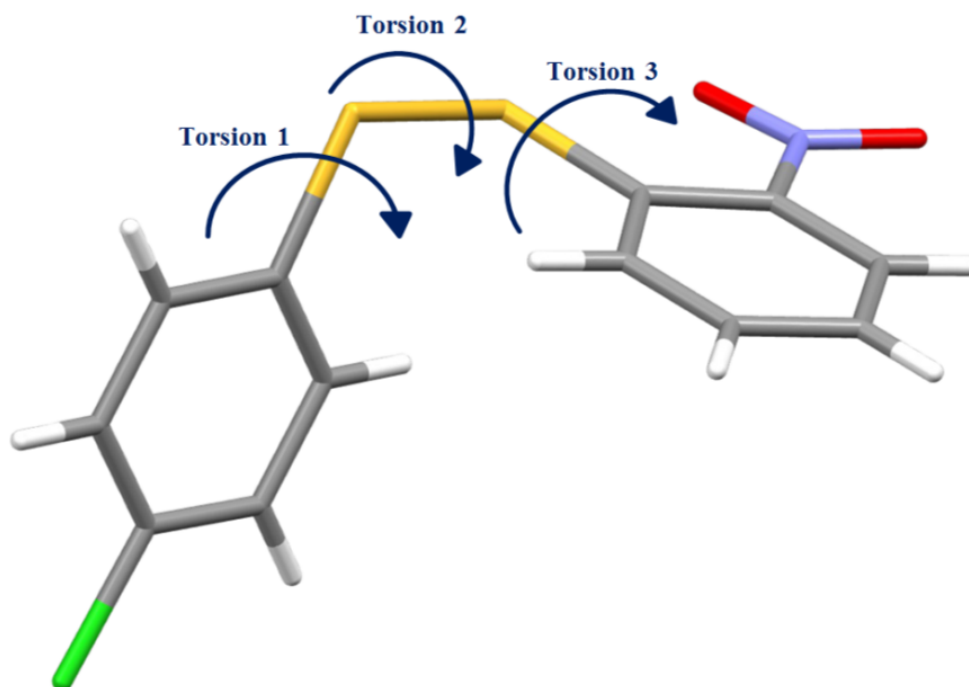


Figure 4.1: The in-crystal geometry of the FUQLIM polymorph (coloured by element). Note the 3 soft torsion angles (*Torsion 1*, *Torsion 2* and *Torsion 3*) formed by the disulfide bridge that connects the two aromatic systems.

4.2 Motivation for a Flexible Molecule Search Procedure

The test molecule for this chapter is bis(2-nitrophenyl,4-chlorophenyl)disulfide but will be referred to by its CSD reference code FUQLIM [143, 144], Figure 4.1. There are currently 3 known polymorphs of FUQLIM: A (FUQLIM), B (FUQLIM01) and C (FUQLIM02) where only the former 2 are presently recorded in the CSD. The latter was obtained through private communication from experimental collaborators who discovered polymorph C.

Each polymorph possesses a unique molecular conformation in the crystalline environment, Figure 4.2. The molecular conformer of polymorph A possesses a more planar shape than the molecular conformations of polymorphs B and C that allows a lower lattice energy within the crystal structure. The molecular conformations of polymorphs B and C are extremely similar to one another but are in contrast to A as the disulfide torsional angles in the former allow the nitro-benzyl ring to twist and become more isolated from the chloro-benzyl system. The differences between the molecular conformations of B and C are small but can be accounted for by the extent of the nitro-benzyl twist.

Nonetheless, these conformational differences gave rise to the previous CSP calculations by Bygrave *et al.* [145] on FUQLIM reported that a rigid body structure generation

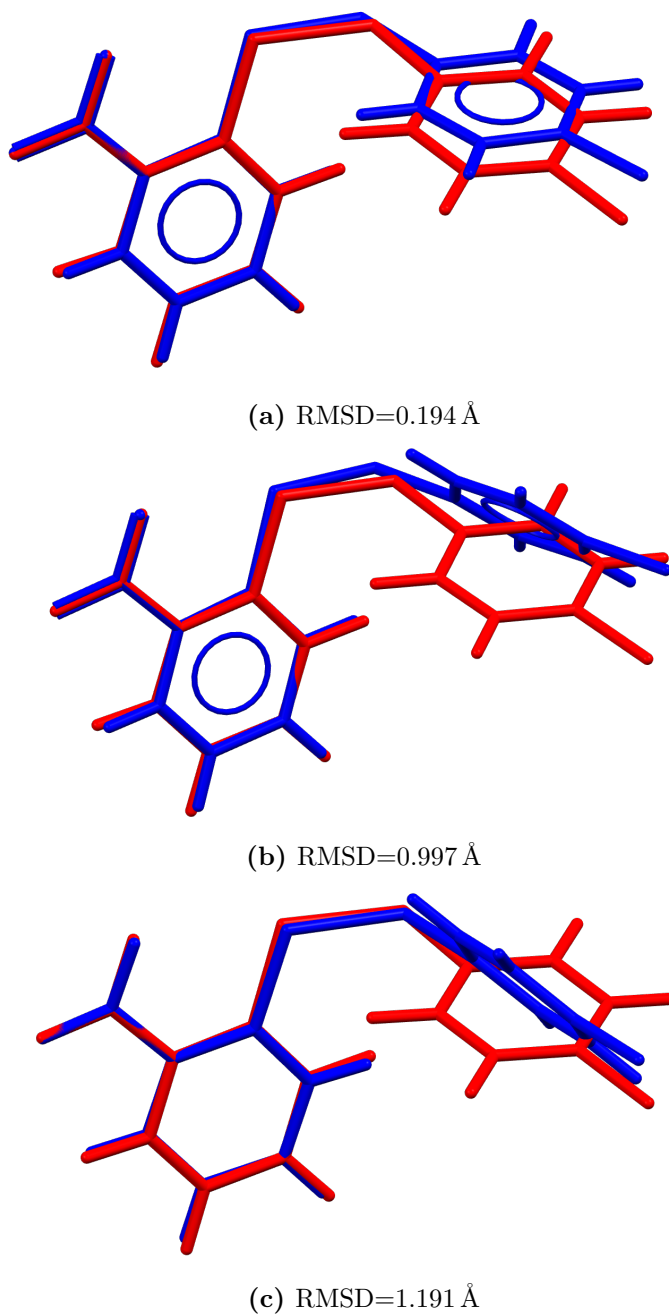


Figure 4.2: Illustration of the gas phase FUQLIM molecular conformation (red) against the in-crystal conformation (blue) for the 3 known polymorphs A, B and C in figures (a), (b) and (c) respectively.

procedure only using gas phase conformer fails to predict the existence of polymorphs B and C. That is, these 2 crystal structures are not present in the final list of structures that is yielded from the CSP process. This was attributed to the distortion of the gas phase conformer in the crystalline environment.

Figure 4.2 shows the in-crystal conformations for all 3 of the known polymorphs overlaid by the gas phase conformer where the former conformers geometry optimise to the latter

in vacuum. The geometrical differences between the gas phase and the conformation of polymorph A are insignificant (RMSD=0.194 Å). However, this observation cannot be transferred to polymorphs B and C whose in-crystal molecular conformations are more significantly distorted (RMSD=0.997 Å and 1.191 Å, respectively) in comparison to their gas phase conformers.

Bygrave surmised that, since the molecular conformations were held rigid during the structure generation phase, the molecular conformations could become ‘trapped’ inside their respective potential energy wells. Therefore the molecular conformations required to yield polymorphs B and C could not be energetically sampled during the energy minimisation phase that was performed using CrystalOptimizer.

An alternative approach to thinking about this issue is that the regions of the PES that would have led to polymorphs B and C were not sampled. This is due to the lack of flexibility during the structure generation process where only energetically stable, gas phase conformers are used for the CSP calculations. The geometries of these conformers are not able to be varied during the structure generation phase and therefore the PES is not wholly being explored.

The PES in Figure 2.10b demonstrates this point where the observed crystal structure lies at the global minimum but the structure generation process only allows crystal structures to be generated in the higher energy local minimum, not about each minimum. Therefore during the energy minimisation phase of the CSP process, crystal structures are not expected to overcome any potential energy barriers to explore the other regions of the conformational space.

In addition, to explore this region of the PES, the molecular conformer would be required to enter an ‘un-physical’ transition state where its atoms would clash with the atoms of neighbouring molecule(s) before returning to a valid, stable crystal structure.

The PES that describes the lattice energy of the crystal structure can change its topology depending on the molecular conformation. The in-crystal molecular conformers of polymorphs A, B and C lie in 3 separate potential energy wells that are separated by potential energy barriers.

However, it is interesting to note that this barrier does not exist for the isolated molecule. Therefore, if the starting geometry for a gas phase molecular geometry optimisation was the in-crystal polymorph A conformation, the final geometry would be the red geometry illustrated in Figure 4.2. Likewise, the molecular conformations of polymorphs B and C would also geometry optimise in the gas phase to this red molecular conformation displayed in figure 4.2.

Principal Displacement	Force Constant
1	0.9
2	2.4
3	4.2
4	43.1
5	45.3
6	62.2

Table 4.1: The 6 lowest force constants, in mDyne \AA^{-1} , for the gas phase conformer of FUQLIM calculated by B3LYP-GD3BJ/6-311G** level of theory.

The inaccuracy for sampling the conformational space about each molecular conformer demonstrates the need for a modified structure generation process where molecular flexibility is taken into account at every stage of the CSP process. In particular for the FUQLIM case, during the structure generation phase.

4.3 Degrees of Freedom for FUQLIM

The FUQLIM molecule possesses 3 soft torsion angles that connect the 2 aromatic ring systems together. These torsions are distributed over the CSSC disulfide bridge and are defined as follows: C-C-S-S from the phenyl-chloride ring (*Torsion 1*), C-S-S-C (*Torsion 2*) and C-C-S-S from the nitro-phenyl ring (*Torsion 3*).

Of course other torsion angles exist in this molecule but these 3 soft torsions are those that are responsible for large geometrical distortions. Therefore they are more significant when attempting to convert between the in-crystal and gas phase geometries of the FUQLIM molecule.

Table 4.1 shows the 6 lowest force constant values for the principal displacements of the FUQLIM molecule calculated at the B3LYP-GD3BJ/6-311G** level of theory. The large increase in force constant from principal displacement 3 to 4 allows this study to be limited to combinatorial displacements along these first 3 only. This is based purely on chemical intuition at this point but Chapter 6 attempts to identify a force constant value tolerance which allows for more accuracy in choosing which principal displacements are required for this methodology.

These first 3 principal displacements are visualised in Figure 4.3. All 3 perform significant distortions about *Torsion 1*, *Torsion 2* and *Torsion 3* as well as other smaller geometrical distortions. Upon this visualisation, it can be assumed that, as well as the

large increase in force constant, that these 3 principal displacements in combination are capable of producing an accurate representation of the in-crystal molecular geometry when using the gas phase conformer as a starting point. In addition, only implementing 3 principal displacements will vastly reduce any computation that is required during the search procedure (as opposed to all $3N - 6$ principal displacements).

4.4 Molecular Conformational Space

For the comparative feature of this chapter, 2 conformational spaces must be defined and searched: the principal and torsional displacement spaces.

4.4.1 Defining the Conformational Spaces

The in-crystal molecular geometry for FUQLIM was extracted from the CSD and optimised with respect to the intramolecular energy before conducting a principal displacement calculation. The latter was performed by diagonalising the Hessian matrix of the molecular energy. Both of these procedures were performed at the DFT-D (B3LYP-GD3BJ/6-311G**) level of theory. The principal displacements with the 3 lowest force constants were extracted along with their respective eigenvectors, Figure 4.3.

Previous research by Thompson & Day [120], discussed in Section 2.2, states that the largest difference in intramolecular energy between the gas phase and in-crystal geometries is 22.5 kJ mol^{-1} . Therefore the potential energy surface must be fitted to energy bounds beyond this value to allow a good fit within the space that is bound by the 22.5 kJ mol^{-1} . Hence an energy bound of 30 kJ mol^{-1} above the equilibrium energy was selected.

To determine the shape of the space, and hence the bounds of the displacement that needed to be sampled, a Simplex [72] brute force numerical minimisation of DFT-D energies were required to determine the 30 kJ mol^{-1} energy bounds of each principal displacement in both positive and negative displacement directions. These displacement values for the first, second and third principal displacements exist at the minimum and maximum displacements from $(-9.40 \text{ \AA} \text{ to } 5.71 \text{ \AA})$, $(-6.19 \text{ \AA} \text{ to } 5.18 \text{ \AA})$ and $(-2.89 \text{ \AA} \text{ to } 6.18 \text{ \AA})$ respectively. The shape of the conformational space is displayed in Figure 4.4a. To sample the intramolecular energy of the space defined by *Torsion 1*, *Torsion 2* and *Torsion 3* angles, DFT-D torsional energy scans were performed. It was found that *Torsion 1* was able to perform a full revolution of 2π radians without exceeding the 22.5 kJ mol^{-1} energy limit. *Torsion 2* and *Torsion 3* can tolerate up to $\pm\frac{\pi}{3}$ radians before breaking the energy limit. This conformational space is shown in Figure 4.4b.

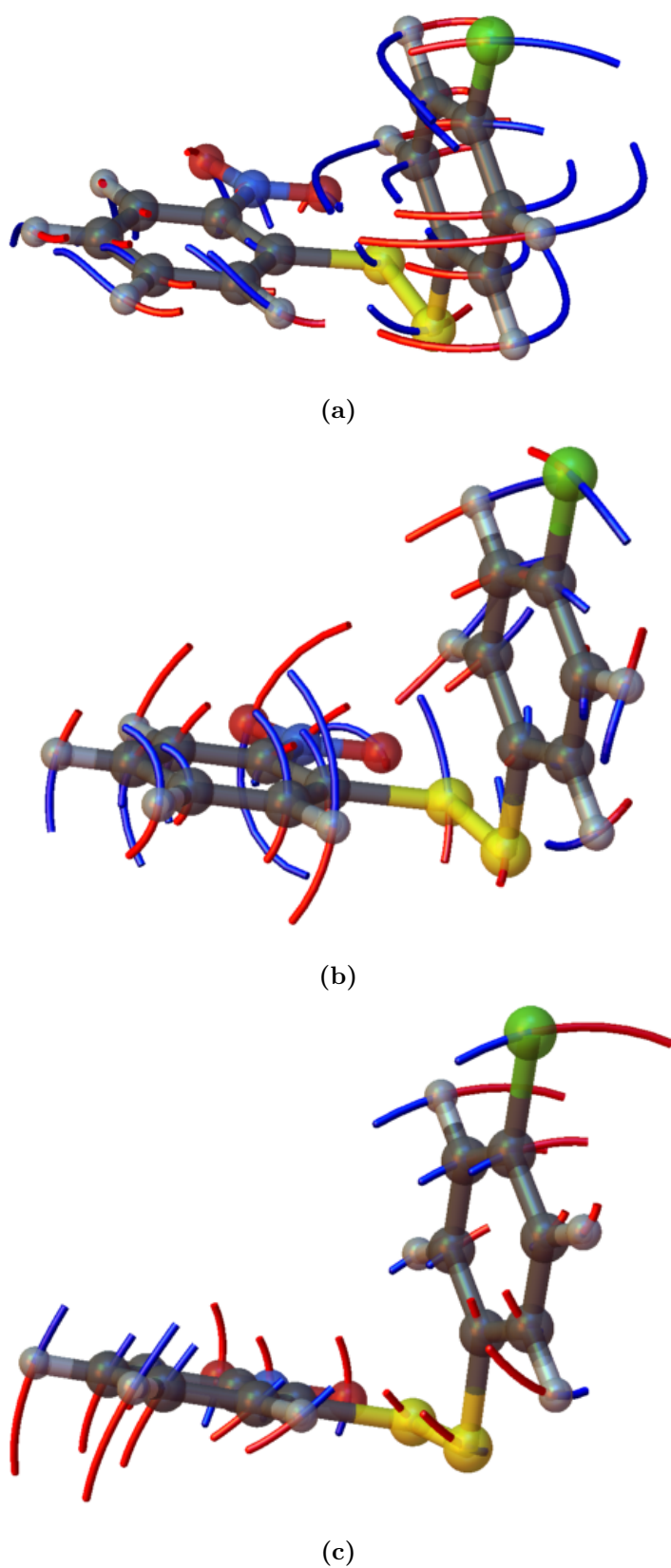
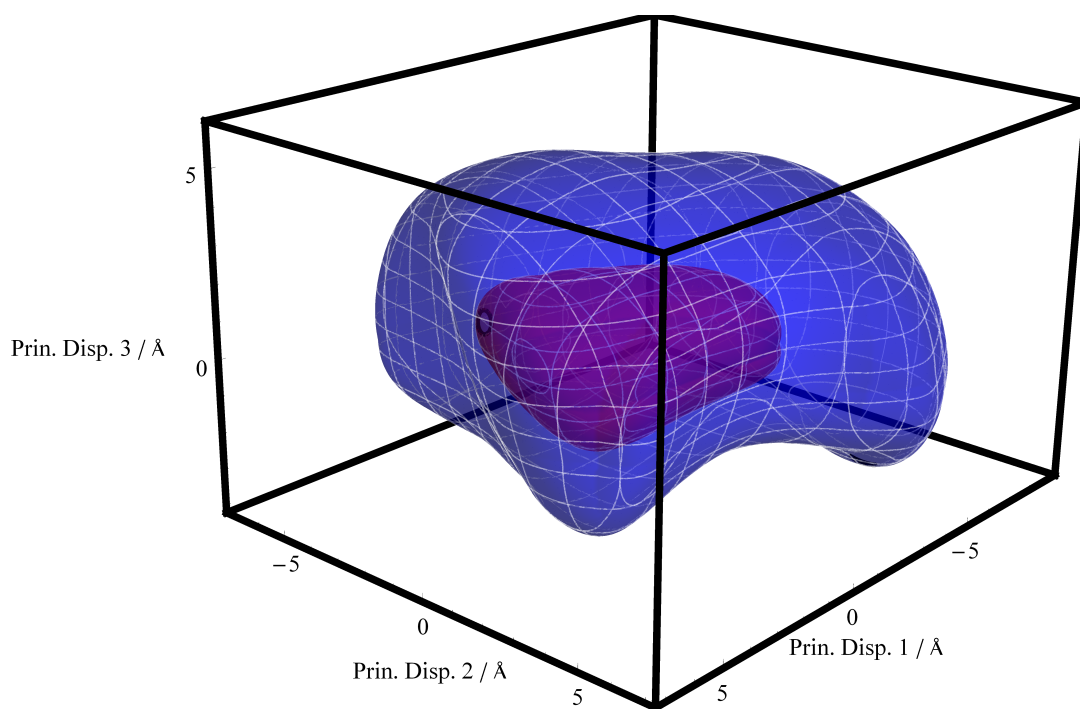
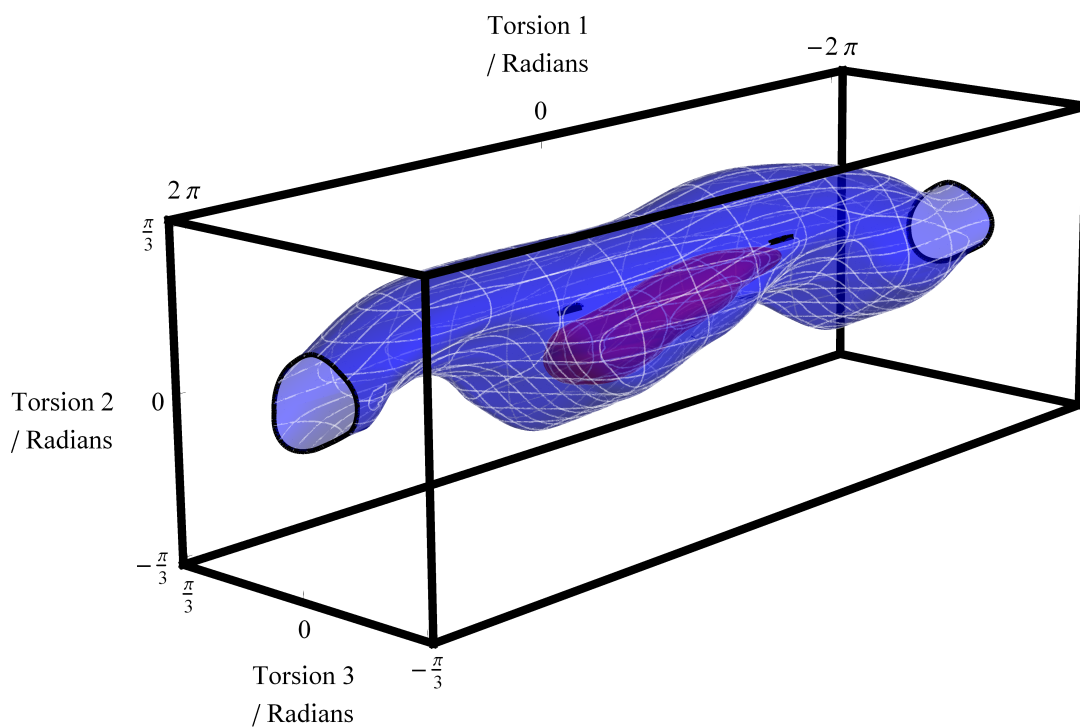


Figure 4.3: The principal displacements for FUQLIM with the first, second and third lowest force constants in (a), (b) and (c) respectively. The curved red and blue lines represent the paths of the atoms in the positive and negative directions, respectively, of the principal displacement in curvilinear space about the gas phase geometry.



(a) Principal Displacement Space



(b) Torsion Displacement Space

Figure 4.4: Principal and torsional displacement spaces for 22.5 kJ mol⁻¹ and 5.0 kJ mol⁻¹ energy limits above the energy of the equilibrium geometry for the FUQLIM molecule highlighted in blue and purple respectively.

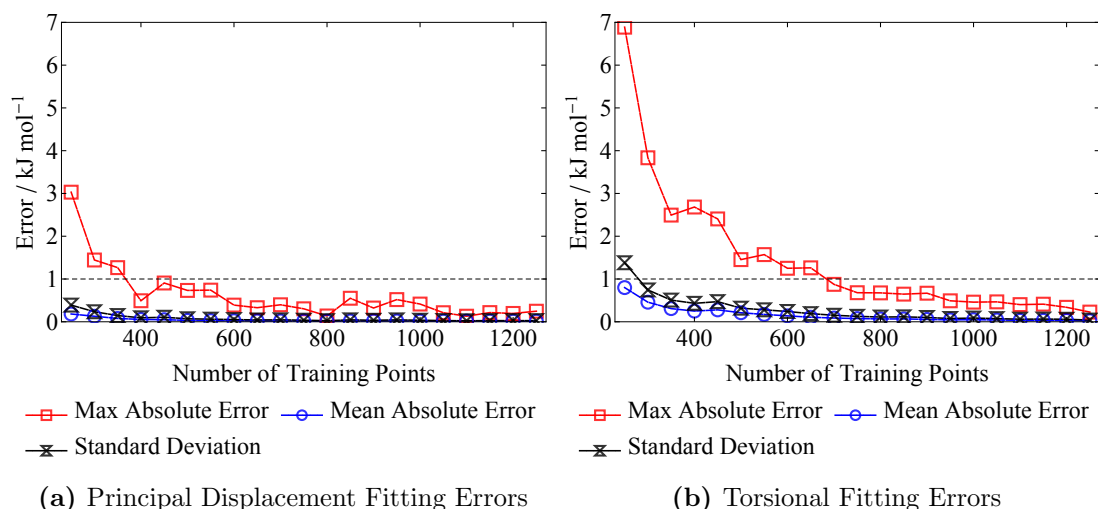


Figure 4.5: The fitting errors in the intramolecular energy when using varying numbers of training points. The dashed line is intended to easily identify to the reader when the maximum absolute error reduces below 1 kJ mol^{-1} .

4.4.2 Potential Energy Surface Fitting

The most theoretically intensive process in this chapter is the fitting of the PES. This procedure will now be explained in detail starting with the calculation of the data that is required for this calculation.

Single point energy calculations were performed on each displaced conformation within the conformational space, defined in Section 4.4.1, using the DFT-D (B3LYP-GD3BJ/6-311G**) level of theory. From these calculations, the intramolecular energy and point charge data are collected and can now be utilised to fit a PES.

A set of points generated from a Sobol sequence was partitioned into a training set (of varying numbers of points) and a 250 point test set where the latter consists of the first 250 points generated by the Sobol sequence that yield energies within 22.5 kJ mol^{-1} of the energy of the gas phase conformer. The former is used to fit the PES and the latter is used to measure the error in the fit between the exact point and the fitted point.

By increasing the size of the training set, the convergence of the error in the intramolecular energy can be observed. Figure 4.5 compares the convergence of the mean, absolute maximum errors and standard deviation errors in the intramolecular energy when fitting the PES to the 3 principal displacements and the 3 torsion angles both mentioned above.

More specifically, Figure 4.5a shows a rapid convergence with only 400 training points and the maximum absolute error also converges to below 1 kJ mol^{-1} within approximately 400 training points. This is in stark contrast to when a PES is fitted to the intramolecular energy derived from the displacement about the torsion angles. The

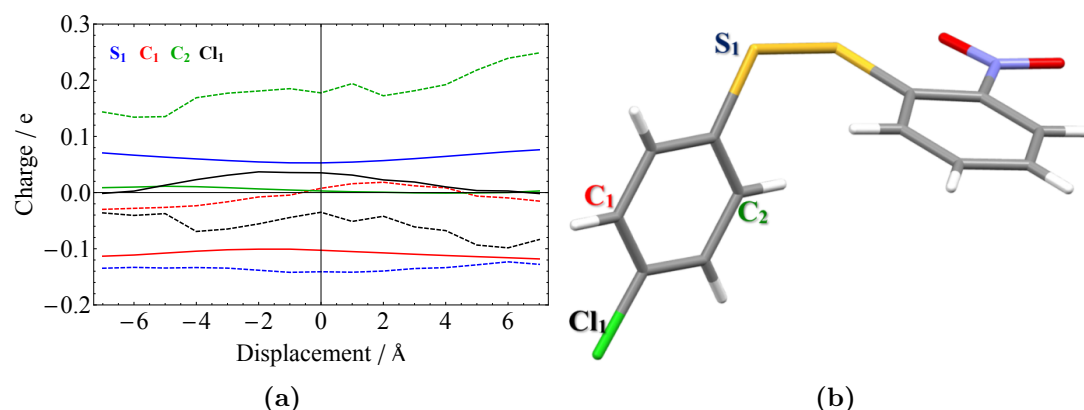


Figure 4.6: (a) shows the variation in calculated atomic charges on 4 selected atoms, (b), as the molecular geometry is displaced along the first principal displacement. The MULFIT and CHELPG charges are presented as solid and dashed lines, respectively, where the latter indicates greater stability and smoother changes with the molecular geometry.

mean and maximum absolute error reduced below 1 kJ mol^{-1} at 700 and 750 training points respectively.

Figure 4.5b also shows the uncertainty of using a low number of training points for a torsional fit as the initial maximum absolute errors are approximately 7 kJ mol^{-1} . In addition to the intramolecular energy, the atomic points charges were also fitted to a model. Atomic charges were fitted to an electrostatic potential using the MULFIT program [146, 147] up to and including the octupole moment in a set of multipoles derived from the GDMA program.

The MULFIT charges were shown to give a smoother and less sensitive response to changes in the molecular geometry than the CHELPG charges [148] that are calculated using GAUSSIAN09, Figure 4.6a. This allows a smoother model to be yielded and therefore reduces the error in the fitting procedure as there are less erratic deviations in the point charges.

4.4.3 Sampling the Conformational Spaces

For the principal displacement space, the sampling of the n -dimensional hypersphere, Figure 4.4a, was conducted using the approach of Stafford [149] to generate a distribution of points, $D(\vec{x})$, as a function of a random number, x :

$$D(\vec{x}) = f(\vec{x}) \cdot s_2^{-\frac{1}{2}} \cdot \Gamma\left(\frac{n}{2} \cdot \frac{s_2}{2}\right)^{\frac{1}{n}} \quad (4.1)$$

where $f(\vec{x})$ is an inverse cumulative Gaussian distribution that generates points along each axis, $s_2^{-\frac{1}{2}}$ is the Euclidean norm of $f(\vec{x})$. The product of these 2 terms will generate

a uniform distribution of points on the surface of a hypersphere with a unit radius [150]. The Gamma function distributes the points throughout the n -dimensional hypersphere of radius 1. In this instance, $n = 3$.

Once these points have been generated, a model is used to map the points onto the relevant conformational space, Figure 4.4, within the intramolecular energy cutoff. This is performed by introducing a scaling factor, α , that ensures the molecular distortion lies within the conformational space and the energy cutoff:

$$E(\alpha D(\vec{x})) = E^{\text{cutoff}}. \quad (4.2)$$

This method is also computationally inexpensive as the use of Gaussian processing, Section 2.4.5.1, only requires the first order derivatives of the energy with respect to the molecular distortions. Therefore Equation 4.2 can be efficiently solved using the BFGS algorithm and guarantees a distortion that will be bounded by E^{cutoff} . This sampling method can be generalised to any values of E^{cutoff} and number of dimensions. However it is observed that it is sensitive to the shape of the conformational space.

Whilst the sampling of the principal displacement space was conducted as was just explained the sampling of the torsional space had to be modified. Since the torsional space is periodic for 22.5 kJ mol^{-1} , *Torsion 1* can be linearly sampled and the other two torsions form a cross-section in the shape of a disc that is sampled using a simple 2D hypersphere (a disk) by implementing the methods from Equations 4.1 and 4.2.

4.5 Generalised Methodology

A generalised overview of the method is now presented and numbered sequentially as follows:

1. Perform a geometry optimisation of the in-crystal molecular conformation to obtain an equilibrium geometry.
2. Perform a principal displacement calculation on the equilibrium conformer.
3. Displace the molecule along each principal displacement individually and extract the quantity of displacement required to yield an energy that lies at a given amount above the equilibrium geometry.
4. Generate a set of Sobol points between 0 and 1 and use each point to calculate a proportion of displacement between the bounds yielded from step 3.

5. Create a list of all possible combinations of the displacements from the Sobol numbers generated in step 4 and displace the molecule along its principal displacements by those quantities.
6. Calculate the energy of each of these displaced conformations and extract the point charges from the corresponding molecular conformations.
7. Fit a PES to a proportion of the energy data and another to a proportion of the point charge data.
8. Perform structure generation of each molecular conformation about the potential energy well where the equilibrium conformer resides.
9. Perform flexible-molecule energy minimisation of each crystal structure.
10. Perform clustering of the crystal structures yielded from step 9.
11. Obtain a final list of crystal structures and structurally compare these to the observed structure(s).

The 2 novel aspects associated with this methodology are firstly, the fitting of the intramolecular energy and the atomic point charges using a Gaussian processing model and secondly, the sampling of the intramolecular DOFs during the structure generation phase.

The PES is fitted to the intramolecular energy of the molecule as it is displaced along combinations of the chosen principal displacements. The intramolecular energy is evaluated over a coarse, quasi-randomly sampled grid of points. This allows interpolation between the data points and hence any intramolecular energy value for a given set of displacements within the fitted space can be extracted. This extraction occurs without the need to perform a single point energy calculation which vastly reduces the computational cost of this method.

A surface is also fitted to the atomic point charges. This is also to increase the efficiency of the lattice energy minimisation procedure performed by DMACRYS. For this procedure, the point charges are used in place of the multipole expansion as this, firstly, increases the speed of the optimisation and also prevents the storage of the multipoles in a local axis frame. Whilst the multipoles yield more accurate results than atomic point charges, it is desirable but not necessary to include them in this research as the focus is the proof-of-concept of this method and would reduce its computational efficiency.

It is important to note that a new PES must be fitted for each stable conformer that exists for a given molecule as this algorithm is designed to only accurately model the

Flexibility Model	Space Group	Valid Crystal Structures
Torsions	$P\bar{1}$	6565
Principal Displacements	$P\bar{1}$	6304
Torsions	$P2_1/c$	6556
Principal Displacements	$P2_1/c$	6539

Table 4.2: Number of valid crystal structures that were successfully optimised from a set of 10,000 trial crystal structures for the FUQLIM molecule.

potential energy well on the PES that that particular conformer resides in. The PES is an important feature of this methodology and so the molecular conformational space must be well sampled to create a set of points that ensures all regions of the PES are accurately fitted.

4.6 Crystal Structure Prediction

From the gas phase conformer of the FUQLIM molecule, the torsional and principal displacement conformational spaces were searched such that 10,000 trial crystal structures were generated for each of the 2 relevant space groups where polymorphs A and B/C occur ($P\bar{1}$ and $P2_1/c$, respectively) using the Global Lattice Energy Explorer [94].

The total lattice energy of each crystal structure was minimised with respect to the displacements along the molecular DOFs in either torsional or principal displacement space. Thereby the molecule was only permitted to optimise about either the 3 torsion or 3 principal displacement DOFs.

The intramolecular energy and its corresponding point charges at each step of the minimisation was extracted from the fitted PESs and fed into DMACRYS for a rigid molecule lattice energy minimisation with respect to the crystal packing variables (lattice vectors, molecular positions and molecular orientations). An approximate success rate of 65% was observed for these crystal structures whose results are summarised in Table 4.2.

Approximately 45% of the crystal structures that were produced failed to converge to a energetically stable minimum. It was found that the in-house software that performed the molecular distortions along the set of principal displacements would commonly produce molecular geometries that were not chemically sensible, for example, overly elongated bond lengths. This causes DMACRYs to become unstable during the crystal structure energy minimisation that leads to either a lack of lattice energy convergence or yield a failed minimisation.

Polymorph	Torsion		Principal Displacements					
	(0,3)	(3,3)	(1,1)	(2,2)	(0,3)	(1,3)	(2,3)	(3,3)
A	ΔE	0.0	0.0	0.0	0.0	0.0	0.0	0.0
	$RMSD_{30}$	0.383	0.383	0.339	0.297	0.365	0.377	0.402
	Hits	210	324	513	400	343	446	405
B	ΔE					+8.9	+11.1	+8.1
	$RMSD_{30}$	-	0.495	-	-	-	0.301	0.545
	Hits		7				2	1
C	ΔE							+10.8
	$RMSD_{30}$	-	0.415	-	-	-	-	0.367
	Hits		10					3

Table 4.3: A summary of the success of predictions for polymorphs A, B and C using the principal displacements and torsion flexibility. Results are broken down using different settings whereby (S,O) combinations represent the molecular flexibility dimensions in the search (S) and optimisation (O) procedures. Upon a successful prediction, the total lattice energy above the global minimum, ΔE , (kJ mol^{-1}), $RMSD_{30}$ (\AA), and the number of structure hits, ‘Hits’, that were found during the search are listed. A hyphen represents 0 Hits and hence an unsuccessful attempt to predict the polymorph.

4.6.1 How Much is Molecular Flexibility Needed?

Before presenting the results of the methodologies, it will be interesting to know if indeed all 3 torsions or all 3 principal displacement DOFs were required to find all 3 of the known polymorphs of FUQLIM. Therefore the structure generation and energy minimisation phases were repeated but allowing the molecule to possess different degrees of flexibility throughout the search (S) and optimisation (O) procedures.

The nomenclature for the sections that follow will consist of the form (S,O) where S and O are the numbers of DOFs that were used in the search and optimisation procedures, respectively. As an example, using a rigid molecule search and a 2-DOF flexible optimisation procedure is referred to as (0,2) whereas a fully flexible search and fully flexible optimisation procedure would yield (3,3).

The results from the method outlined in Section 4.6 are summarised in Table 4.3 and will be presented by comparing the results for the (2,3), (2,2), (1,3) and (1,1) against the (0,3) and (3,3) methodologies.

4.6.2 (3,3) Molecular Flexibility

The justification of the principal displacement flexible molecule search procedure requires the comparison of the (3,3) against the (0,3) methodologies. That is, where the former

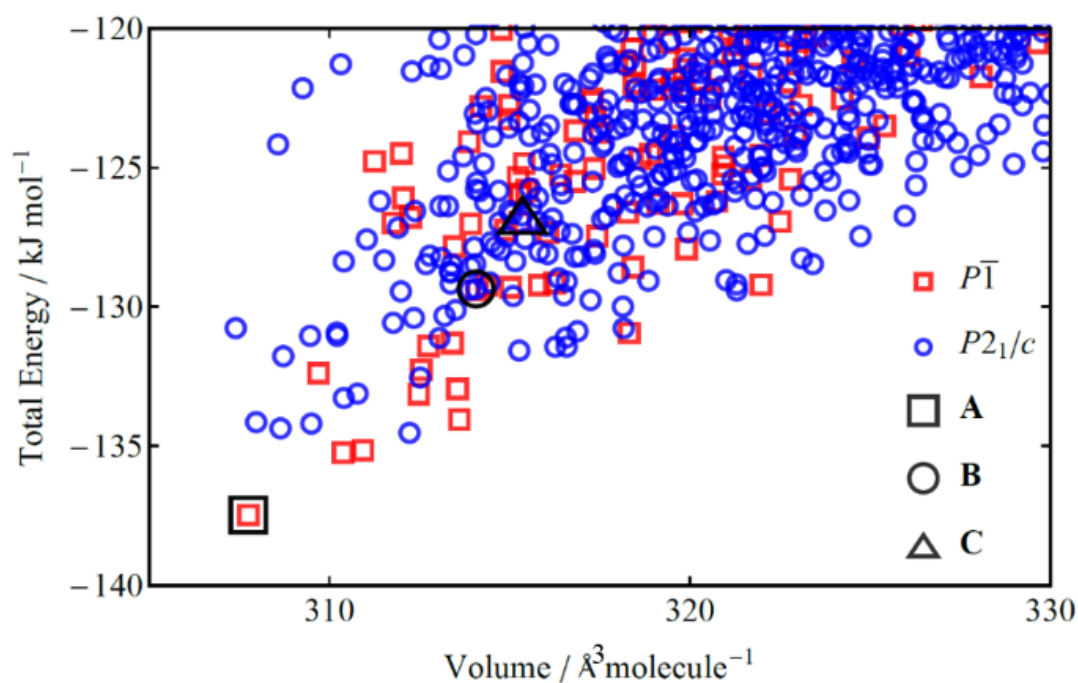
possesses a fully flexible search and lattice energy minimisation procedure and the latter possesses a rigid search but fully flexible lattice energy minimisation procedure.

Both (3,3) principal and torsion displacement methodologies correctly locate the crystal structures of all 3 A, B and C polymorphs. This was confirmed by performing structural comparisons of the calculated list of structures against the observed structures. Their positions on the energy landscape are shown in Figures 4.7a and 4.7b, for the torsion and principal displacement methodologies, respectively. The details of this can be gleaned from Table 4.3 in that polymorph A occurs as the global energy minimum and was found 444 times when implementing principal displacements and 324 times in the torsion search. Polymorphs B and C occurred at 8.1 kJ mol^{-1} and 10.8 kJ mol^{-1} above polymorph A and was found 11 and 3 times, respectively. This is in contrast to both of the (0,3) methods for principal displacements and torsions where only polymorph A is found, Figure 4.8. In addition, polymorph A is also found fewer times when using a rigid search procedure.

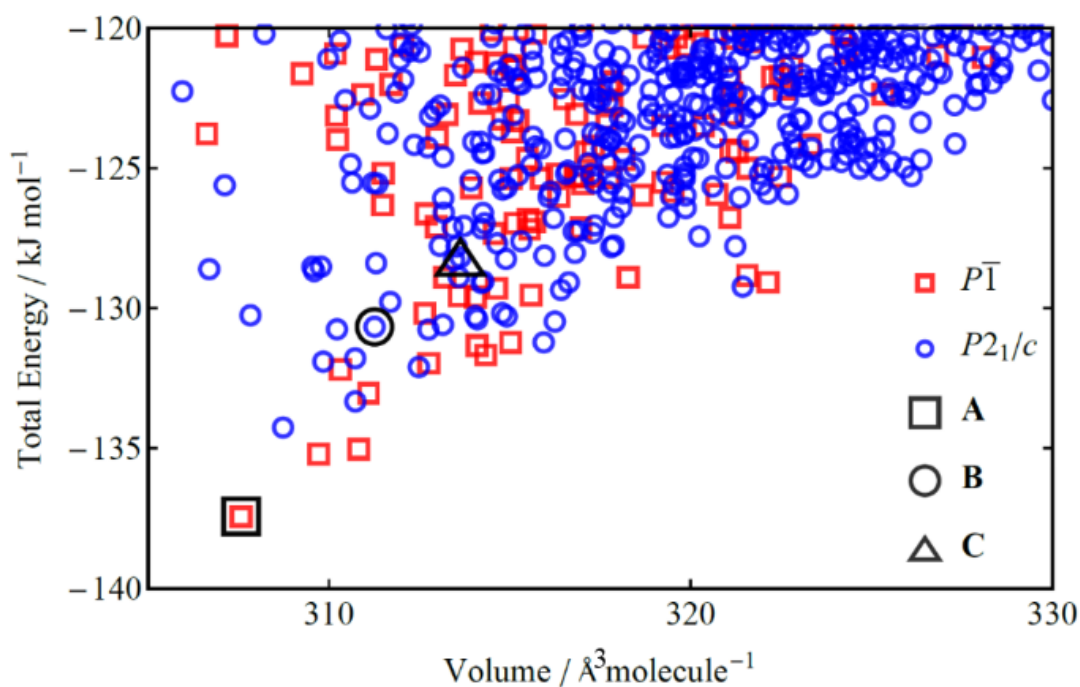
All 4 sets of results show a strong tendency to predict polymorph A as it always occurs as the global energy minimum, possesses a low RMSD value and is produced by approximately 3-7% of valid energy minimisations in the $P\bar{1}$ space group.

Polymorph B requires some degree of flexibility during the search procedure to be correctly predicted. Implementing the (3,3) principal and torsion displacement methodologies yields polymorph B at 6.8 kJ mol^{-1} and 8.1 kJ mol^{-1} above polymorph A, respectively. The energy difference between these 2 predicted crystal structures for polymorph B is 1.3 kJ mol^{-1} . This energy difference arises from the subtle geometrical differences yielded from displacing the molecule using different methodologies. The displacement of the molecule by pure torsions does not allow the molecule to flex in response to these change whereas the principal displacement methodology allows all of the other molecular DOFs to respond to changes in a single variation in a specific DOF. It is therefore perhaps unsurprising that the principal displacement methodology yields a lower ΔE value.

The RMSD values possess a larger range than those for polymorph A where the principal displacements give a substantially closer match to the observed crystal structure than the torsion displacements at 0.214 \AA and 0.485 \AA , respectively. Each crystal structure was found multiple times although significantly fewer times than the analogous results for polymorph A with principal displacements finding the observed crystal structure 4 more times than when using torsion angles.

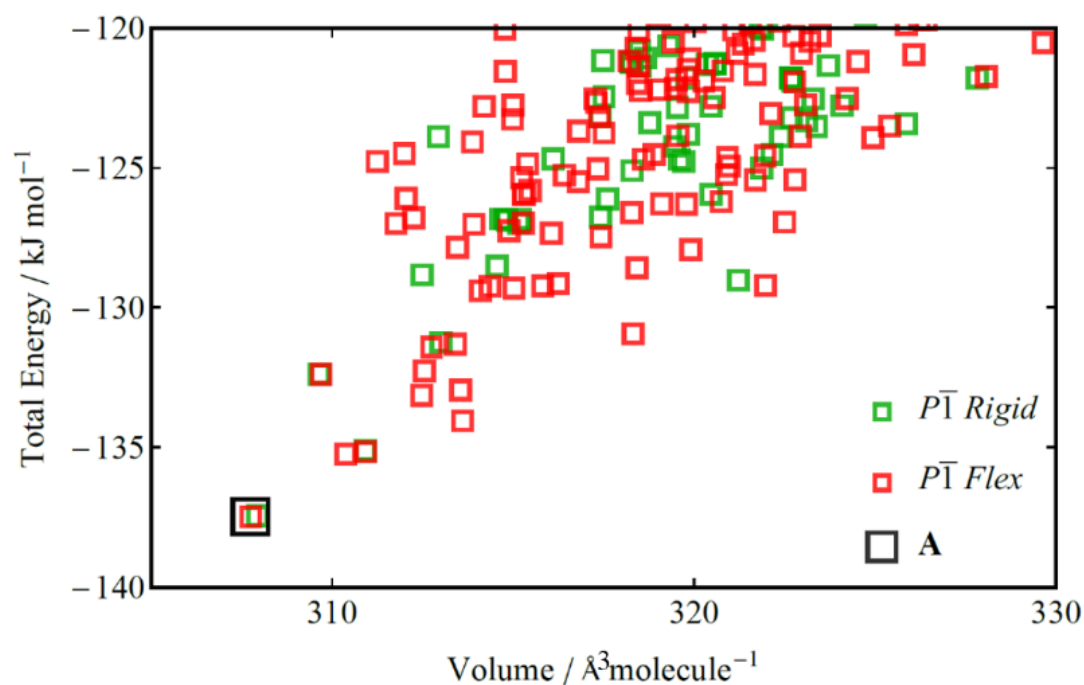


(a) CSP landscape for FUQLIM for the (3,3) principal displacement search where A, B and C show the location of the structures corresponding to the 3 known polymorphs. It is important to note the number of crystal structures between the global energy minimum structure and the polymorphs as well as the energy differences between the structures.

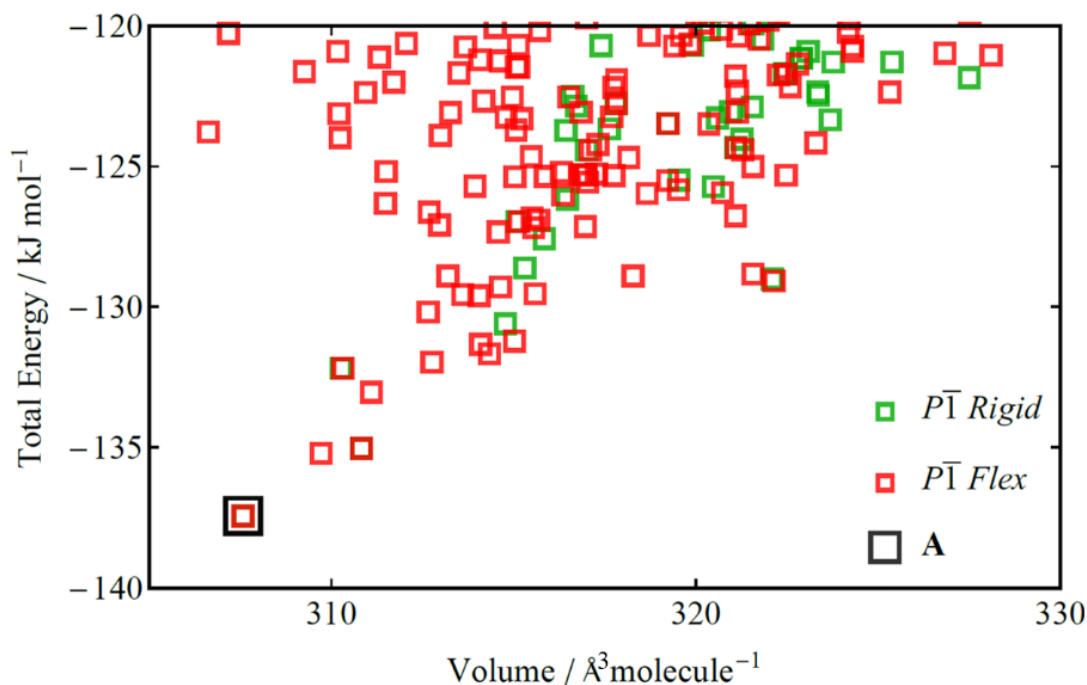


(b) CSP landscape for FUQLIM for the (3,3) torsion displacement search where A, B and C show the location of the structures corresponding to the 3 known polymorphs. It is important to note the number of crystal structures between the global energy minimum structure and the polymorphs as well as the energy differences between the structures.

Figure 4.7



(a) CSP landscape for FUQLIM for the (3,3) and (0,3) principal displacement searches where A shows the location of the structure corresponding to one of the 3 known polymorphs. It is important to note the number of crystal structures between the global energy minimum structure and the polymorphs as well as the energy differences between the structures.



(b) CSP landscape for FUQLIM for the (3,3) and (0,3) torsional displacement searches where A shows the location of the structure corresponding to one of the 3 known polymorphs. It is important to note the number of crystal structures between the global energy minimum structure and the polymorphs as well as the energy differences between the structures.

Figure 4.8

The angle of *Torsion 1* differs in the crystal structure for the principal and torsion displacements by 15.8° which accounts for the increased accuracy in the former methodology. The other two torsion angles both differ by $< 5^\circ$. Nonetheless, both methodologies provide multiple structure hits for polymorph B demonstrating that either methodology provides sufficient flexibility in the search to sample the relevant areas of conformational space.

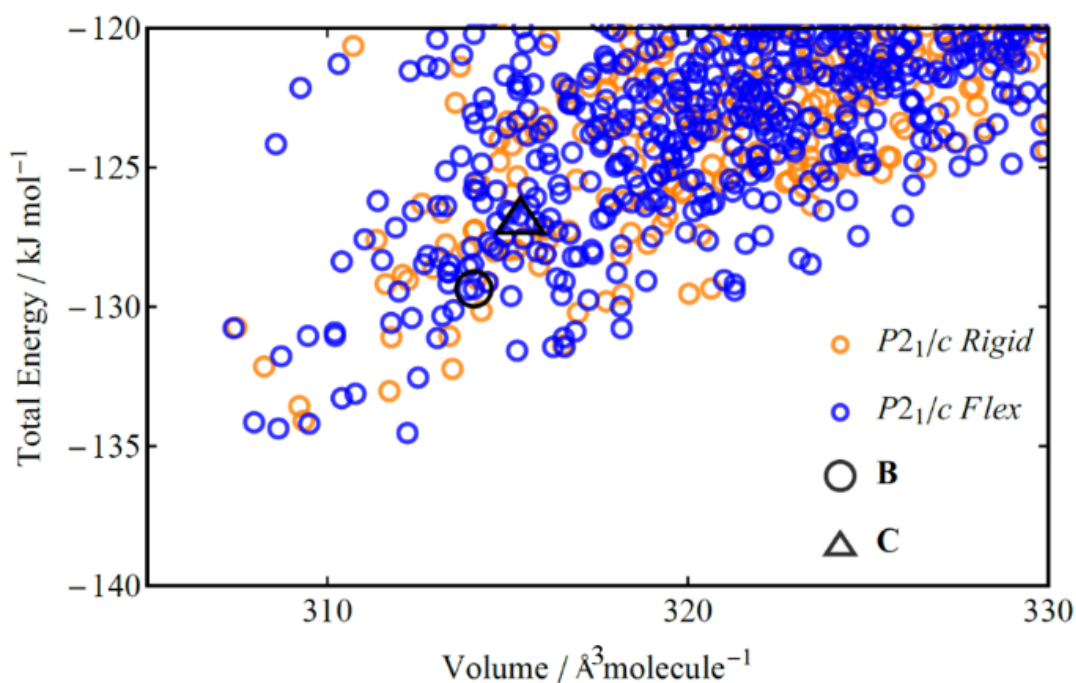
The importance of the inclusion of molecular flexibility is most prominently highlighted by the results for polymorph C. This requires all three DOFs during both the structure generation and lattice energy minimisation and is found using both principal displacement and torsion angle methodologies. Polymorph C occurs at 9.3 kJ mol^{-1} and 10.8 kJ mol^{-1} above polymorph A for the torsion and principal displacements respectively. In this instance, the torsional method leads to 3 more hits than the principal displacements but again gives a less accurate crystal structure with an RMSD 0.048 \AA greater than the latter.

The CSP landscape shown in Figure 4.9a and Figure 4.9b is an enlargement of the region where polymorphs B and C occur for the principal displacement method. The analogous CSP landscapes for the torsional displacements are shown in Figure 4.10. It was found that polymorphs B and C are present for the (3,3) methodology but not the (0,3). This is in contrast to the earlier Figure 4.8 where it can be observed that both methodologies find polymorph A.

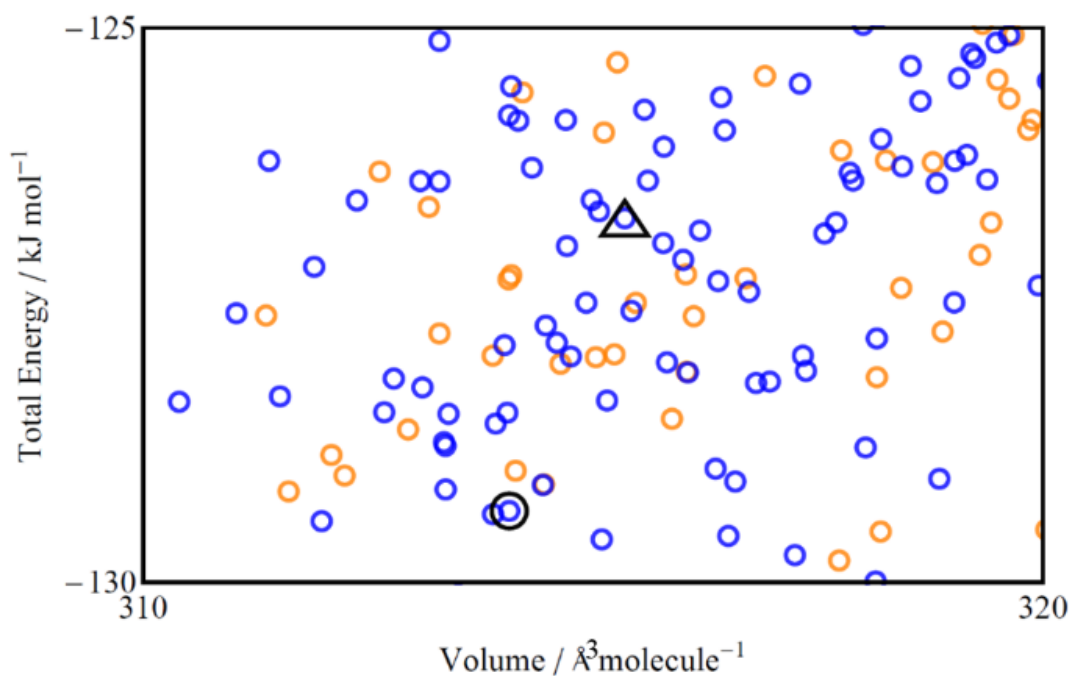
Figures 4.7a and 4.7b show the contrast in CSP landscapes when using different methodologies. Each data point is a unique crystal structure where more unique structures exist in the $P2_1/c$ space group.

These differences can be attributed to several major factors:

1. The DOFs of the molecule only allow the molecules to distort in certain ways. Allowing molecular flexibility during the structure generations phase of the CSP process allows additional minima to be found.
2. Each crystal structure is based on the lowest energy structure in a clustered set of structures. Therefore multiple structures minimise into the same potential energy well but convergence criteria and numerical noise must be minded during the minimisation procedure. This can cause minima to ‘exist’ by these criteria that would be structurally identical to other structures if the PES was more accurately described.
3. Although both (3,3) search procedures find all 3 polymorphs in this instance, there is always an assumption that the PES is not completely sampled. This can

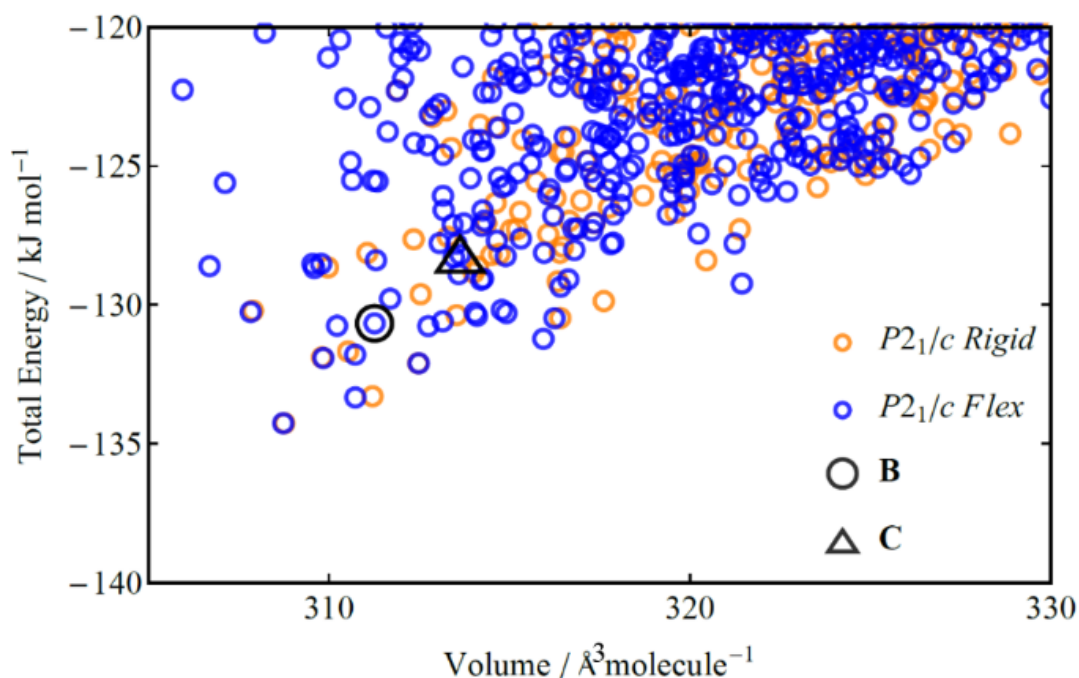


(a) CSP landscape for FUQLIM for the (3,3) and (0,3) principal displacement searches where B and C show the location of the structures corresponding to two of the 3 known polymorphs. It is important to note the number of crystal structures between the global energy minimum structure and the polymorphs as well as the energy differences between the structures.

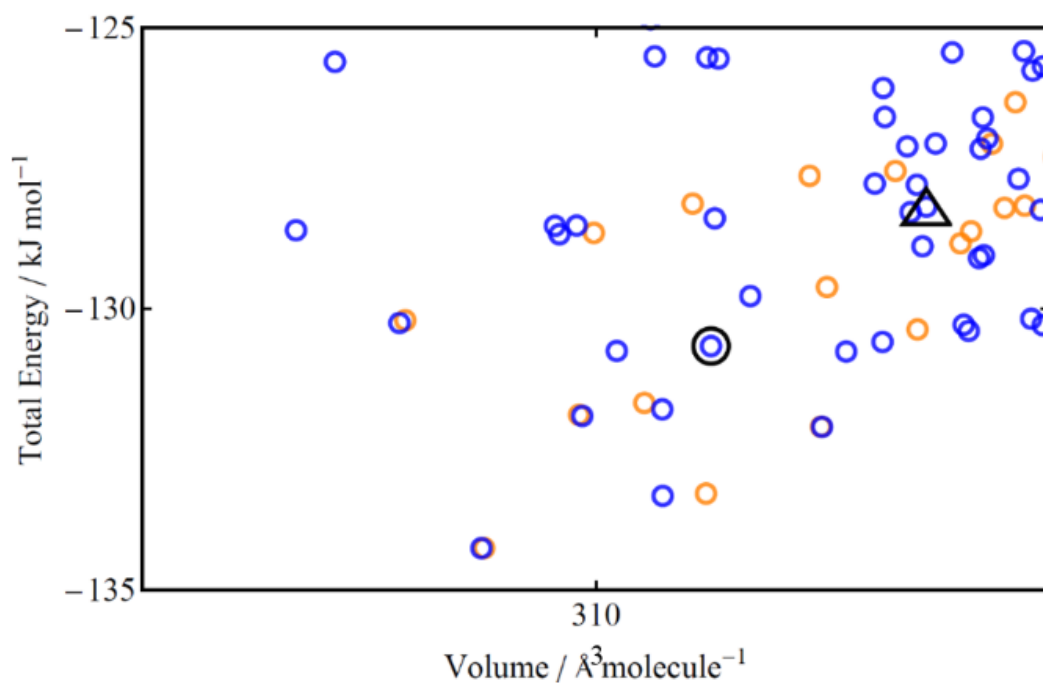


(b) A zoom of the CSP landscape for FUQLIM the (3,3) and (0,3) principal displacement searches where B and C show the location of the structures corresponding to 2 of the 3 known polymorphs. It is important to note the number of crystal structures between the global energy minimum structure and the polymorphs as well as the energy differences between the structures.

Figure 4.9



(a) CSP landscape for FUQLIM for the (3,3) and (0,3) torsional displacement searches where B and C show the location of the structures corresponding to 2 of the 3 known polymorphs. It is important to note the number of crystal structures between the global energy minimum structure and the polymorphs as well as the energy differences between the structures.



(b) A zoom of the CSP landscape for FUQLIM for the (3,3) and (0,3) torsional displacement searches where B and C show the location of the structures corresponding to 2 of the 3 known polymorphs. It is important to note the number of crystal structures between the global energy minimum structure and the polymorphs as well as the energy differences between the structures.

Figure 4.10

Polymorph	Crystal Structure	Lattice Properties						
		<i>a</i> (Å)	<i>b</i> (Å)	<i>c</i> (Å)	$\alpha(^{\circ})$	$\beta(^{\circ})$	$\gamma(^{\circ})$	<i>V</i> (Å ³)
A	Torsion	6.99	7.91	11.50	83.31	77.44	86.12	615.12
	Observed	7.04	7.83	11.37	82.58	80.69	83.19	611.43
	Prin. Disp	6.98	7.90	11.51	83.46	77.69	86.44	616.98
B	Torsion	6.98	13.58	14.43	90.00	112.68	90.00	1261.34
	Observed	7.10	13.46	13.84	90.00	109.73	90.00	1245.53
	Prin. Disp	6.99	13.71	14.04	90.00	110.99	90.00	1256.77
C	Torsion	7.08	10.92	16.33	90.00	96.31	90.00	1255.24
	Observed	7.15	11.17	15.93	90.00	94.69	90.00	1268.01
	Prin. Disp	7.07	11.43	15.73	90.00	96.81	90.00	1261.98

Table 4.4: Comparison of the properties of the observed versus the predicted crystal structures yielded from the torsion angles and principal displacement methodologies for the 3 known polymorphs of FUQLIM.

usually be quantified by observing how many hits each crystal structure possesses. For the (3,3) principal and torsion displacements, the low lattice energy regions ($< -130.0 \text{ kJ mol}^{-1}$) of the PES were well sampled where polymorph A was hit 324 and 444 times respectively. This is not as clear for polymorphs B and C as each polymorph was hit less times which could suggest either a lack of sampling in these regions, a region of the PES possessing many minima or the crystal structures residing in narrow minima with steep walls.

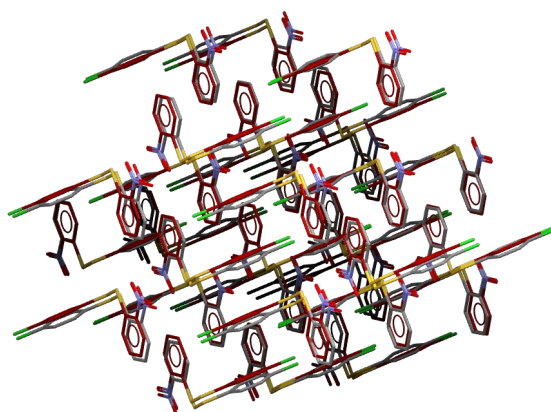
4.6.2.1 Structural Comparisons

Table 4.4 shows the comparison of properties of the observed crystal structure against those from the torsion angle and principal displacement methodologies. Figure 4.11 and 4.12 also shows overlays of the crystal structures yielded from the torsion and principal displacement methodologies with their corresponding observed structures.

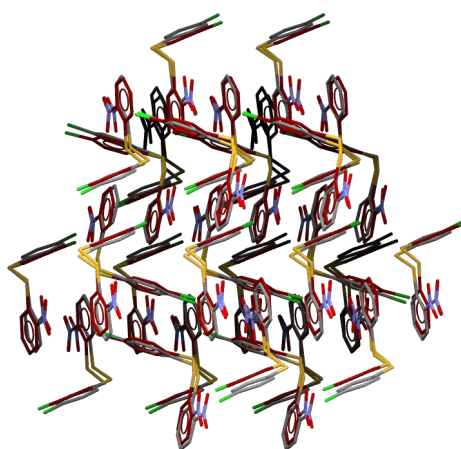
Both methodologies yielded crystal structures that give an RMSD of $< 0.5 \text{ Å}$ relative to their observed counterparts. Therefore all 3 polymorphs are found in both of these procedures. The extent of the accuracy of each methodology is also comparable.

Both methods give RMSD values of approximately 0.384 Å , Figures 4.11a and 4.12a for torsion angles and principal displacements, respectively

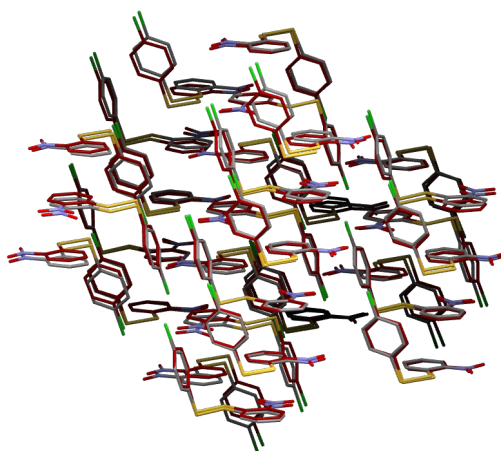
Polymorph B is generally well reproduced from both methods which, again, is reflected in the RMSD values where torsion angles gives an RMSD value of 0.495 Å which is 0.281 Å greater than when using principal displacements, Figures 4.11b and 4.12b respectively.



(a) RMSD = 0.383 Å

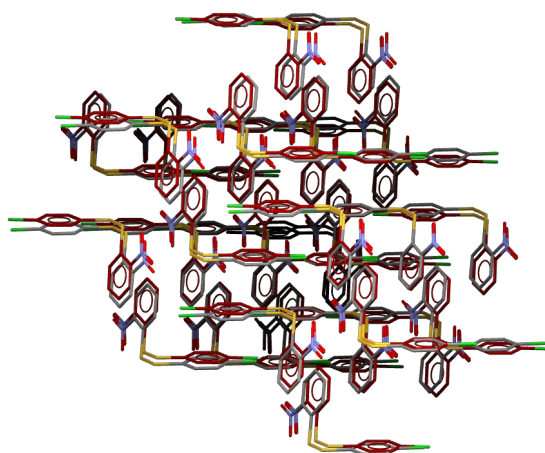


(b) RMSD = 0.495 Å

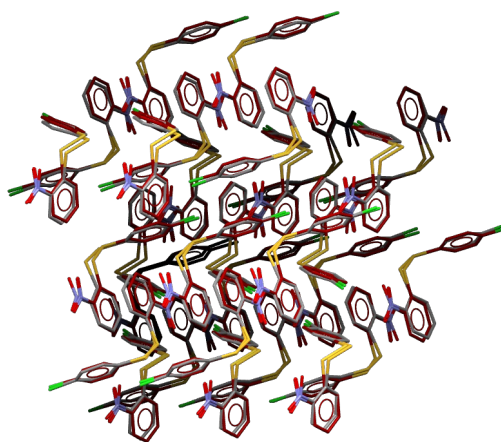


(c) RMSD = 0.415 Å

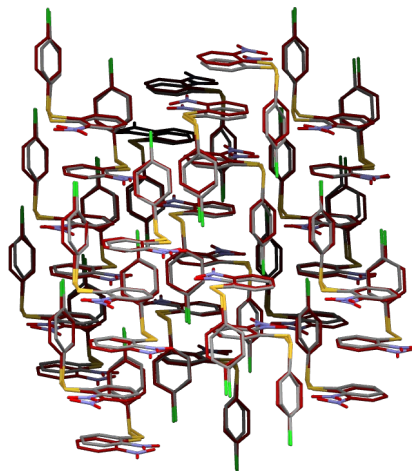
Figure 4.11: Crystal structures yielded using the torsion angle method (carbons in red) for FUQLIM polymorphs A (a), B (b) and C (c). Hydrogen atoms have been removed for clarity.



(a) RMSD = 0.384 Å



(b) RMSD = 0.214 Å



(c) RMSD = 0.367 Å

Figure 4.12: Crystal structures yielded using the principal displacement methodologies (carbons in red) for FUQLIM polymorphs A (a), B (b) and C (c). Hydrogen atoms have been removed for clarity.

The methods are also consistent when predicting polymorph C and is reflected in the RMSD values between the two methodologies as the principal displacement method gives only a 0.05 Å lower RMSD than the torsion angle method, Figures 4.12c and 4.11c respectively.

4.6.3 1-DOF and 2-DOF Flexibility

Principal displacements offer a logical and systematic procedure for increasing the number of DOFs that can be included in the search procedure. This is in contrast to internal coordinates where the user must select the relevant DOFs based on chemical intuition (unless all DOFs are chosen which can prove an unnecessary computational expense). The analysis of this point was performed by altering the number of DOFs in the search and optimisation procedures.

Table 4.3 shows that all 3 DOFs for either methodology are required to correctly predict all 3 polymorphs. Simply using a (1,1) or (2,2) approach proved insufficient in locating polymorphs B and C on the PES. Therefore, it is the inclusion of the third principal displacement that is required at the optimisation phase that allows polymorph C to be found.

Therefore it can be implied that this principal displacement performs a crucial movement to afford the observed crystal structure. This is in contrast to what would have been predicted from the visualisation of the principal displacements in Figure 4.3 where the first principal displacement would be expected to perform the bulk of the movement to convert from the gas phase into the in-crystal geometry. To further this analysis, performing the (1,3) and (2,3) combinations allows the correct prediction of polymorphs A and B. However, B is only found once in the (2,3) case and twice in the (1,3) case whereas the (3,3) case yields 11 hits.

Perhaps a logical assumption would be that the (2,3) case should yield more hits than the (1,3) case as more DOFs are included. However, as was discussed in Section 4.6.2, the PES changes shape with the inclusion of every new DOF. The accuracy of the shape of the PES will increase as more DOFs are added but that does not guarantee that more hits will be encountered. In addition, the difference in these small numbers of hits are not statistically significant as they have most likely been afforded by chance and not as a direct result of the method.

4.7 Evaluation of the Sampling Procedure

Figure 4.13 displays the collection of graphs that show where in the Sobol sequence the first crystal structure occurred in the search procedure that minimised to the polymorphs for both torsion and principal displacement methodologies.

The comparison of the sampling of the $P\bar{1}$ space group between the two methods are visualised in Figures 4.13a and 4.13b. Both methods give good convergence with the majority of low energy crystal structures occurring within the first 2,000 valid trial structures.

Furthermore, the global, minimum energy crystal structure was found for both methodologies within the first 200 crystal structures that were validly minimised. This number gives a positive result for CSP and shows that the low energy regions of the $P\bar{1}$ space group are well sampled within only a small number of valid structures. This also shows that the PES defined for this molecular system within this space group possesses wide potential energy wells due to the relatively low number of unique crystal structures that are found. Although new crystal structures are found at higher numbers of valid crystal structures, these all occur in the higher energy regions of the PES and are energetically far from the global energy minimum.

The lack of unique structures also shows that the PES is smooth such that the newly generated crystal structures are not easily trapped in deep potential energy wells with steep walls. Therefore, upon energy minimisation, these crystal structures possess clear and well defined paths to their respective local minima.

Figures 4.13c and 4.13d compare the total lattice energy against the number of valid minimisations within space group $P2_1/c$. The annotated points show that rank of the structure with respect to the total lattice energy and indicates where the first structure was generated from the Sobol sequence that minimised to one of the polymorphs.

Both figures show that the global energy minimum is found within the first 3,000 valid crystal structures for both methods. Whilst this is significantly higher than the 300 structures required for the $P\bar{1}$ space group, the $P2_1/c$ space group possesses many more minima. Therefore it is perhaps unsurprising that the global energy minimum occurs later in the search.

Space group $P2_1/c$ also possesses more unique structures that is suggestive of a more volatile PES with more minima than space group $P\bar{1}$. The lack of convergence of these plots is a result that shows the search requires a higher number of structures to fully explore the PES as unique, low energy crystal structures are still being found at the end of the search. This is also suggestive that the minima on the PES possess steeper

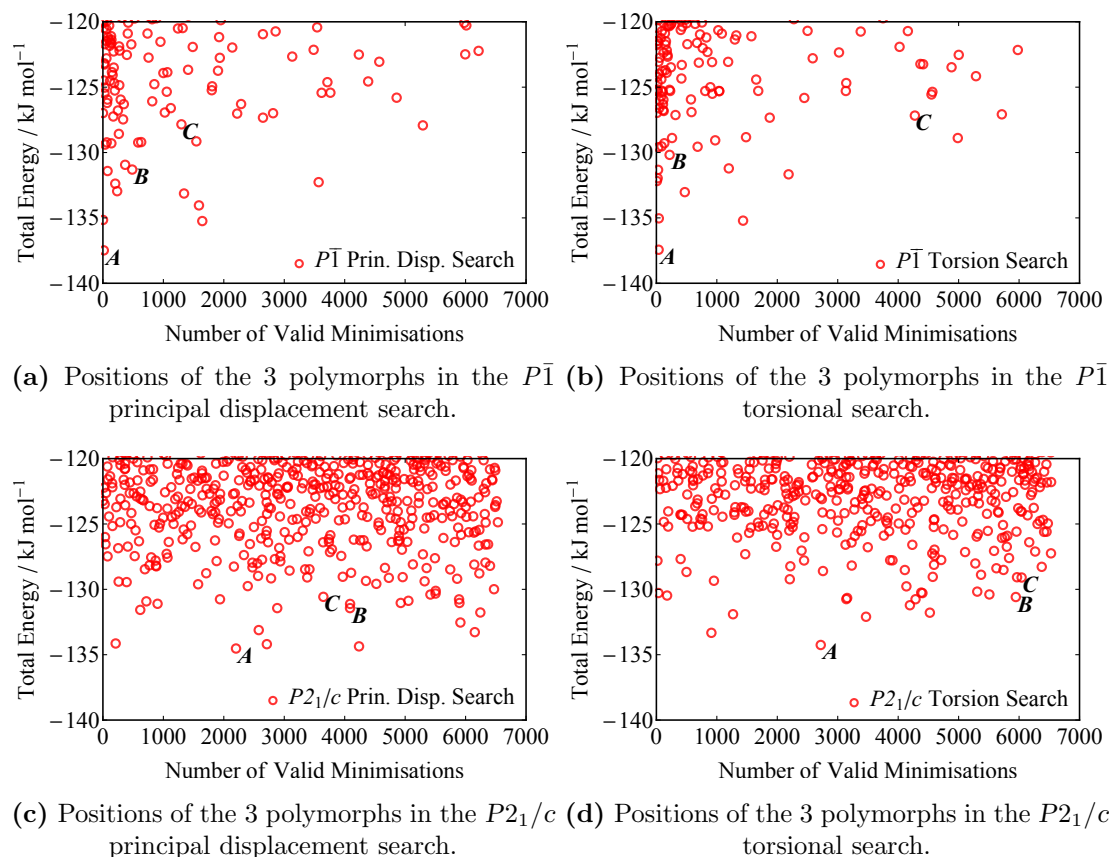


Figure 4.13: All 4 plots show the number of unique crystal structures being generated for a given total energy against the number valid minimisations.

walls and can only be found when more specific regions of the conformational space is sampled. This brings to the discussion of the weighting of the number of crystal structures to include per space group. However, this research is not focussing that issue but should remain a feature of both methodologies that should be considered.

Clearly, although space group $P2_1/c$ did not converge as unique low energy structures were still being found at the end of the search, the sampling was sufficient enough to find all 3 polymorphs which is clearly an improvement on previous methods [145].

Directly comparing the 2 methodologies, the torsional method appeared to allow both space groups to converge more quickly. This is not necessarily a positive point as, although the principal displacement and torsional PESs will differ, unique crystal structures were still being found as higher number of structures generated. Statistically, as more crystal structures are generated one would expect a higher probability of finding the observed structure(s), although this is not a guarantee.

4.8 Conclusions

The research in this chapter has demonstrated that the incorporation of molecular flexibility in the form of molecular principal displacements into CSP is not only possible, but also enhances the accuracy of the results.

The case of FUQLIM has been presented to which now all 3 polymorphs A, B and C are found in the final list of crystal structures. This is in contrast to previous research [145] where only polymorph A was found when implementing molecular flexibility solely at the lattice energy minimisation phase of the CSP process. It was found that using current methods for treating molecular flexibility, where flexibility is allowed about selected torsion angles, at the search stage would have sufficed for finding all 3 polymorphs when including the most chemically intuitive DOFs. However, the development of implementing molecular principal displacements as an alternative proves that this method is not only comparable to the current *status quo*, but can also yield more chemically accurate crystal structures.

The principal displacement, state-of-the-art methodology remains in its infancy however and still possesses a major blind spot; the decision of which principal displacements are needed to be included without possessing any prior knowledge of the number of polymorphs their respective crystal structures.

The research associated with this problem will have to be delayed by a chapter as the methodology required to conduct these calculations yielded a whole other section of novel research.

Chapter 5

Molecular Geometry Interconversion Using Principal Displacements

5.1 Introduction

The conclusions of Chapter 4 highlight a level of ambiguity that remains as to how to define the principal displacements required to define a search space without prior knowledge of the target crystal structure(s). The utilisation of the set of FUQLIM crystal structures demonstrated the proof-of-concept of the principal displacement search procedure and provided a promising outlook for the methodology.

However, the methodology still relies on the user knowing the in-crystal molecular conformation (*conformations* if the system exhibits conformational polymorphism) such that the conformational search space will be large enough to encompass both the gas phase and in-crystal geometries. Ironically however, if all of the geometries are already known, it renders the technique redundant and any CSP calculations will not be needed as the target crystal structure(s) are already known. Therefore more information must first be provided on how the search space will be defined by observing which principal displacements contribute this geometry conversion, how many principal displacements are required and how far does one traverse along each principal displacement. The answers to these questions are required to give the greatest degree of certainty that the search performed in CSP studies encompasses the in-crystal molecular geometries for all possible low energy crystal structures.

To provide an opening into the answering of these questions, a test set of gas phase molecular conformers accompanied by their in-crystal conformational counterparts will be required to commence the investigation. This will be followed by a set of calculated principal displacements that perform the geometry interconversion from that of the gas phase to the in-crystal molecular geometry. By observing which principal displacements that contribute to this geometry interconversion, a set of rules can begin to be derived for use within the methodology in CSP.

However, the research conducted to derive these rules will be delayed by a chapter as the theoretical basis required to perform the geometry interconversion requires a deeper knowledge of molecular principal displacement theory than originally thought. The results of the methods implemented in this chapter can be found in Chapter 6.

The present chapter will commence by defining this test set of molecular crystals before presenting 3 different methodologies for performing a geometry interconversion using molecular principal displacements. The merits and failings of each method will be analysed to aid in the selection of which method is most suitable for the purposes of this research.

5.2 The Test Set of Molecules

The test set of molecules that were involved in this study were a modified set of those used by Thompson & Day in past research [120]. This test set of 13 molecules is shown in Figure 5.1.

The test set is modified from that of Thompson & Day as it excludes 2 molecules, HIBGUV and HAJYUN. This is due to the presence of halogen atoms within these molecules. Halogen atoms are able to be dealt with by the version of the W99 potential that is being implemented but there is less confidence in the parameters for these atom types. Therefore allowing these molecules to continue through the processes could lead to inaccuracies in the geometries that result. Since the test set is large and diverse enough for preliminary investigations, the exclusion of these molecules will not limit the conclusions. These rejected molecules, HIBGUV and HAJYUN, are shown in Appendix B, Figure B.1.

The polymorphic extent of each molecule is highlighted in the captions in Figure 5.1 with the numbers representing the number of known polymorphs, the total number of independent molecules in the crystal structure (summed over all of the included polymorphs) and the number of unique in-crystal conformers found in all of the included polymorphs, respectively.

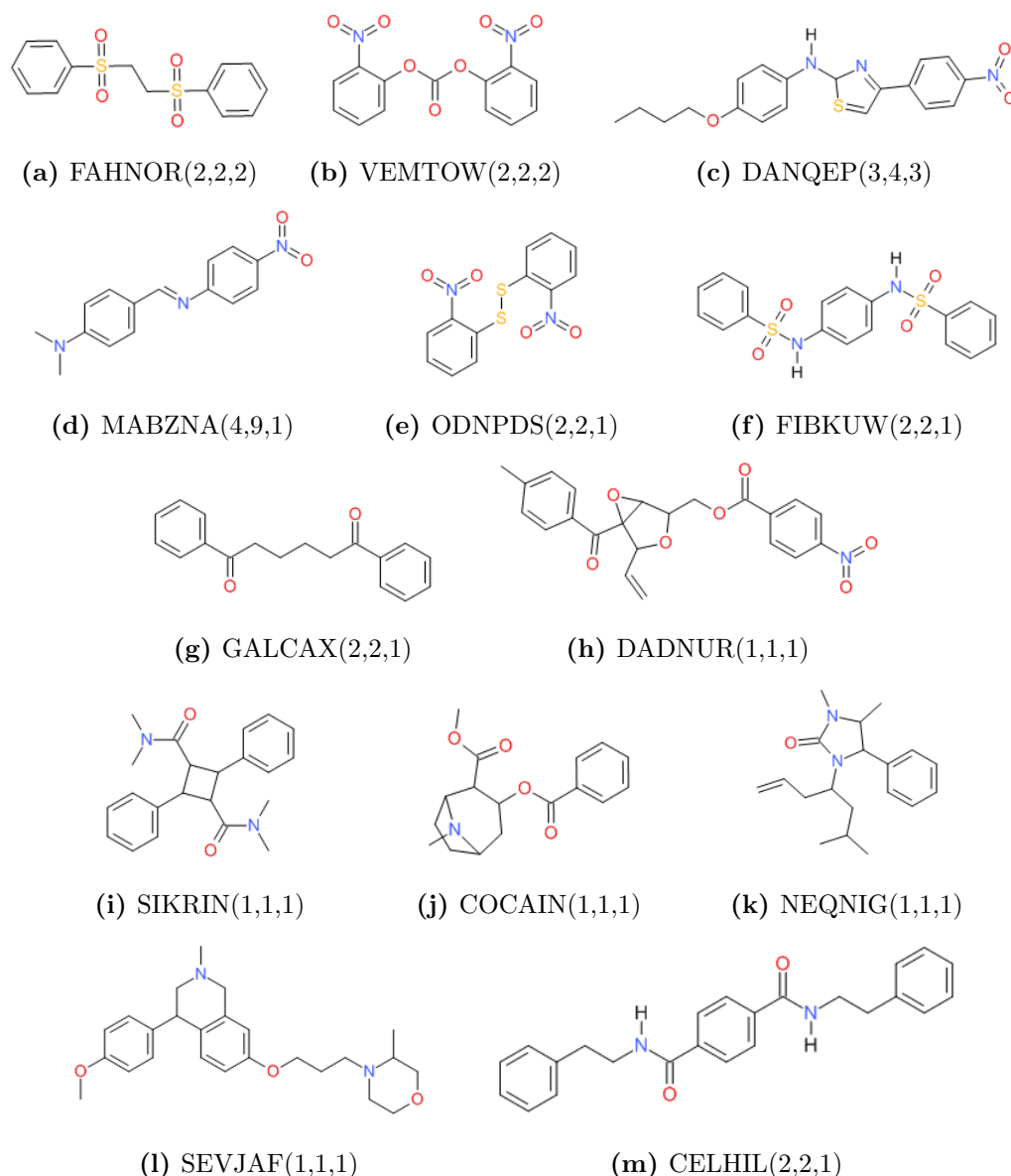


Figure 5.1: The 13 molecules included in this study. The molecules are referred to by their CSD reference codes and the 3 numbers in parentheses following each molecule refer to the number of known polymorphs, the number of independent molecular geometries (summed over all known polymorphs) and the number of unique conformers found in all known polymorphs, respectively.

Molecules FIBKUW, GALCAX, ODNPDS, CELHIL, MABZNA exhibit packing polymorphism as they possess only 1 unique conformer in both of their polymorphs. FAHNOR, VEMTOW and DANQEP exhibit conformational polymorphism as they possess more than 1 unique in-crystal molecular conformation over their polymorphs.

The remaining molecules SIKRIN, COCAIN, NEQNIG, SEVJAF and DADNUR do not exhibit any known polymorphism. The presence of polymorphism does not affect this investigation and merely adds to the number of crystal structures that are presented in the results.

Molecule	Number of Atoms	Number of Rotatable Bonds
CELHIL	28	8
DADNUR	30	8
DANQEP	26	8
GALCAX	20	7
SEVJAF	30	7
FIBKUW	26	6
NEQNIG	22	6
VEMTOW	22	6
COCAIN	22	5
FAHNOR	20	5
ODNPDS	20	5
MABZNA	20	4
SIKRIN	26	4

Table 5.1: The number of atoms and rotatable bonds of the molecules in the test set ordered by the extent of molecular flexibility.

Table 5.1 lists the number of atoms and the number of rotatable bonds present in the molecules in the test set. All of the molecules possess a similar number of atoms (20 to 30) but the shape of the molecule can vary greatly. One of the most prominent examples of this in the test set are the geometrical differences between COCAIN and DANQEP where the former is more spherical and the latter possesses a long, thin, more planar geometry.

The extent of molecular flexibility can be easily quantified by the number of rotatable bonds in each molecule, Table 5.1. A rotatable bond is defined as a ‘soft’ torsion angle that possesses the ability to be displaced when a relatively small force is applied to it. There is no robust definition of ‘soft’ and so it is assigned purely using chemical intuition. For example, amide bonds are not classed as rotatable as it is commonly found that these bonds largely remain in a *trans* configuration [151].

The number of rotatable bonds in the molecules in the test set range from 4 to 8 therefore a varied level of flexibility is exhibited by the test set. The distribution of rotatable bonds in the molecules also range from being localised in one part of the molecule to being spread about the molecule. This contrast is highlighted by comparing the structural formulas of SIKRIN and CELHIL where the rotatable bonds of the former exist about a central, four-membered ring and where they are spread throughout the molecule in the latter.

The molecules in the test set possess the ability to form intermolecular hydrogen bonds but do not possess the flexibility to form any intramolecular hydrogen bonds. More specifically, a selection of these molecules do possess the relevant atom types to be capable of hydrogen bonding, however the molecules do not possess the level of flexibility required to naturally bring these atom types into close enough proximity to each other to form any intramolecular hydrogen bonds either.

The presence of intramolecular hydrogen bonds can vastly affect the ΔE_{conf} and ΔE_{strain} values. Therefore, for the purposes of analysing ΔE_{strain} , this study will include molecules that do not possess intramolecular hydrogen bonding. The ΔE_{strain} analysis for these molecules can be found in Chapter 6. For now, this chapter will purely focus on the methods for converting between two molecular conformations.

For these molecules, the objective is now to find a set of principal displacements that can convert between these gas phase and in-crystal molecular geometries; both of which are already known. The set of required principal displacements, and the amount of distortion along each one, will inform the application of the flexible molecule CSP method introduced in Chapter 4. The next sections discuss the failures and successes of the different methodologies for performing this geometry conversion.

5.3 Pure Cartesian Approach

The simplest approach is to formulate this problem of finding a set of principal displacements to convert between two molecular geometries into a pure Cartesian solution. This methodology (and the others discussed in later sections of this chapter) could be used to convert the geometries of *any* two molecular conformations (not just the in-crystal and gas phase geometries that this work pertains to). Henceforth, the in-crystal and gas phase molecular geometries will be referred to as the ‘target’ and ‘base’ geometries, respectively.

5.3.1 Theory

The target molecular geometry, \vec{q}_t , can be thought of as a perturbation, $\Delta\vec{q}$, of the base molecular geometry, \vec{q}_b . In relation to CSP, \vec{q}_b is the geometry optimised gas phase molecular coordinates and \vec{q}_t are the atomic coordinates of a molecule in a lattice energy minimised version of the crystal structure.

Therefore this methodology, as well as being the simplest, is also the most logical starting point for the series of methodologies. Formulating this into an equation is by use of the

Cartesian coordinates of both molecular geometries:

$$\vec{q}_t = \vec{q}_b + \Delta\vec{q} \quad (5.1)$$

or, more explicitly in vector form:

$$\begin{pmatrix} x_{t1} \\ y_{t1} \\ z_{t1} \\ \vdots \\ x_{tn} \\ y_{tn} \\ z_{tn} \end{pmatrix} = \begin{pmatrix} x_{b1} \\ y_{b1} \\ z_{b1} \\ \vdots \\ x_{bn} \\ y_{bn} \\ z_{bn} \end{pmatrix} + \begin{pmatrix} \Delta x_1 \\ \Delta y_1 \\ \Delta z_1 \\ \vdots \\ \Delta x_n \\ \Delta y_n \\ \Delta z_n \end{pmatrix} \quad (5.2)$$

where, \vec{q}_t and \vec{q}_b are already known. This leaves only one unknown, $\Delta\vec{q}$. The principal displacements of the base geometry are still required to perform the base-target interconversion. This information, although not readily available, is easy to calculate from the methods described in Section 2.2.1.4, and can then be used when rearranging Equation 5.1 which is then followed by the decomposition of $\Delta\vec{q}$:

$$\vec{q}_t - \vec{q}_b = \Delta\vec{q} = \vec{s} \cdot \mathbf{C} \quad (5.3)$$

where \mathbf{C} is a $(3N - 6 \times 3N)$ matrix containing information about the principal displacements of the base geometry which are calculated using DFT. More specifically, each column of \mathbf{C} is a principal displacement of the base geometry and each row is a given atomic contribution to that particular principal displacement. \vec{s} represents a column vector where the elements represent the corresponding contribution of the principal displacements. The complete equation for this methodology can be written as:

$$\vec{q}_t = \vec{q}_b + \vec{s} \cdot \mathbf{C}. \quad (5.4)$$

From the previous discussion, it is now clear that \vec{s} is the only unknown in this equation. However, \mathbf{C} is a non-square matrix and therefore is not directly invertible. Hence, Equation 5.4 cannot be exactly rearranged to make \vec{s} the subject. It is important to note that matrix pseudo-inverse techniques do exist [152–154] but in this instance, another approach can be utilised.

Fortunately, with reference to Equation 5.3, there exists a least squares solution [155]. Equation 5.4 can be reformulated into a least squares problem that possesses the general formula, thus:

$$\vec{A} = \mathbf{B} \cdot \vec{x} \quad (5.5)$$

where \mathbf{B} is a coefficient matrix, \vec{A} contains the ordinate variable values and \vec{x} possesses the least squares solution. This equation is not exact and will provide an approximate solution for \vec{x} but it is computationally efficient and requires only one function evaluation. From a computational perspective, the least squares Equation 5.5 is solved by normalising the minimised Euclidean distance:

$$\left\| \vec{A} - \mathbf{B} \cdot \vec{x} \right\|. \quad (5.6)$$

Although this is a minimisation procedure, it rapidly converges within less than 10 iterations. Applying this methodology to Equation 5.3, a least squares formulation can be derived:

$$\Delta \vec{q} = \vec{s} \cdot \mathbf{C}. \quad (5.7)$$

Therefore a column vector \vec{s} can be obtained which will describe the contribution of each principal displacement to the geometry interconversion. It is also important to note that the computational expense of this methodology is negligible due to only simple matrix-algebraic operations being performed.

However, this methodology is short-sighted as it completely ignores the subtlety of the differences between linear and curvilinear space described in Section 2.2.1.3. Although this method is mathematically sound it is not scientifically robust. Therefore, even with the relatively small molecular distortions that these molecules are exposed to, the linear displacements for the principal displacements are not natural distortions of the molecule. It could be argued that this is a moot point, and it perhaps is for small displacements, however there are more robust methodologies available.

5.4 RMSD Minimisation

A more robust approach is to attempt to minimise the RMSD between the target and base conformations as a function of the principal displacements, which are applied in curvilinear space. An immediate issue with this method is the issue of a numerical minimisation technique in that it is, quintessentially, a numerical minimisation.

More specifically, these techniques require multiple evaluations of the same function and generally do not scale well when more dimensions are added to the minimisation; both of which add to the computational expense. Nonetheless, a numerical minimisation approach allows for the subtlety of curvilinear space and therefore, if the computational expense can be kept manageable, this method is superior to the *Pure Cartesian Approach* described in Section 5.3.

5.4.1 Theory

The algorithm for performing this *RMSD Minimisation* is simple and is shown diagrammatically in Figure 5.2.

The selection of the principal displacement parameters causes an issue with this methodology. Each molecule in the test set possesses between 78 and 168 principal displacements. Even the former is far too many principal displacements to include in the minimisation of the RMSD. Therefore the user must specify a list of principal displacements that are either up to a given force constant cutoff, or select a set of principal displacements using a user-defined set of criteria.

In addition, the start points for each principal displacement selected are also an issue. This RMSD minimisation is a local minimisation, so the starting point has a strong influence on the final result. However, since the molecular distortions for these base-target combinations are relatively small, a start point of 0.0 Å displacement for each principal displacement could be implemented (although this has no other scientific basis thus far) and the target conformation could still be found.

However, this methodology will need to be extrapolated to molecules that are potentially more flexible and possess a larger atom count, hence possessing more principal displacements. This will require larger molecular distortions to match the base to the target geometry and therefore a larger step size by the minimiser, with respect to the quantity of each principal displacement.

The type of minimiser will also affect the evaluation of this method. For this research, the Simplex algorithm [72] will be implemented as this does not require the computationally expensive step of computing the gradients of the function that is being minimised.

5.4.2 Hydrogen Peroxide: a Preliminary Example

To test that the *RMSD Minimisation* technique can be used, a trivial, and arguably the most simple, example was chosen. Hydrogen peroxide, Figure 2.3, possesses only 1 torsion angle, 2 bond angles and 3 bond length vectors hence possessing only 6 principal displacements which can be expressed as combinations of the bond length, bond angle and torsion angle distortions. This is a very manageable number to deal with for this simple case as opposed to 2 or 3 orders of magnitude more from the other molecules in the test set.

The torsion angle will possess the lowest force constant and therefore is the main contributor to the lowest energy principal displacement. In this simple example, the HOOH

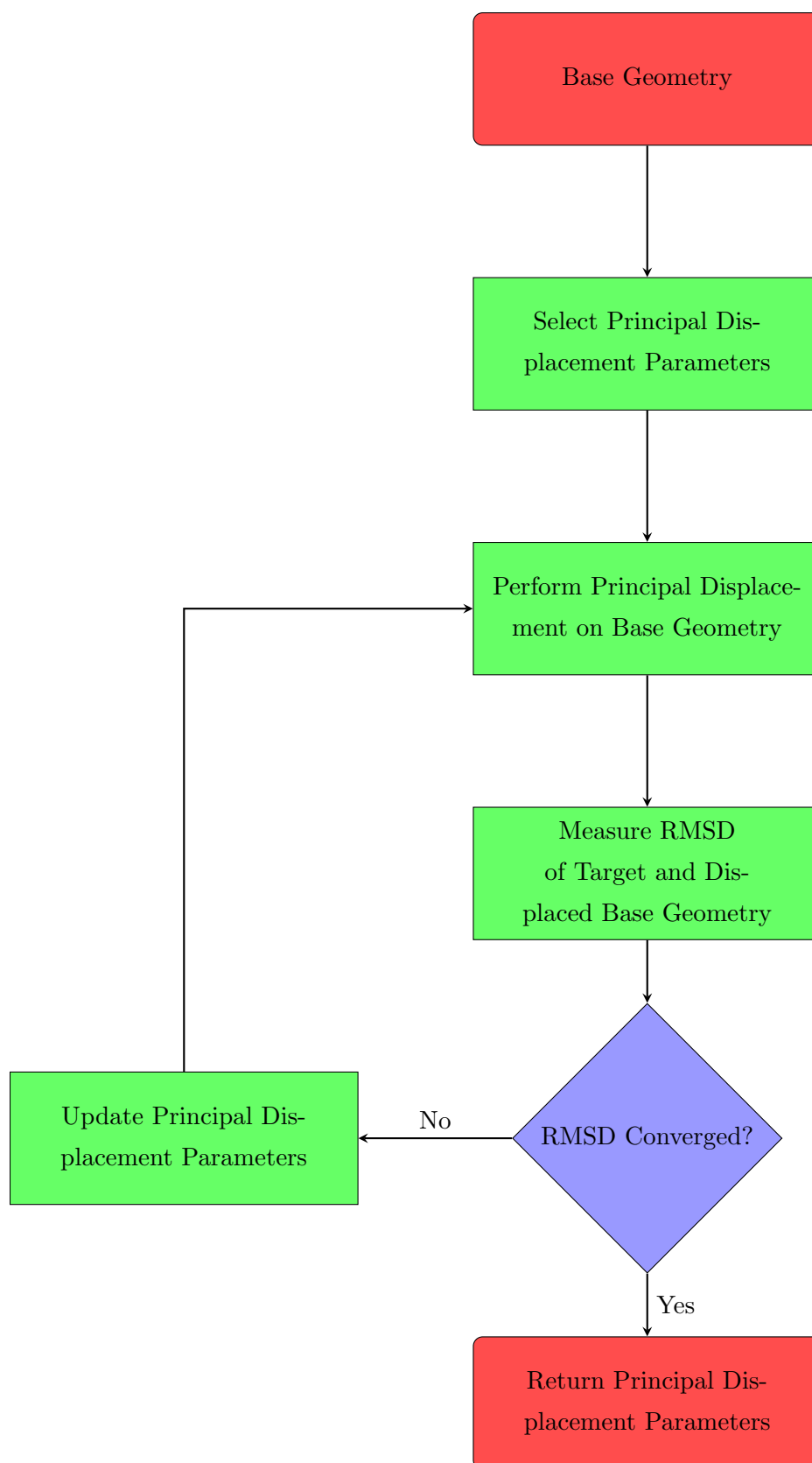
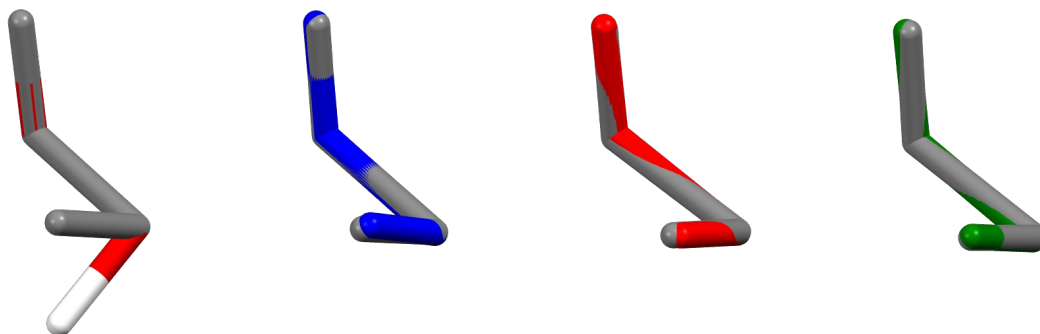
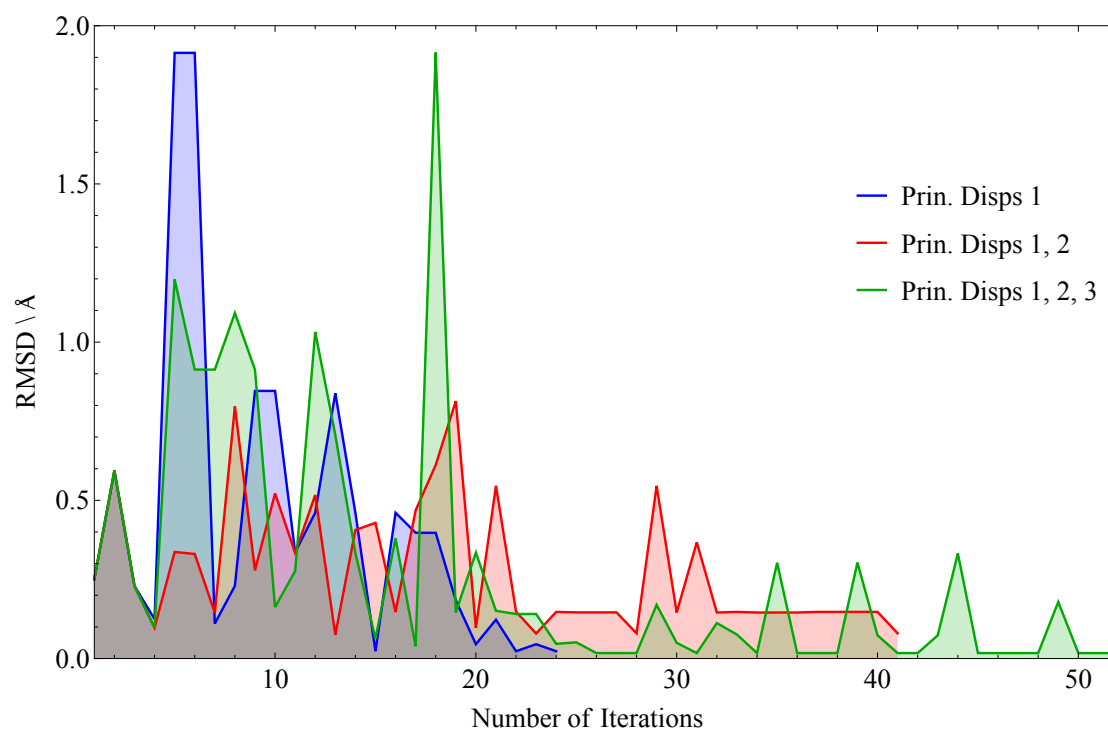


Figure 5.2: Description of RMSD minimisation procedure work flow. The RMSD convergence criteria is a user-defined value.



(a) Geometrical comparisons of the hydrogen peroxide target geometry (grey) against the base geometry (coloured by element) and the displaced geometries yielded from using the first (blue), first and second (red) and the first, second and third (green) principal displacements in the *RMSD minimisation* technique.



(b) The RMSD value at every step of the minimisation procedure when using the first (1), blue; first and second (1,2), red; and first, second and third (1,2,3), green principal displacements.

Figure 5.3

torsion angle is displaced from the base value (120.1°) by $+60.0^\circ$ to create the target geometry, Figure 5.3a, leaving all bond lengths and bond angles unmodified. The base geometry of hydrogen peroxide and its lowest energy principal displacement are then used to minimise the RMSD between the base and the target geometries.

Section 2.2.1.1 previously mentioned that torsional movements provide the largest geometrical distortions for the lowest energy cost. Therefore it is logical to assume that the principal displacements with lower force constants are probable to more aggressively reduce the RMSD.

Figure 5.3b shows the reduction in the RMSD with each iteration of the minimisation procedure when using the first (1); first and second (1,2); and first, second and third (1,2,3) principal displacements. More iterations are required as more principal displacements are added to the minimisation procedure due to the increase in dimensionality of the space where the minimisation is occurring. The convergence criteria was a reduction in RMSD of less than 10^{-6}\AA .

The variability of the traces show the sensitivity of the RMSD as the molecule is distorted which shows the history of the path to the minimised molecular geometry. Using (1,2) gives rise to a lower volatility but finds a different minimum to (1) and (1,2,3). This highlights an issue with the minimisation procedure which is that there exists many RMSD minima, the molecular geometry can minimise to a different minimum opposed to the desired global minimum. The minimisation path is not always smooth from the base to the target geometries.

Nonetheless, this procedure finds low RMSD minima and does not prove computationally expensive (at least for this system). The method is also scientifically robust and therefore provides an improvement on the *Pure Cartesian Approach*. However, by combining the idea behind this approach with the theory in Section 2.2.1.2, an even more robust solution can be obtained.

5.4.3 An Alternative to RMSD

The RMSD is a well understood quantity as it is widely recognised as the *status quo* when quantifying differences between molecular geometries. However, if the composition of the RMSD summation was broken down into its individual terms, then each term would give information about the matching of each atom in Cartesian space.

Whilst this is useful, an alternative approach would be to quantify the difference in the DOFs between two molecules, U . This requires the evaluation of the following equation:

$$U = \left| \sum_{i=1}^{3N-6} |q_{t,i} - q_{b,i}| \right| \quad (5.8)$$

where i is the DOF index and N is the number of atoms in the system. In addition, the definitions of $q_{t,i}$ and $q_{b,i}$ change to represent a column vector that now lists the value of

each DOF in the molecule, in terms of bond lengths, bond angles and torsion angles for the target and base geometries respectively. This takes the sum of the absolute difference of each corresponding DOF between the two systems. By decomposing the summation here and observing each term, this shows the source of the differences in molecular geometry; not merely the product of it. For the example of hydrogen peroxide, when measuring the differences between the base and target geometries in Figure 2.2a, one obtains an expression for the RMSD and U values:

$$RMSD_{\text{HOOH}} = \sqrt{\frac{(0 + 0 + 0 + 0.024\text{\AA}^2)}{4}} = 0.078\text{\AA} \quad (5.9)$$

$$U_{\text{HOOH}} = |(0 + 0 + 0 + 0 + 0 + 1.049\text{Rad})| = 1.049\text{radians}. \quad (5.10)$$

Equations 5.9 and 5.10 show an example of this point. The final RMSD and U terms in each summation exist as 0.024\AA and 1.049 radians, respectively. Note that the RMSD only gives information on the distance between each atom. Hence Equation 5.9 shows that atom 4 (in this case a hydrogen atom), is out of place and all other atoms exist in identical positions between the two geometries. This is in contrast to U that possesses 6 terms in the summation, one for each DOF. This shows that the 6th term is 1.049 radians different between the two geometries and all other DOFs remain identical. In this example, the 6th term describes the difference in the torsion angle between the two molecular geometries, which is the source of the deviation; not the result of the deviation as the RMSD provides.

U is by no means a replacement for the RMSD, but provides another perspective when quantifying the differences between two molecular geometries. It is important to realise that measuring both values in tandem provides information on both the source and the result of the deviations in molecular DOFs. This cannot always be appreciated as small differences in a torsion angle can lead to a large RMSD value. The result of changes in DOFs is best measured using the RMSD and so both quantities are useful in their respective areas.

At least for this research, U is more powerful than the RMSD. However, U encompasses an inherent problem in that it is performing the summation over all terms which are partitioned into bond lengths, bond angles and torsion angles, where former exist in units of \AA and the latter two exist in units of radians.

Even though these two different types of units give values that are on the same order of magnitude, they are still not directly comparable (even though they get summed

together anyway). Note that RMSD does not possess this issue as all of the terms in the summation are measured in Å.

As much as a novel quantity that U is, probing into its capability and addressing this inherent issue will not advance further in this thesis. Nonetheless, temporarily implementing U is beneficial for achieving another solution to the geometry interconversion problem.

5.5 Direct Solution

Although an *RMSD Minimisation* procedure is scientifically robust and works for small, simple molecules such as hydrogen peroxide, a more computationally efficient methodology exists. This method commences by reformulating and coupling the ideas behind Equation 5.4, the *RMSD Minimisation* procedure and U . This can be used to create a solution that exists in curvilinear space:

$$U = \min_{\vec{s}} |\vec{q}_t^{\text{nr}} - (\vec{q}_b^{\text{nr}} + \mathbf{B}_{\text{nr}} \cdot (\vec{s} \cdot \mathbf{C}))| \quad (5.11)$$

where \vec{q}_t^{nr} and \vec{q}_b^{nr} are column vectors that possess the $3N - 6$ DOFs for the target and base geometries, respectively, \mathbf{B}_{nr} is the non-redundant Wilson B-matrix, \mathbf{C} possesses the Cartesian eigenvectors for each atom in each principal displacement and \vec{s} is another column vector where each element dictates the quantity of each principal displacement to be used.

By minimising U instead of the RMSD, this method provides an alternative minimisation procedure. However, Equation 5.11 can still be further developed. The eventual goal from this whole chapter is to yield the most accurate representation of the target geometry that is possible by displacing the base geometry along a combination of its principal displacements.

Therefore, setting $U = 0$ assumes that the base and target molecular geometries are identical. This also removes any ambiguity with the units of U as all values within the summation are also 0.

By removing the minimisation procedure and rearranging Equation 5.11, the desired vector, \vec{s} , can be made the subject:

$$0 = \vec{q}_t^{\text{nr}} - (\vec{q}_b^{\text{nr}} + \mathbf{B}_{\text{nr}} \cdot (\vec{s} \cdot \mathbf{C})) \quad (5.12)$$

$$\vec{q}_t^{\text{nr}} - \vec{q}_b^{\text{nr}} = \mathbf{B}_{\text{nr}} \cdot (\vec{s} \cdot \mathbf{C}) \quad (5.13)$$

$$\vec{s} = (\mathbf{C}^\top \cdot \mathbf{C})^{-1} \cdot \mathbf{C}^\top \cdot (\mathbf{B}_{\text{nr}}^\top \cdot \mathbf{B}_{\text{nr}})^{-1} \cdot \mathbf{B}_{\text{nr}}^\top \cdot (\vec{q}_{\text{t}}^{\text{nr}} - \vec{q}_{\text{b}}^{\text{nr}}) \quad (5.14)$$

$$\vec{s} = (\mathbf{C}^\top)^+ \cdot (\mathbf{B}_{\text{nr}})^+ \cdot (\vec{q}_{\text{t}}^{\text{nr}} - \vec{q}_{\text{b}}^{\text{nr}}) \quad (5.15)$$

where the ‘+’ represents a pseudo-inverse of a matrix.

The $(\mathbf{B}_{\text{nr}})^+ \cdot (\vec{q}_{\text{t}}^{\text{nr}} - \vec{q}_{\text{b}}^{\text{nr}})$ term is a constant and yields the exact infinitesimal displacements required to transform the base into the target geometry. This behaviour can be verified by removing given rows from \mathbf{C} (i.e. removing whole principal displacements) and re-solving Equation 5.15, which gives the same values within \vec{s} . Therefore the values of \vec{s} do not vary with the number or type of principal displacements possessed by the molecule. This shows that this solution yields one, and only one, displacement value for every principal displacement.

In addition, the $(\mathbf{B}_{\text{nr}})^+ \cdot (\vec{q}_{\text{t}}^{\text{nr}} - \vec{q}_{\text{b}}^{\text{nr}})$ term now possesses the displacement values in only units of Å. This matrix inversion eliminates the issue of combining molecular DOFs that exist in Å (bond lengths) and radians (bond angles and torsion angles).

Another advantage of this methodology is that, since all of the principal displacements are mutually orthogonal, there exists one, and only one, solution for each principal displacement. This immediately removes the issue of falling into RMSD local minima that may not be the exact (or most the accurate) solution and also vastly reduces the computational cost, as only one function evaluation needs to be performed.

5.6 Results

For each molecule in the test set, the crystal structure(s) were lattice energy minimised using the CrystalOptimizer algorithm that allows the molecular geometry to adjust to the crystalline environment. This employed the B3LYP-GD3BJ/6-311G level of theory for both the intermolecular electrostatic interactions and the intramolecular energy model. The in-crystal molecular (target) geometry was extracted from this energy minimised crystal structure and optimised in the gas phase to yield the base geometry that also employed the B3LYP-GD3BJ/6-311G** level of theory.

The *RMSD Minimisation* method was performed using the Simplex algorithm [72] for each molecule in the test set by minimising the geometrical difference between the base and target molecular geometries. The convergence criteria was a reduction in RMSD of less than 10^{-6} Å. The principal displacements were introduced successively, in order of increasing force constant, to examine the contribution of each principal displacement, if any, to the reduction in the RMSD.

In addition, the *Direct Solution*, Equation 5.15, was also calculated for each molecule in the test set. Owing to the orthogonality of the principal displacements, this method can also follow a successive procedure, adding one displacement at a time, to show how the exact solution to this problem compares to the *RMSD Minimisation* method.

The overall results are presented for each base conformer individually. The results for each base conformer are summarised in two figures: a comparison of the RMSD overlay that employed COMPACK where a molecular match at a distance tolerance of 20% and angle tolerance of 20° was required to provide a satisfactorily reproduced molecular geometry between the base and target geometries for both the *RMSD Minimisation* and *Direct Solution* techniques and the value of each displacements for each principal displacement.

Two specific cases will be presented in this section: ODNPDS and FIBKUW. The former highlights a typical example of the overall test set whereas the latter is a more strained system and therefore allows for a rigorous test of the 2 methods. The data pertaining to the other molecules in the test set are provided in Appendix B, Figure B.2.

5.6.1 ODNPDS

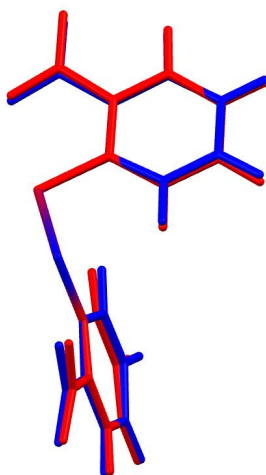
ODNPDS possesses 2 polymorphs that possess similar in-crystal geometries, Figure 5.4a, with an RMSD of 0.079 Å. This small RMSD value classifies these 2 polymorphs as packing polymorphs.

The base-target geometry conversion of the 2 ODNPDS conformers represent a typical example of the data that is observed within the test set of molecules. Figure 5.4 displays the overlay of the base and target geometries for ODNPSD02 and ODNPSD11 as well as an overlay of the 2 target geometries.

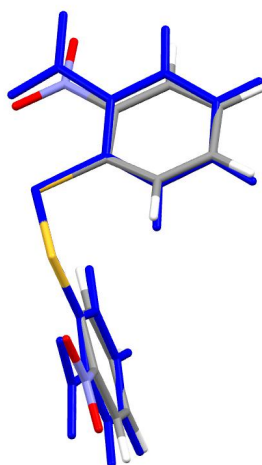
The difference between the base and target geometries for both crystal structures is subtle, with the RMSD of approximately 0.35 Å. The geometrical differences for ODNPDS02 mainly arise from a +6° twist in the SSCC torsion angle. The differences for ODNPDS11 are also due to a SSCC torsion twist of +5°. Additional distortions for both conformers also include a twisting of +7° and +14° of both of the nitro groups out of the plane of the aromatic ring.

Perhaps surprisingly, the CSSC exocyclic torsion angle in both conformers deviates by < 1° from the base geometry. Although this is one of the 3 soft torsion angles in the system, it is not distorted by the crystal packing forces in either crystal structure.

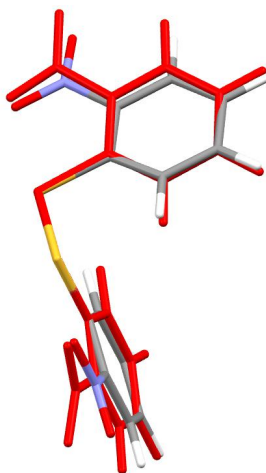
Figures 5.5a and 5.5c display the results of the *Direct Solution* and the *RMSD Minimisation* methodologies for ODNPDS02 and ODNPDS11, respectively. For the *RMSD*



(a) ODNPDS Targets: $\text{RMSD} = 0.079\text{\AA}$



(b) ODNPDS02: $\text{RMSD} = 0.344\text{\AA}$



(c) ODNPDS11: $\text{RMSD} = 0.359\text{\AA}$

Figure 5.4: Overlay of the target and base (coloured by element) geometries for the ODNPDS02 (blue) and ODNPDS11 (red) molecules.

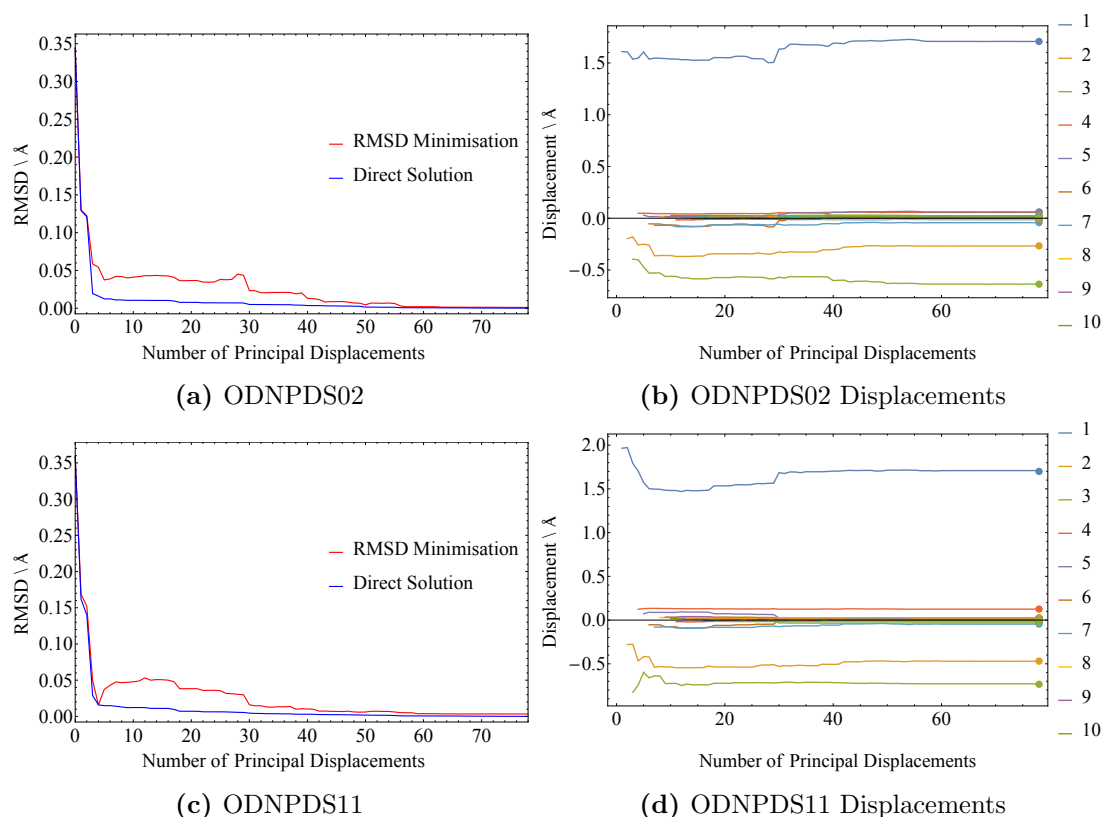


Figure 5.5: (a) and (c) show the reduction in RMSD as more numbers of principal displacements are added into the procedure for ODNPDS02 and ODNPDS11, respectively. (b) and (d) show the displacements for each of the principal displacements with the 10 lowest valued force constants labelled for ODNPDS02 and ODNPDS11, respectively.

Minimisation, each principal displacement was introduced successively whereby the principal displacement with the lowest force constant was optimised in isolation, such that the RMSD was minimised. The next stage was then to use this displacement value as a starting point and now include the second lowest force constant principal displacement in the minimisation, such that both were optimised in tandem to reduce the RMSD. This process was repeated until all principal displacements were included. Although the *Direct Solution* calculates all displacements simultaneously, for a valid comparison, the RMSD was measured at the inclusion of each principal displacement for this method too.

In addition, Figures 5.5b and 5.5d show the displacement values for each principal displacement during each phase of the *RMSD Minimisation* method. To label each principal displacement on these figures would render to graph cluttered, so to avoid this, only the first 10 principal displacements are listed in the legend. These are the principal displacements that generally provide relatively large displacement values and reduce the RMSD by the greatest amounts.

The large points at the end of each trace show the *Direct Solution* values for that particular principal displacement. This solution only yields 1 displacement value for each principal displacement as it is independent of the number of principal displacements that are used.

For ODNPDS02, both methodologies reduce the RMSD by 0.214 Å when using only the first principal displacement. The displacement values, for the *Direct Solution* and *RMSD Minimisation* techniques, differ by 0.001 Å at 1.695 Å and 1.708 Å respectively. Both methodologies continue to further reduce the RMSD using the second principal displacement but the 2 methodologies then diverge upon the inclusion of the third principal displacement. From here, the RMSD is reduced by an additional 0.040 Å by the *Direct Solution*. This is reflected in Figure 5.5b where the displacement values for the first and second principal displacements are affected, such that it hinders the *RMSD Minimisation* which does not find the same RMSD minimum as the *Direct Solution*. This greatly affects the performance of the *RMSD Minimisation* procedure up until principal displacement 57, where both methods now approximately yield the same RMSD value. This also correlates with the displacement values as these are unpredictable at the early stages of the *RMSD Minimisation* procedure but then eventually converge on the value calculated by the *Direct Solution*.

The contribution of each principal displacement decreases as the force constant value increases. This is because the force constant dictates how the molecular energy changes when the molecule is displaced along a, or combination of, principal displacements, where larger force constants incur higher energy costs for a given displacement. Thereby major movements that will reduce the RMSD by the greatest amounts, from larger displacement values, generally occur early in the list. The principal displacements that perform smaller geometrical movements are only used to reduce the RMSD by smaller, even negligible, amounts.

The importance of the principal displacements are ranked as $1 > 3 > 2$ since the RMSD reduces by 0.214 Å and 0.100 Å for the first and third principal displacements, respectively, where the second only reduces the RMSD by 0.010 Å. This highlights an important point in that merely because the force constant is smaller than that of the crystal packing forces, it cannot be assumed that the molecule will be displaced along all of those principal displacements. It is clear that which principal displacements a molecule will be displaced along are also dependent upon other factors. These factors will be expanded on further in Chapter 6.

Another important comparison to make between the 2 methods is the sign of the changes in the RMSD value. Each contribution to the *Direct Solution* always reduces the RMSD

whereas the *RMSD Minimisation* can exhibit occasional increases in the RMSD. This is counter-intuitive as a minimisation technique should always reduce the RMSD value.

However, the explanation lies in the initial step made by the minimiser. This can be large enough to step out of the current RMSD well and find another minimum. This new minimum can lead to a better or worse solution, but in most cases it results in the latter. This highlights another important point in that a local *RMSD Minimisation* technique is not robust enough to yield the best possible solution at low numbers of dimensions.

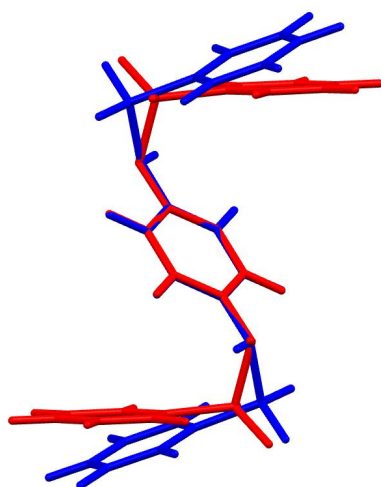
There is no guarantee that a global *RMSD Minimisation* technique would have found the global RMSD minimum. This is in contrast to the *Direct Solution* that allows all displacement values to be calculated simultaneously, leading to the best possible solution.

The analysis of ODNPDS11 draws similar conclusions to that of ODNPDS02. Figures 5.5c and 5.5d show that, again, it is principal displacements 1 and 3 that yield the largest reductions in RMSD. In contrast, the *RMSD Minimisation* method tracks the *Direct Solution* until principal displacement 4. From here, the RMSD increases using the *RMSD Minimisation* method and proceeds to never find the same minimum that is found by the *Direct Solution*. This is due to a difference in the displacement values for the principal displacements with higher force constants. This effect is observed in Figure 5.5c and shows that the *RMSD Minimisation* method can be unstable and yield a solution other than the ideal one; even at these small geometrical differences. This effect can be observed in Figure 5.5d where the displacement values of principal displacements 1, 2 and 3 are unpredictable when less than 10 are included in the minimisation. In particular, the inclusion of the fourth principal displacement causes a reduction in contribution of these first 3 principal displacements.

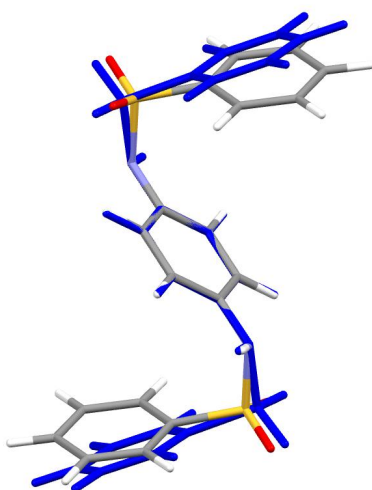
Regardless of this issue with the *RMSD Minimisation*, the difference in RMSD for the final solution for both methodologies is 0.003 Å. This is a negligible quantity that will not affect the crystal packing and therefore is purely an artefact of the calculations.

5.6.2 FIBKUW

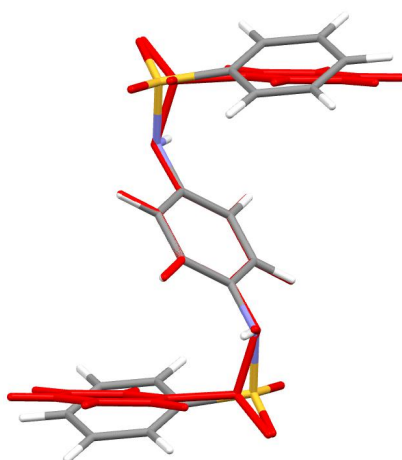
Whilst ODNPDS was chosen as a typical example of what is observed for the majority of the molecules in the test set, the FIBKUW polymorphs provide a rarer example of when larger geometric distortions exist between the base and target geometries. This therefore highlights a more stringent test for both methodologies, Figure 5.6.



(a) FIBKUW Targets: RMSD = 0.307Å



(b) FIBKUW01: RMSD = 0.424Å



(c) FIBKUW02: RMSD = 0.467Å

Figure 5.6: Overlay of the target (coloured) and base (coloured by element) geometries for the FIBKUW01 (blue) and FIBKUW02 (red) molecules.

The RMSD between the base and target geometries are 0.424 Å and 0.467 Å for FIBKUW01 and FIBKUW02, respectively, and are the highest in the test set. These large geometrical distortions arise from rotations about the 6 exocyclic bonds present in the molecule. These are soft torsion angles that are easily distorted by crystal packing forces. Both of the target geometries optimise to the same gas phase conformer and both possess distortions through 15° twists of the terminal carbon rings, Figure 5.6.

The target geometries differ through a 30° rotation of the central carbon ring. The target geometry for FIBKUW02 possesses a more ‘closed’ conformation as the terminal phenyl rings are in closer proximity to the central ring. Whereas FIBKUW01 possesses a more ‘open’ conformation, such that these terminal rings are extended away from the central ring.

The identical nature of the base geometry, but differing natures of the target geometries, also allows this system to test the methodologies to produce different conformers from the same starting point. More specifically, these methods will be using the same set of principal displacements to produce 2 different in-crystal conformations. Hence, a different combination of the same movements will be required to reproduce the target geometries.

Figures 5.7a and 5.7c show the reduction in the RMSD for both methodologies when allowing more principal displacements to be used for the geometry conversion. It can be observed that both methodologies for both base-target conformations find the same respective minima once all principal displacements have been added. However, the *RMSD Minimisation* method consistently yields slightly larger RMSD values after the 5th until approximately the 85th principal displacement.

For FIBKUW01, it is the 5th and 6th principal displacements that reduce the RMSD by 0.232 Å and 0.270 Å for the *Direct Solution* and *RMSD Minimisation* methods, respectively. These contributions are reflected in Figure 5.7b that shows the displacement value for each principal displacement as it is introduced into the *RMSD Minimisation* technique.

It is observed that the 5th and 6th principal displacements possess larger displacement values than the rest of the set. This is with exception to the 1st principal displacement that consistently possesses a large displacement value, which initially swings violently from negative to positive as it compensates for the inclusion of the additional principal displacements. Nonetheless, the values converge onto those that are calculated by the *Direct Solution*.

FIBKUW02 possesses a similar trend to FIBKUW01 where only a small number of principal displacements are required to vastly reduce the RMSD. In this instance, it is the

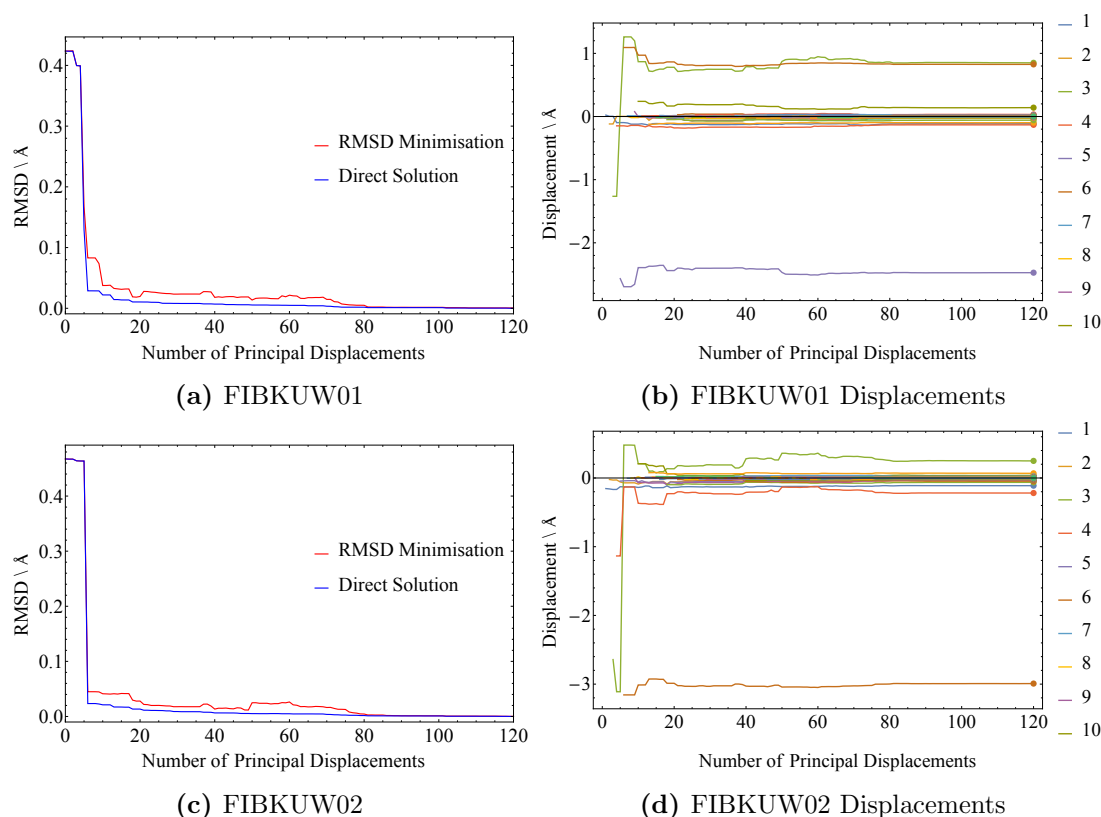


Figure 5.7: (a) and (c) show the reduction in RMSD as more numbers of principal displacements are added into the procedure for FIBKUW01 and FIBKUW02, respectively. (b) and (d) show the displacements for each of the principal displacements with the 10 lowest valued force constants labelled for FIBKUW01 and FIBKUW02, respectively.

6th principal displacement only. Both methodologies acknowledge that this movement is required to reduce the RMSD; albeit the RMSD minimiser does not converge to the desirable RMSD value with as few principal displacements as the *Direct Solution*.

Again, this is reflected in Figure 5.7d where the 6th principal displacement possesses the largest value at -3.44 \AA . It is interesting to note that the 3rd principal displacement possesses a -3.11 \AA value before it violently changes direction upon the inclusion of the 6th principal displacement. This suggests that the former is unsuccessfully attempting to perform the function of the latter until the actual movement required is included.

The 2 target conformations of FIBKUW are highly strained but show that a simple solution can still be yielded. Therefore it is not dependent on the extent of geometrical differences between the base-target pair, as the solution can comprise of one or several principal displacement combinations.

5.7 Conclusions

DFT-D calculations were performed on a test set of molecules where the in-crystal molecular geometries were extracted before this was geometry optimised in the gas phase. This formed a set of target (in-crystal) and base (gas phase) geometries with the goal of using the principal displacements of the latter to perform a geometry conversion into the former.

Firstly, a *Pure Cartesian Approach* was presented but was not considered scientifically robust as this method failed to appreciate the difference between curvilinear and linear space. Nonetheless, this provided an important theoretical basis for the *Direct Solution* technique.

Two scientifically valid methodologies were then derived: one performed a numerical minimisation of the RMSD between the base and target geometries where each principal displacement was introduced successively and the other calculated a *Direct Solution* for all of the principal displacements simultaneously.

Although both methods were performed on all of the molecules in the test set, the specific cases of ODNPDS and FIBKUW were analysed in more detail and they provided an example of a typical and a more extreme difference between the base and target geometries respectively. Both methods were successful in finding an almost exact solution to the problem.

However, the *RMSD Minimisation* procedure proved relatively computationally expensive. This is due to the large number of function calls that were required. This is in contrast to the *Direct Solution* that requires only one function evaluation and hence is the more computationally efficient method.

In addition, the *RMSD Minimisation* method generally yielded worse geometrical matches at lower numbers of principal displacements than the *Direct Solution*. The values required for the lower principal displacements could also change rapidly for the RMSD minimiser which could easily place the displaced gas phase conformer into the wrong area of the RMSD hypersurface. Whereas, once again, the *Direct Solution* provided the values that would be found eventually by the RMSD minimiser in one function evaluation.

Although both methods were generally successful, it is clear that the *Direct Solution* is far superior in terms of computational expense and stability, than the *RMSD Minimisation* method. Hence, this methodology can be used to analyse the role of principal displacements such that rules can be formulated for their implementation within CSP.

The questions posed at the end of Chapter 4 can now begin to be answered when converting the gas phase into the in-crystal geometry notably: which principal displacements are required, how many are required and how far must be traversed along each one? Arguably, a hint of these answers has already been provided in the Section 5.6 of this chapter, however this small test set of molecules remains limited and to truly derive meaningful answers to these questions, a larger set of molecules must be used.

The following chapter attempts to decompose and more fully understand this ΔE_{strain} term by means of the *Direct Solution* presented in this chapter. This is with the intention of using the answers for implementation of the principal displacement crystal structure generation method devised in the Chapter 4.

Chapter 6

Decomposition of Molecular Strain in Crystals

6.1 Introduction

Chapter 5 demonstrated that using a *Direct Solution*, Equation 5.15, allowed successful interconversion between two molecular conformations that is also computationally efficient. However, the question still remains from Chapter 4 as to which principal displacements and what is the extent of the distortion to the molecular geometry that are imposed by the crystalline environment. Therefore a large test set of flexible molecules will need to be collected and each molecule subject to the calculations described in the previous chapter.

In addition, this test of molecules will also need to possess more exocyclic, soft torsional angles as well as more atoms per molecule than would usually be computationally practical for a flexible molecule CSP procedure at present. This is also due to a curiosity to define a new boundary for which Equation 5.15 can still operate comfortably and with scientific rigour. Performing similar analysis on these new, more flexible molecules in addition to those collected by Thompson & Day [120], will provide a more detailed insight into describing molecular distortions by the crystal packing forces using molecular principal displacements.

Furthermore, information on the ΔE_{strain} values, the energy required to strain the molecule from its gas phase to its in-crystal conformation, for these molecules can also be collected on-the-fly which can energetically quantify the extent of molecular distortions in the crystal. Part of what follows is the largest reported collection of calculated

ΔE_{strain} values for molecular organic crystals, which reveals the energetic behaviour of flexible molecules when subject to these crystal packing forces.

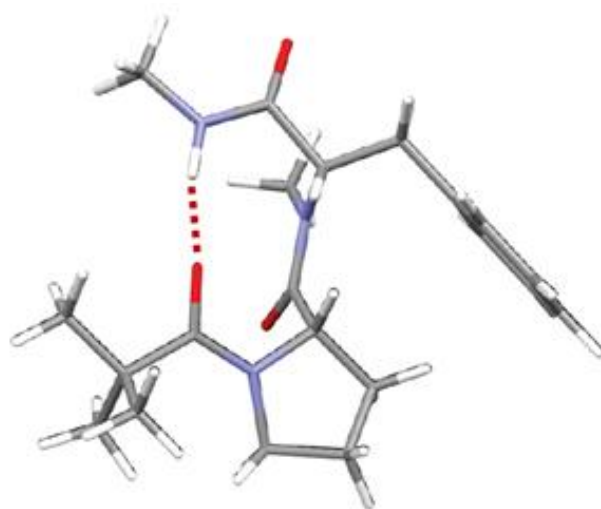
6.2 Defining Non-Bonded Intramolecular Interactions

A major focus of this chapter is the effect of the formation of non-bonded intramolecular interactions on the ΔE_{strain} and their effect on the geometry of the molecular conformer. To be more specific, these refer to interactions that occur between atoms of the same molecule that are not covalently bonded together. These interactions are categorised into three flavours for the purposes of this research: hydrogen bonds, polar interactions and non-polar interactions. An example of these three types of non-covalent intramolecular bonds are visualised in Figure 6.1.

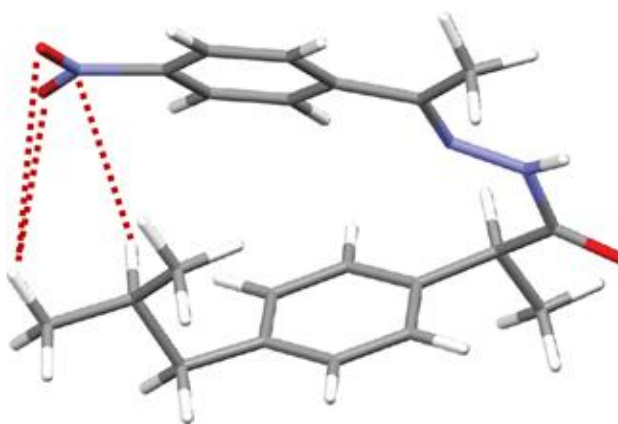
Figure 6.1a shows an example of a non-covalent intramolecular interaction in the form of a hydrogen bond. This is formed between the carboxyl and amine groups of the molecule and is the strongest type of the 3 forces. Figure 6.1b shows a non-covalent, polar interaction between a nitro functional group and a benzene ring. This type of interaction is not as strong as a hydrogen bond but still aids in stabilising the molecular conformer. Figure 6.1c shows an example of a non-covalent, non-polar interaction where both of the aromatic and aliphatic functional groups are non-polar and form the weakest type of interaction in this categorisation scheme.

These non-covalent intramolecular interactions are significant as their formation can affect the conformation of a molecule. If the molecule is sufficiently flexible and possesses the relevant atom types to form one of the three non-covalent intramolecular interactions, then the attractive force that results between the two functional groups will bring them into close proximity such that a molecular energy minimum is reached. This energy minimum will be lower in energy and result in a more stable molecular conformer than if these non-covalent intramolecular interactions did not exist.

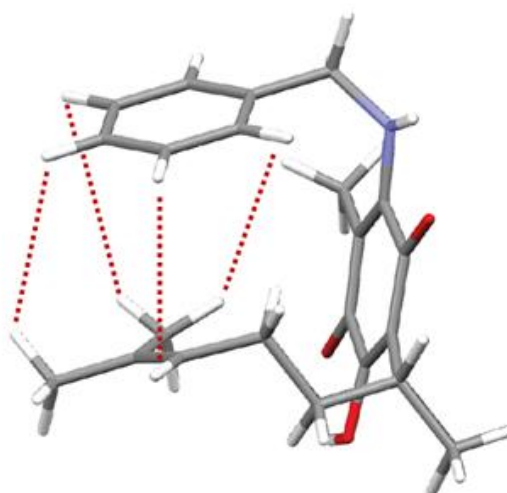
The ΔE_{strain} value is calculated by taking the energy difference between the in-crystal conformation and its corresponding gas phase conformer. Therefore if a particular type of non-covalent intramolecular interaction is present for the gas phase conformer and not in the in-crystal conformation, the ΔE_{strain} will be larger than if this interaction was not formed. More specifically, and in relation to Figure 2.10b, these non-covalent intramolecular interactions are formed between two molecular conformations that reside in the same potential energy well.



(a)



(b)



(c)

Figure 6.1: (a), (b) and (c) showing examples of non-covalent intramolecular hydrogen bonding, polar interactions and non-polar interactions, respectively, highlighted by the red, dashed lines.

6.3 Methodology

6.3.1 Obtaining a Test Set of Molecules

The CSD possesses a wealth of crystal structures consisting of many different packing arrangements and combinations of molecules. Therefore it was necessary to apply a strict search criteria when searching for the test set of molecules required for this study. This was not only to find exactly the type of molecules that were needed, but to also obtain a realistic number of crystal structures for which the calculation could be performed, with respect to computational expense.

The search criteria, along with an explanation for each constraint exists as follows:

1. **Number of atoms per molecule > 25** - A simple requirement to encourage larger molecules to be included. This follows an assumption that larger molecules possesses a greater chance of being more flexible but nonetheless will increase the computational cost.
2. **Number of rotatable bonds per molecule > 2 and < 8** - This refers to the number of exocyclic bonds per molecule, again, to encourage molecules that are more flexible to be included in the study but also to limit computational expense.
3. **Oxygen atom count < 5** - A balance exists between desiring the molecule to form non-covalent intramolecular interactions and possessing too many, given that the molecule is flexible enough. The latter could yield the molecule to be less receptive to crystal packing forces. This constraint attempts to control this issue by allowing non-covalent intramolecular interactions to occur but only in moderation.
4. **Number of hydrogen bond donors > 0** - This approaches constraint 3 but from the opposite perspective and therefore encourages the molecules to have the capability to form hydrogen bonds.
5. **Number of aromatic bonds < 15** - Although the molecules are desired to possess a large number of atoms, this does not guarantee that the molecule will be flexible as many atoms that can be part of rigid, aromatic groups. Aromatic moieties in CSP are often classed as rigid bodies as they are only distorted by a negligible amount in the crystalline environment. This constraint seeks to limit this behaviour by facilitating larger molecules with limited aromaticity hence promoting flexibility.

6. **Number of rings < 4** - This refers to both aromatic and non-aromatic systems and, again, seeks to exclude molecules with large numbers of atoms that do not possess soft dihedral angles.
7. **Molecules must be uncharged** - This also excludes zwitterionic molecules too and formal charges on molecules are often a challenge for current CSP methodologies. This study seeks to isolate and investigate the flexibility issue within CSP and not focus on other challenges within the field.
8. **Molecules only contain atoms types: H, C, O, N and S** - The crystal structure calculations are well defined for these atom types whereas the inclusion of more 'exotic' atom types can lead other challenges (that are beyond the scope of this thesis). It is desirable to test the limits of the principal displacement methodology whereas the inclusion of other atoms types can lead to ambiguity in the results of those molecules. However, no limit was set on the composition of molecules with respect to these atom types (with the exception to point 3 that limits the number of oxygen atoms).

Further constraints were then placed on the crystal structures yielded from this set of molecular criteria:

1. **Number of molecules per asymmetric unit = 1** - This constraint allows for convenience of the CSP calculations when implementing DMACRYS in that crystal structures possessing a $Z' > 1$ can yield unit cells with large numbers of atoms that substantially add to the computational cost. Conversely, a $Z' < 1$ structure can cause software issues within DMACRYS with respect to symmetry as less than one molecule exists in the asymmetric unit. This constraint also excludes hydrated crystals and co-crystals.
2. **No disorder in the system** - Ambiguity in the atomic positions therefore led to uncertainty in the crystal structure. To accurately check as to whether the calculations performed on the crystal structure are valid, this ambiguity needs to be removed and hence this constraint is applied. For this study, the removal of these molecules was conducted manually by eye.

Both sets of these constraints yielded a set of 149 crystal structures (the CSD reference codes for are listed in Appendix C), Set 1.

An interesting feature of a selection of molecules in this set is that they are capable of forming some type of non-covalent intramolecular interaction. Further analysis of this property in Set 1 will be performed in Section 6.3.2. Nonetheless this builds on the

molecules used by Thompson & Day [120] that are only capable of forming intermolecular bonds, if any.

The occurrence of non-covalent intramolecular interactions could result in higher ΔE_{strain} values as these bonds will most likely, assuming the molecule is flexible enough, be present in the gas phase but not necessarily present in the in-crystal conformations. Therefore these bonds will need to be broken to convert the gas phase into the in-crystal geometry. This requires more energy to perform and hence could yield higher ΔE_{strain} values.

In addition, the principal displacements required to strain the gas phase into the in-crystal geometry could require higher valued force constants. Therefore more of the principal displacement space would need to be explored adding to the computational cost of the calculation.

Since the ΔE_{strain} and principal displacement calculations have already been performed on the molecules in the set of Thompson & Day [120], Set 2, these molecules will be added to the 149 molecules in Set 1. This yields 175 molecules that all possess similar characteristics.

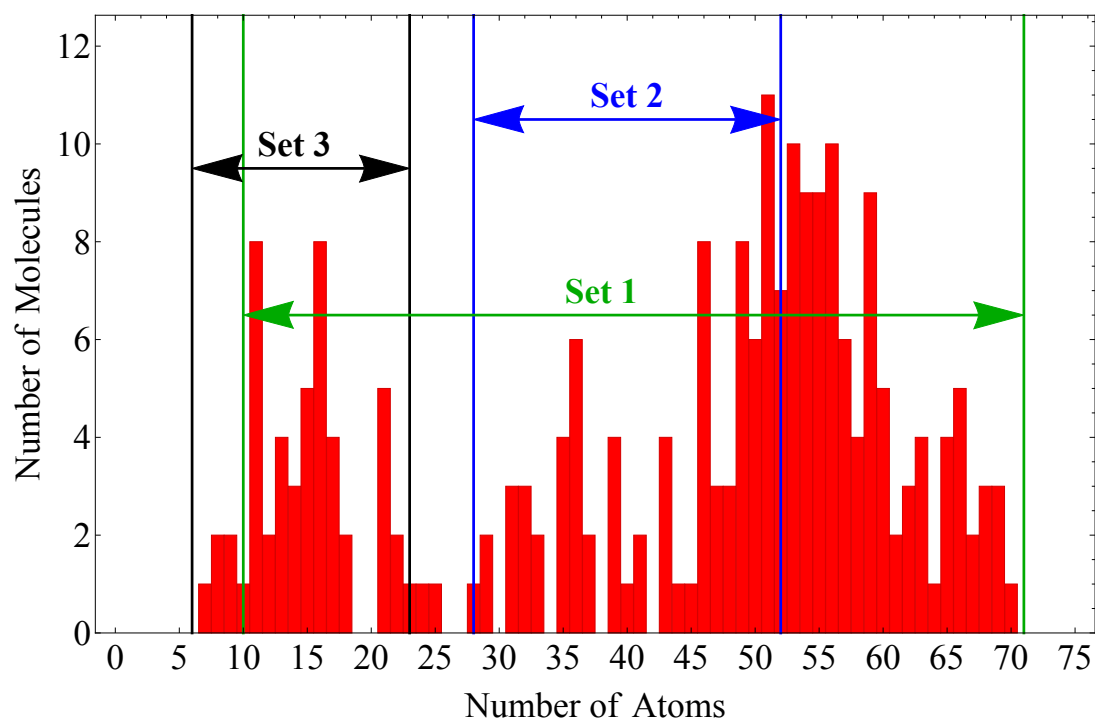
Whilst 175 molecules give a large enough sample size, CSP calculations have already been performed on the 54 molecules presented in Chapter 3 also. These molecules are smaller and less flexible than those in Sets 1 and 2. However, these molecules possess between 0 and 2 rotatable bonds and should possess ΔE_{strain} values that are lower than those of the large molecules.

To keep the search criteria consistent, it is necessary to omit the 5 polymorphs of glycine from the set of 54 molecules to yield 49 molecules, Set 3, that are added to the 175 molecules of Sets 1 and 2 to, finally, form a complete test set that consists of 224 molecules.

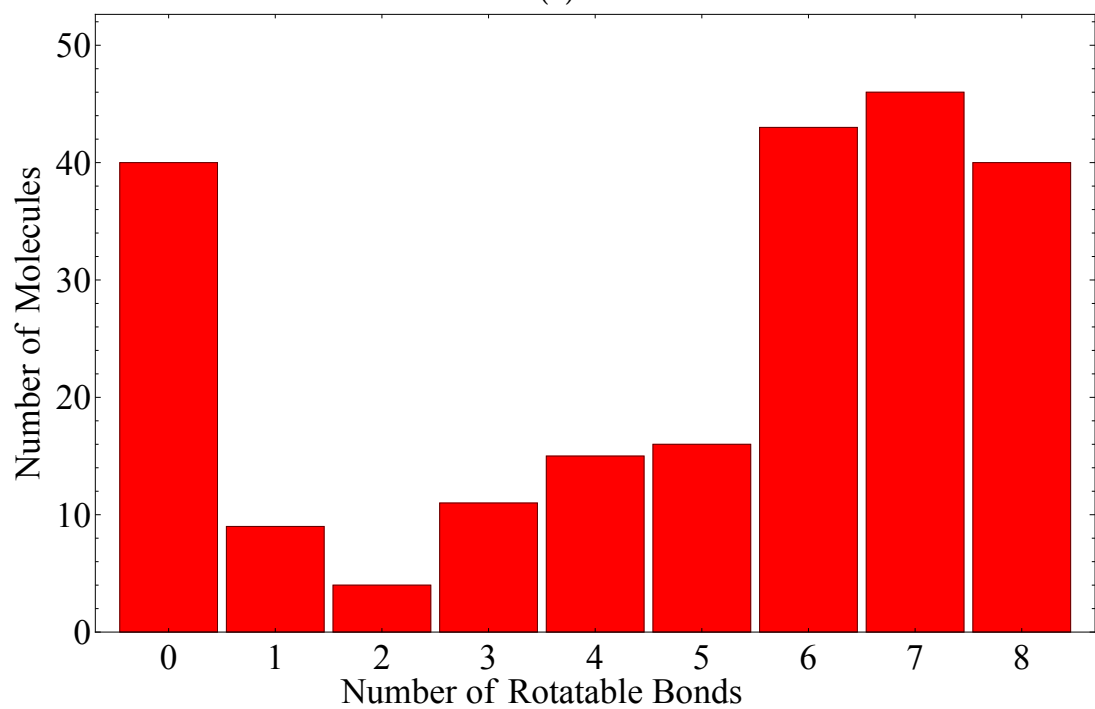
6.3.2 Analysis of the Test Set

This large test set of 224 molecules possess a variety of molecular systems of differing shapes and sizes. Therefore it is useful to understand the distribution of some basic statistics since there have been two other sets of molecules added to the initial 149 that originally existed.

Figure 6.2a shows the distribution of the number of atoms of all 224 molecules in the test set. The range of the number of atoms vary between 6 and 70 with the molecules from Set 3 dominating the lower values.



(a)



(b)

Figure 6.2: (a) shows the distribution of the number of atoms with the ranges of molecular Sets 1, 2 and 3 and (b), the number of rotational bonds in the 224 molecule test set.

The distribution is generally evenly sampled with only a small amount of atom counts not being included. These values exist around the middle region between approximately 19 and 24 atoms, although it is not essential to sample every value multiple times. The atom count ranges for the 3 Sets of molecules are shown in Figure 6.2a, where Set 1 spans almost the entirety of the total range.

Figure 6.2b now attempts to quantify the potential for the 3 Sets of molecules to exhibit flexibility, as opposed to observing the sizes of the systems based on atom count. The number of rotational bonds range from 0 to 8. This allows the molecules to possess very small to very large amounts of flexibility and should give a varied range of ΔE_{strain} values.

The distribution in this plot contains evenly sampled areas, notably molecules with 0, 6, 7 or 8 rotatable bonds that possess approximately 40 molecules in each bin. The number of molecules that possess 1, 2, 3 or 4 rotatable bonds are less common but still consist of approximately 10 to 15 molecules. This is with the exception of the 2 rotatable bonds bin which has only possesses 3 molecules.

Again, it is the more flexible molecules that are important to analyse here so a deficiency in the smaller rotational bond numbers is a not a significant disadvantage. They are indeed an added bonus that gives a slight insight into these molecules that possess limited flexibility and would not have been available otherwise; thus padding out the results more substantially.

As mentioned in Section 6.3.1, a proportion of these molecules possess the ability to form both non-covalent intra- and intermolecular bonds. This not only includes the possession of hydrogen bond donor and acceptor atoms, but also the ability to allow these atoms to become in close proximity to one another such that some type of non-covalent intramolecular interaction is formed. Therefore a moderate level of flexibility is required or that the donor-acceptor pairs are located close enough to one another within the molecule.

Of the 224 molecules, there is a total of 60 molecules that possess the correct functional groups in the relevant positions to form some type of non-covalent intramolecular interaction. 35 of these 60 molecules that could form a type of non-covalent intramolecular interaction have the ability to form a hydrogen bond whereas an additional 7 and 18 molecules are capable of forming polar and non-polar bonds, respectively.

With application to this research, this is referring to a fact that a molecule will possess non-covalent intramolecular interactions in the gas phase that are not present in the in-crystal geometry. It is not guaranteed that all 60 of these molecules will possess this

property but those that do will experience an additional stabilisation and hence possess larger ΔE_{strain} values than if these interactions were not present.

This 224 test set of molecules represent a wide variety of organic molecules in general. Of course there exists molecules that are larger and more flexible than these included but for the purposes of this research, the results yielded will pave the next step to understanding the affect of crystalline packing forces on molecular conformations.

6.3.3 Calculating ΔE_{strain} and Molecular Principal Displacements

The observed crystal structure of each molecule was obtained from the CSD. Each crystal structure was subject to an initial CrystalOptimizer calculation that employed a B3LYP-GD3BJ/6-311G** level of theory for both the intermolecular electrostatic interactions (using atomic multipoles from a DMA) and the intramolecular energy model. The converged in-crystal molecular geometry was extracted and subject to a gas phase geometry optimisation using the same B3LYP-GD3BJ/6-311G** level of theory followed by a calculation of the principal displacements for the optimised geometry using the Hessian whose 2nd order derivative elements were calculated numerically.

The ΔE_{strain} was calculated by subtracting the energy of the in-crystal conformation from the energy of the gas phase conformer that will always yield a positive value as it is assumed that the former is always strained away from the latter.

Using the principal displacements of the gas phase conformer, the *Direct Solution* (Equation 5.15) was implemented to calculate the contribution of each principal displacement to the interconversion between the gas phase and in-crystal molecular geometries.

6.3.4 Reproducing the Observed Crystal Structure

Whilst the in-crystal geometry can be accurately reproduced using the principal displacements of the gas phase geometry, it is undesirable to implement all principal displacements during the flexible molecule structure generation phase. This is due to the conformational search procedure outlined in Chapter 4 that explores the many dimensions of conformational space, where each dimension is a principal displacement.

Hence, if all principal displacements are included, the dimensionality of the space that is required for searching is too large and hence renders the technique computationally expensive to the point where it becomes unusable. Therefore it is useful to know the minimum number of principal displacements required to create an approximate, but accurate enough, in-crystal geometry.

The required accuracy of the molecular geometry is hence tested by performing multiple rigid body lattice energy minimisations on various displaced gas phase conformers. The work flow for this procedure is visualised in Figure 6.3 and will now be discussed in detail.

Each crystal structure is extracted from the CSD and a flexible molecule energy minimisation is performed to yield crystal structure **(B)**. The in-crystal geometry, **(1)**, is then extracted and geometry optimised in the gas phase to produce conformer **(2)**. The ΔE_{strain} is then calculated by taking the energy difference between conformers **(1)** and **(2)**. Conformer **(2)** is then also subject to a principal displacement calculation which is then used to yield a set of principal displacements that converts the geometry of **(2)** into **(1)**.

To produce an approximate molecular geometry, **(3)**, the principal displacements for each molecule were ordered by ascending force constant value. Conformer **(1)** was displaced along the first principal displacement by the amount specified from the *Direct Solution*. If the RMSD between **(3)** and **(1)** was $< RMSD_{\text{tol}}$, the approximate geometry was accepted and inserted into crystal structure **(B)** (in place of **(1)**). However, if the $RMSD > RMSD_{\text{tol}}$, the following principal displacements were sequentially introduced until the $RMSD < RMSD_{\text{tol}}$.

A set of RMSD tolerances, $RMSD_{\text{tol}}$, were chosen from 0.1 Å to 1.0 Å in increments of 0.1 Å. This range was chosen under the assumption that an $RMSD_{\text{tol}}$ that lies outside of these bounds is too extreme (where the in-crystal approximation is either too tight or too loose). Hence it would be expected that a ‘tipping-point’ will be observed in this range where a high proportion of crystal structures will and will not be found when the $RMSD_{\text{tol}} = 0.1$ Å and 1.0 Å, respectively. This process afforded 10 crystal structures per molecule, each containing a molecular geometry that approximates to the exact in-crystal geometry.

Each of these crystal structures was subjected to a rigid-body, lattice energy minimisation before being compared to crystal structure **(B)**. The red arrow in Figure 6.3 shows that, providing the $RMSD_{\text{tol}}$ was tight enough, the flexible molecule energy minimised crystal structure, **(B)**, will be reproduced. The comparisons of crystal structures **(B)** and **(C)** were performed using COMPACK where a 30/30 molecule match at a distance tolerance of 20% and angle tolerance of 20° was required to provide a satisfactorily reproduced crystal structure.

Following this procedure will show how close an approximation to the in-crystal geometry is required before the original crystal structure is satisfactorily reproduced. This firstly gives insight into reducing the number of principal displacements that are required for

the geometry conversion from gas phase to in-crystal, but it also yields insight in how sensitive the crystal structure is to varying changes in the molecular geometry.

6.4 Results

The amount of data generated for each molecule from the methodologies just described was vast. Therefore a large amount of analysis can commence that gives ample insight into the behaviour of flexible molecules in the crystalline environment.

The molecules throughout this chapter are consistently partitioned into two categories. These categories arise from the fact that some of the 224 molecules form some type of non-covalent intramolecular interaction upon the geometry optimisation of the in-crystal geometry. Therefore these molecules will experience an additional energetic stabilisation and hence an increase in the ΔE_{strain} value to beyond what would have been yielded if these non-covalent intramolecular interactions were not present. Hence the molecules that possess any form of non-covalent intramolecular interactions are highlighted during the analysis in order to accurately draw conclusions from the results.

In addition, of the molecules that do form non-covalent intramolecular interactions, these molecules are further partitioned into whether the interactions are polar, non-polar or hydrogen bonding. These types of interactions are of differing strengths and will contribute different amounts to the ΔE_{strain} value. The CSD reference codes for the molecules that exist in these intramolecular interacting groups are listed in Appendix C.

6.4.1 ΔE_{strain} Comparison: CrystalOptimizer versus CRYSTAL09

Before the commencement of the ΔE_{strain} analysis for the 224 molecule test set, it is important to obtain a sense for these values from a familiar, smaller set of the molecules. This occurs in the form of the 26 molecular conformers from Chapter 5.

Previous ΔE_{strain} values have already been computed for these molecules [120] using the CRYSTAL09 [68] periodic DFT software package using a B3LYP/6-31G* level of theory with an empirical dispersion correction proposed by Civalleri *et al.* [109]. However the ΔE_{strain} results from this chapter were solely computed from the energies calculated by the CrystalOptimizer algorithm, $\Delta E_{\text{strain}}^{\text{CrysOpt}}$. Hence before the analysis of all of the ΔE_{strain} values, a comparison will be made between the ΔE_{strain} values yielded from both the CrystalOptimizer and CRYSTAL09 methodologies. These results are presented in Table 6.1 (and visualised in Figure 6.4) and are listed in order of decreasing number

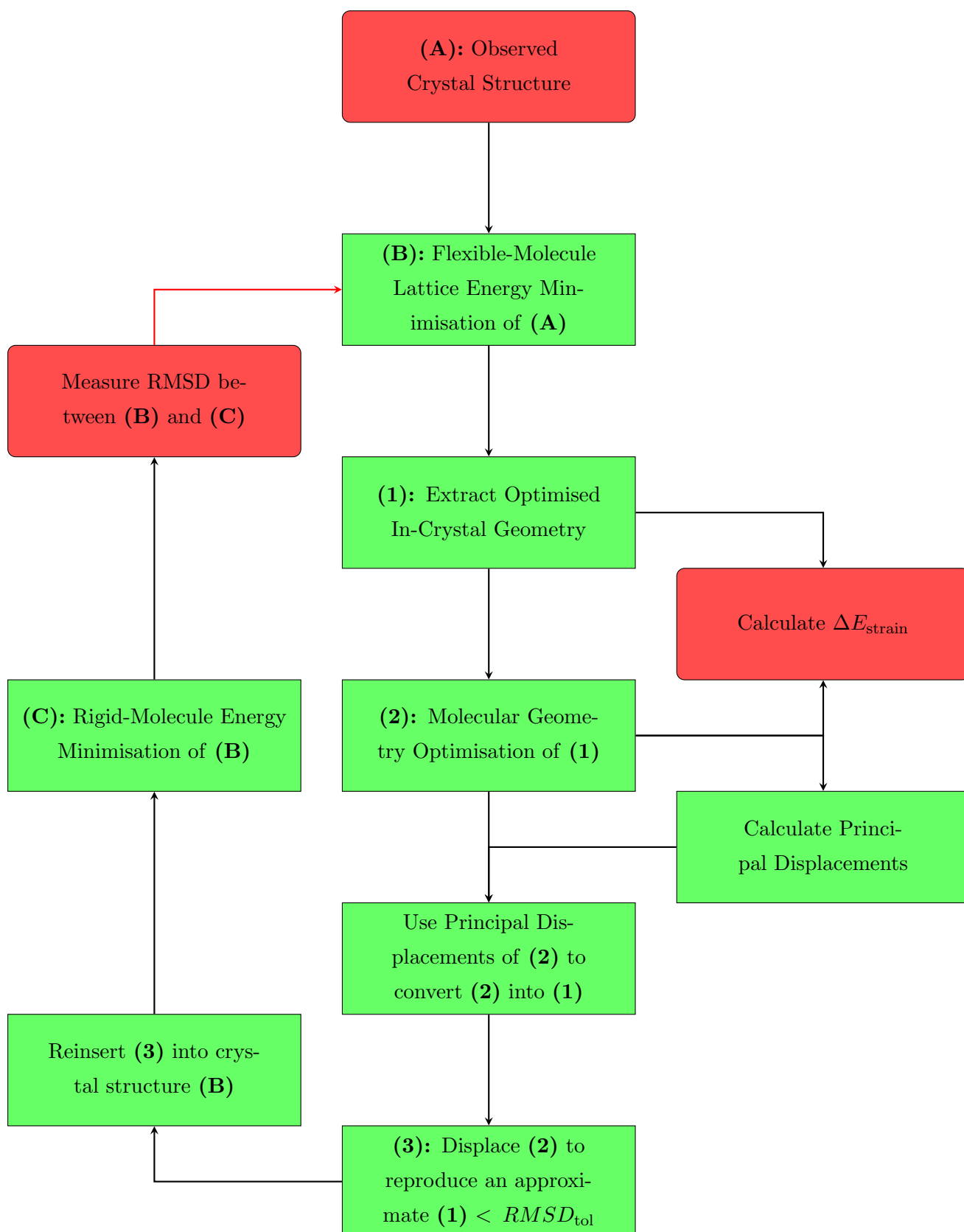


Figure 6.3: The work flow for measuring the maximum RMSD tolerance ($RMSD_{tol}$) required to approximate the in-crystal geometry to accurately reproduce (B). Crystal structures and molecular geometries are labelled with letters and numbers, respectively.

Molecule Reference	Atom Count	Rotatable Bonds	$\Delta E_{\text{strain}}^{\text{CRYSTAL}}$ kJ mol ⁻¹	$\Delta E_{\text{strain}}^{\text{CrysOpt}}$ kJ mol ⁻¹	$\Delta\Delta E_{\text{strain}}$ kJ mol ⁻¹
CELHIL01	28	8	18.99	13.07	5.92
DADNUR	30	8	16.63	9.16	7.47
CELHIL	28	8	14.41	8.93	5.48
DANQEP	26	8	13.14	10.18	2.96
DANQEP01	26	8	11.63	10.14	1.49
DANQEP02 _a	26	8	7.86	4.32	3.54
DANQEP02 _b	26	8	6.15	4.70	1.45
SEVJAF	30	7	7.11	3.52	3.59
GALCAX	20	7	3.77	1.55	2.22
GALCAX01	20	7	2.42	0.75	1.67
FIBKUW02	26	6	21.58	15.79	5.79
FIBKUW01	26	6	18.34	7.22	11.12
VEMTOW	22	6	6.18	3.47	2.71
VEMTOW01	22	6	3.52	3.59	-0.07
NEQNIG	22	6	2.77	1.57	1.20
FAHNOR05	20	5	9.20	4.53	4.67
FAHNOR	20	5	5.83	1.90	3.93
ODNPDS02	20	5	5.46	4.17	1.29
COCAIN	22	5	4.64	1.87	2.77
SIKRIN	26	4	14.65	5.95	8.70
MABZNA01 _b	20	4	3.98	1.05	2.93
MABZNA01 _a	20	4	3.96	2.20	1.76
MABZNA02	20	4	3.68	1.39	2.29
MABZNA _a	20	4	3.18	1.41	1.77
MABZNA _b	20	4	3.14	3.28	-0.14

Table 6.1: ΔE_{strain} values, in kJ mol⁻¹, for CRYSTAL09, $\Delta E_{\text{strain}}^{\text{CRYSTAL}}$, and CrystalOptimizer, $\Delta E_{\text{strain}}^{\text{CrysOpt}}$, DFT approaches for the 26 molecules. In addition, the number of atoms and rotatable bonds per molecule are also presented.

of rotational bonds and then by the ΔE_{strain} values reported by Thompson & Day ($\Delta E_{\text{strain}}^{\text{CRYSTAL}}$) within each group.

From both sets of ΔE_{strain} values, it is clear that there appears to be no correlation between the ΔE_{strain} value and the number of rotatable bonds or the atom count in the systems as extreme ΔE_{strain} values exist within each rotatable bond group.

Within the $\Delta E_{\text{strain}}^{\text{CRYSTAL}}$ set, the ΔE_{strain} values range from 2.42 kJ mol⁻¹ to 21.58 kJ mol⁻¹ which compares to the 0.75 kJ mol⁻¹ to 15.79 kJ mol⁻¹ for the $\Delta E_{\text{strain}}^{\text{CrysOpt}}$ set. This

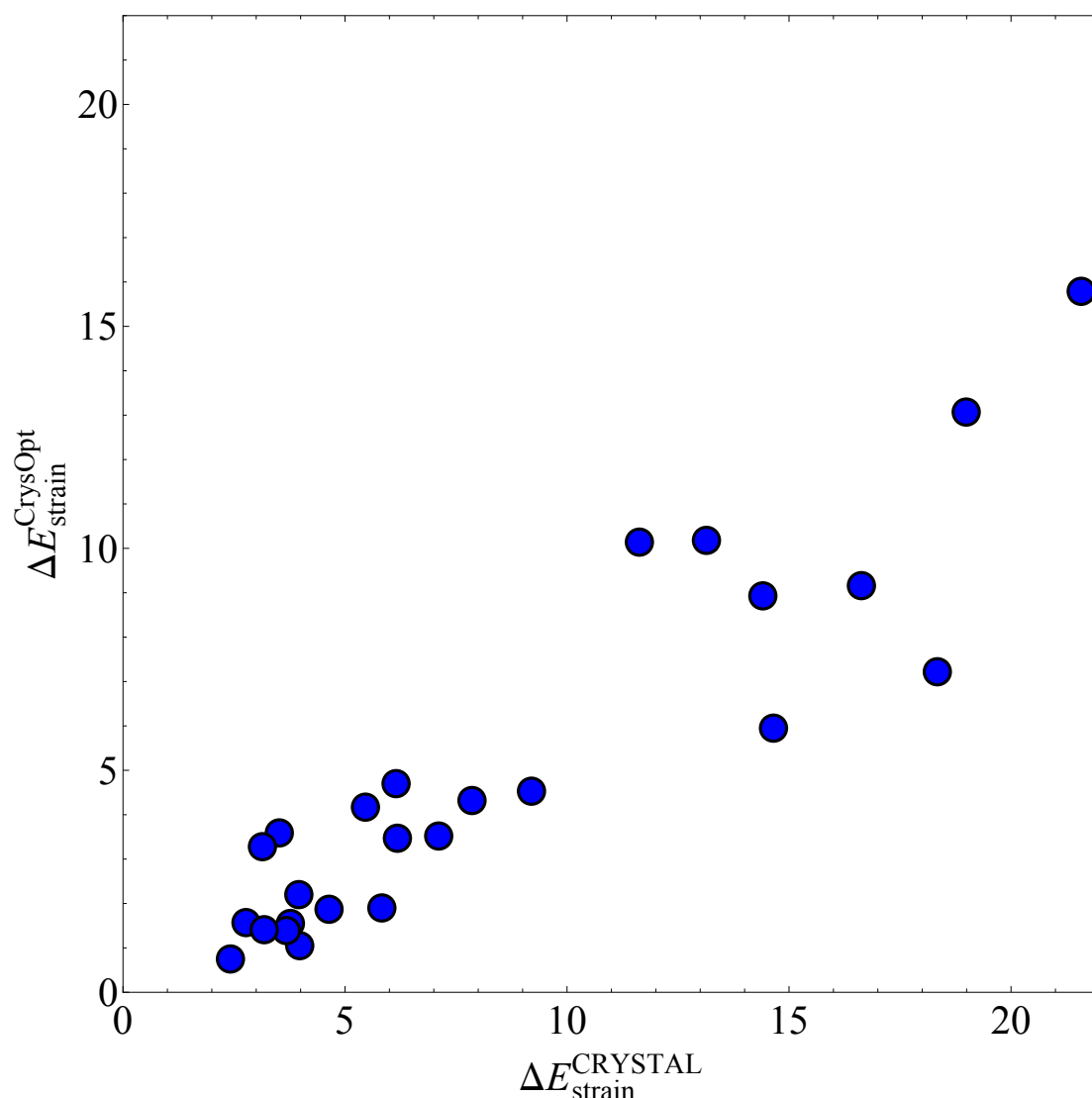


Figure 6.4: Comparison of ΔE_{strain} values, kJ mol^{-1} , calculated from the CRYSTAL09, $\Delta E_{\text{strain}}^{\text{CRYSTAL}}$, and CrystalOptimizer, $\Delta E_{\text{strain}}^{\text{CrysOpt}}$, software packages.

presents a slight difference in the range of ΔE_{strain} values but the molecules yielding these extreme values are identical for both sets.

The difference in the two sets arises from the difference in the methods used to calculate the ΔE_{strain} values between the two identical molecules. To aid in this comparison, a $\Delta\Delta E_{\text{strain}}$ has been calculated. With the exception of MABZNA_b and VEMTOW01, all of the $\Delta\Delta E_{\text{strain}}$ values are positive since the ΔE_{strain} values from the $\Delta E_{\text{strain}}^{\text{CRYSTAL}}$ set are always larger than the $\Delta E_{\text{strain}}^{\text{CrysOpt}}$ set. The cases of MABZNA_b and VEMTOW01 are not extreme exceptions to this general trend as they possess $\Delta\Delta E_{\text{strain}}$ values of -0.14 and $-0.07 \text{ kJ mol}^{-1}$, respectively.

Therefore the reduction in energy upon optimisation of the in-crystal geometry is greater when using CRYSTAL09 than CrystalOptimizer. However it is not clear whether

the crystal structures are less stable or the gas phase conformers are more stable for $\Delta E_{\text{strain}}^{\text{CRYSTAL}}$ over $\Delta E_{\text{strain}}^{\text{CrysOpt}}$ as either or both of these will affect the ΔE_{strain} . Nonetheless, CRYSTAL09 purely implements a DFT methodology whereas CrystalOptimizer uses a mixture of DFT and force-field approaches where both methods incur different errors which justifies the differences in ΔE_{strain} values.

The $\Delta\Delta E_{\text{strain}}$ values are generally less than approximately 5 kJ mol^{-1} . This is a sensible difference as this is within an acceptable tolerance for CSP calculations at present [156]. The $\Delta\Delta E_{\text{strain}}$ values for CELHIL01, DADNUR, SIKRIN and FIBKUW01 with the $\Delta\Delta E_{\text{strain}}$ values that exist at 5.92 kJ mol^{-1} , 7.47 kJ mol^{-1} , 8.70 kJ mol^{-1} and $11.12 \text{ kJ mol}^{-1}$, respectively, lie outside of this 5 kJ mol^{-1} tolerance. This highlights that the differences in the $\Delta\Delta E_{\text{strain}}$ values between the methodologies can be more extreme.

However, the comparison of these results shows that, for the majority of cases, there is little difference between the two DFT methods. However, it is a widely recognised that periodic DFT, CRYSTAL09, is more computationally expensive than, a standard DFT approach coupled with a force-field method, CrystalOptimizer, assuming the same level of DFT theory is implemented for both methods. Although one would expect the former to yield more accurate results, similar and comparable results can be obtained from the latter that is a computationally cheaper approach.

6.4.2 ΔE_{strain} Analysis

The ΔE_{strain} values for all 224 molecules are shown in Figure 6.5a and in descending order in Figure 6.5b.

The ΔE_{strain} values range from $0.001 \text{ kJ mol}^{-1}$ to $36.610 \text{ kJ mol}^{-1}$ and where geometrical differences in the former are imperceptible upon visualisation of the two geometries.

The shape of the distribution is one that decays exponentially with approximately half of the molecules possessing an $\Delta E_{\text{strain}} < 5.98 \text{ kJ mol}^{-1}$ and 48 molecules possessing an $\Delta E_{\text{strain}} < 1.0 \text{ kJ mol}^{-1}$. The mean ΔE_{strain} is 7.83 kJ mol^{-1} which is 1.85 kJ mol^{-1} larger than the median. This shows the extent of crystal packing forces to affect the energy of a molecular system which are relatively weak and do not energetically distort the molecule by a visually perceptible amount.

Further detail can be gleaned from the molecules in this set that form a non-covalent intramolecular interaction upon geometry optimisation. It is clear that these molecules possess ΔE_{strain} values that are higher than molecules that do not form any type of non-covalent intramolecular interaction.

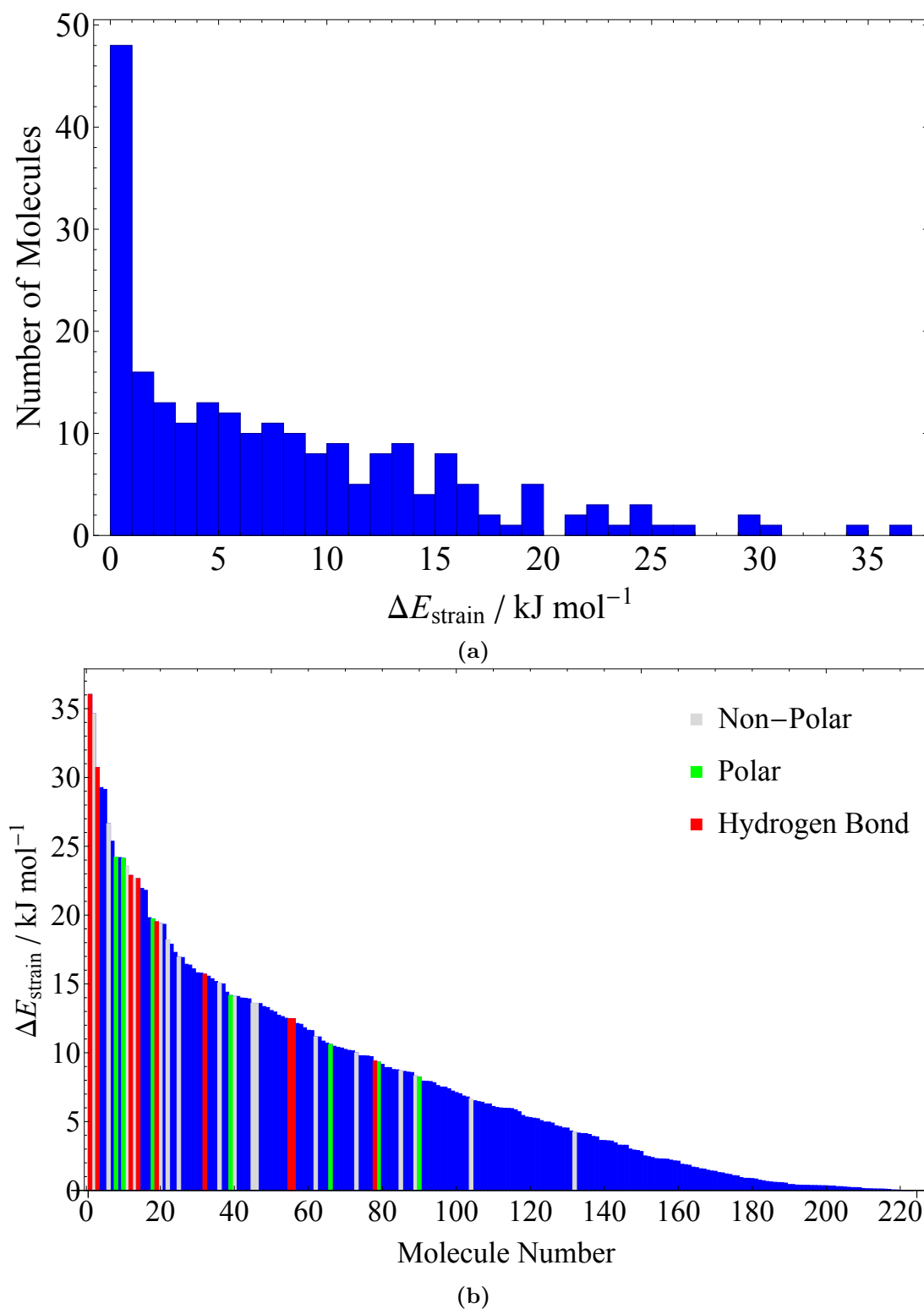


Figure 6.5: (a) a histogram of the ΔE_{strain} values separated by 1 kJ mol^{-1} bin widths. (b) distribution of ΔE_{strain} values, in kJ mol^{-1} , in descending order for the 224 molecule test set (molecules that form a non-covalent intramolecular interaction are highlighted in differing colours).

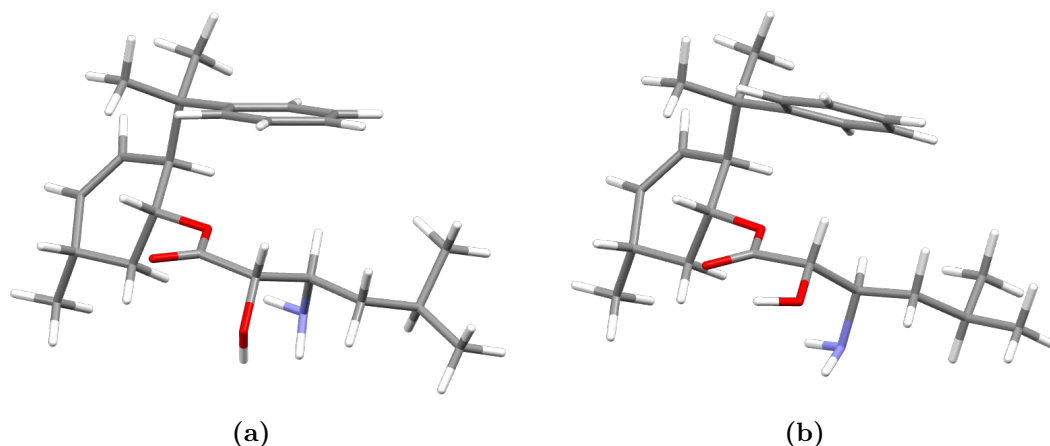


Figure 6.6: (a) and (b) shows the in-crystal and gas phase geometries, respectively, of the KIMSOO molecule. Note the twist of the amine and alcohol groups that allows 2 intramolecular hydrogen bonds to be formed.

The molecules that form a polar, non-polar or hydrogen bond possess ΔE_{strain} mean values of $15.70 \text{ kJ mol}^{-1}$, $15.73 \text{ kJ mol}^{-1}$ and $20.21 \text{ kJ mol}^{-1}$, respectively. These values are higher than the mean ΔE_{strain} for the set of molecules that do not form non-covalent intramolecular interactions upon the geometry optimisation of the in-crystal conformation. This result derives from the formation of non-covalent intramolecular interactions that lowers the energy of the gas phase conformer and hence increasing the ΔE_{strain} .

Whilst the mean polar and non-polar ΔE_{strain} values are approximately similar, the mean hydrogen bond value exists at an approximately 4 kJ mol^{-1} larger value. This is an expected, but interesting, result as hydrogen bonds are stronger forces that will result in additional stabilisation.

The molecule with the largest ΔE_{strain} value ($36.61 \text{ kJ mol}^{-1}$) is KIMSOO whose in-crystal and gas phase geometries are displayed in Figures 6.6a and 6.6b, respectively. This molecule illustrates an interesting point in that the 2 geometries do not appear to be too dissimilar upon first glance. Slight geometrical distortions exist that affect the benzene ring and also the isopropyl group at the terminus of the molecule. Whilst these geometrical distortions will contribute to the ΔE_{strain} , they are not significant enough to account for its entirety. In actual fact, it is the formation of 2 hydrogen bonds that give rise to the large ΔE_{strain} value. Upon geometry optimisation of the in-crystal geometry, the hydroxyl group twists around to form a hydrogen bond with the carbonyl group but, in addition, this allows the amine group to realign and form another hydrogen bond with the oxygen on the alcohol group. This molecule therefore forms 2 intramolecular hydrogen bonds that are required to be broken to yield the in-crystal geometry, which gives rise to the large ΔE_{strain} value.

6.4.3 Molecular Properties versus ΔE_{strain}

So far the ΔE_{strain} values have been analysed and it was proven that the formation of a non-covalent intramolecular interaction led to higher ΔE_{strain} values. An interesting question is whether an ΔE_{strain} value can be predicted prior to any calculations being performed based purely on the information contained within the in-crystal conformation. The most obvious place to start is by observing how the ΔE_{strain} varies with the number of atoms in the system, Figure 6.7a. The large cluster of data points when the number of atoms is < 20 in the figure are the 54 small organic molecules included in the 224 molecule test set. This shows that molecules of up to 20 atoms are generally more rigid and are not capable of forming non-covalent intramolecular interactions due to their lack of flexibility to bring donor and acceptor atoms into close enough contact with one another.

For molecules that possess more than 20 atoms, there is little correlation between the number of atoms and the ΔE_{strain} value. The molecules that form non-covalent intramolecular interactions generally exist at higher ΔE_{strain} values but are spread across the width of the graph. This plot shows that the number of atoms in a molecule cannot be used to predict the ΔE_{strain} value. This is due to the various functional groups that the molecule can possess. For instance, a benzene ring possesses 12 atoms but possesses an extremely small ΔE_{strain} because the molecule is rigid and is not easily distorted by crystal packing forces. Therefore some functional groups will hinder the size of the ΔE_{strain} value. Hence purely using the number of atoms to predict the ΔE_{strain} value is not viable. This also implies that the number of atoms is not predictive of molecular flexibility either.

Molecular flexibility can more accurately be defined by the number of rotatable bonds in the system and when plotted against the ΔE_{strain} values, a more prominent correlation arises, Figure 6.7b. It is now observed that the range of ΔE_{strain} values increase with the number of rotatable bonds in the molecule. The precise value of the ΔE_{strain} still remains unpredictable but it can now be observed what the maximum value for the ΔE_{strain} will be as a function of the number of rotatable bonds in the molecule. The line drawn on this figure attempts to predict this maximum ΔE_{strain} , $\Delta E_{\text{strain}}^{\text{max}}$, value for molecules that do not form any type of non-covalent intramolecular interactions. This straight line possesses the equation:

$$\Delta E_{\text{strain}}^{\text{max}} = (4 \text{ kJ mol}^{-1}) \cdot (\text{number of rotatable bonds}) + 2.5 \text{ kJ mol}^{-1}. \quad (6.1)$$

This line also bounds approximately 99% of all of the ΔE_{strain} values of the molecules. Only molecules that form non-covalent intramolecular interactions exist outside of this

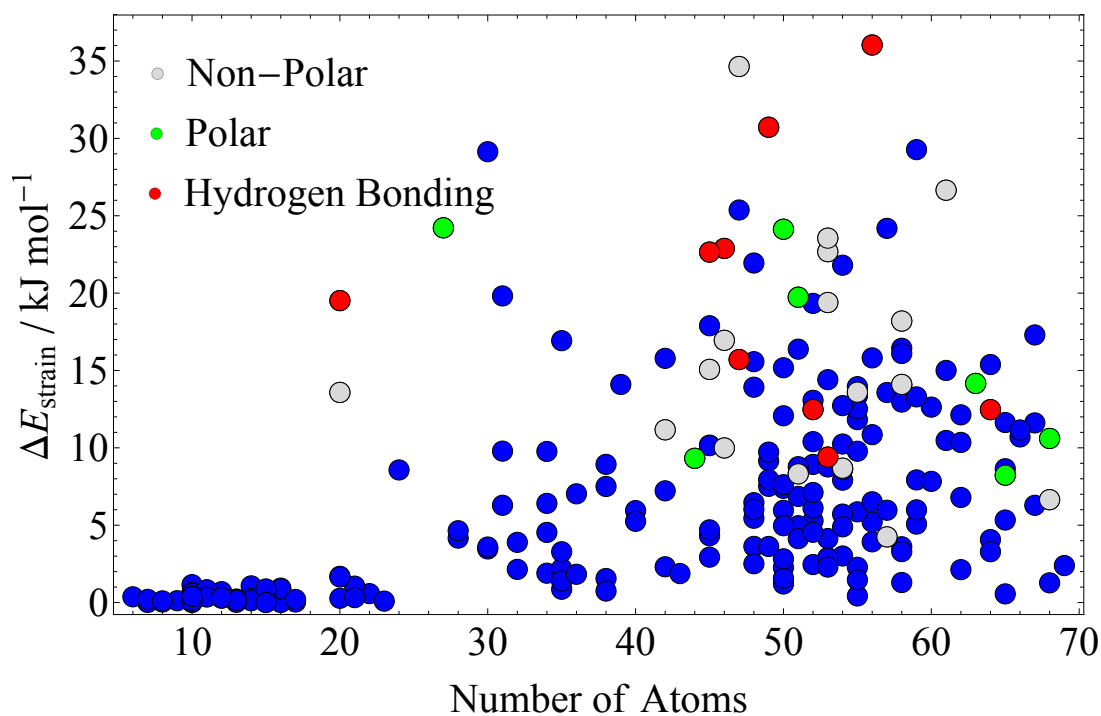
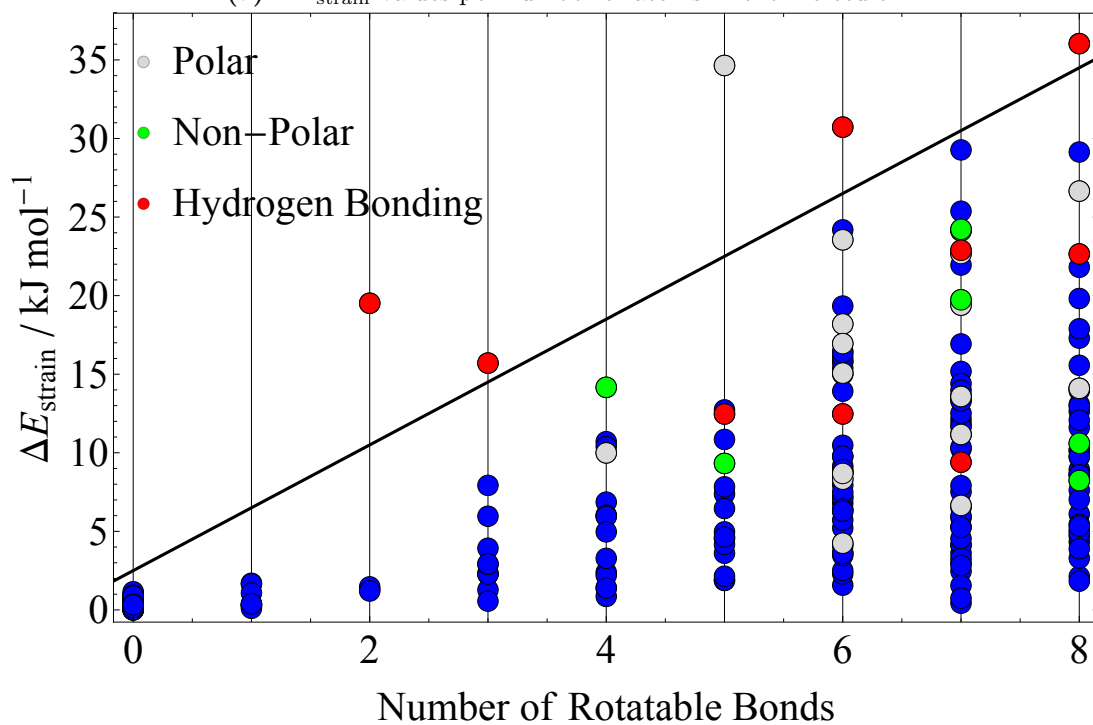
(a) ΔE_{strain} values per number of atoms in the molecule.(b) ΔE_{strain} values per number of rotatable bonds in the molecule. The line with equation $\Delta E_{\text{strain}}^{\text{max}} = (4 \text{ kJ mol}^{-1}) \cdot (\text{number of rotatable bonds}) + 2.5 \text{ kJ mol}^{-1}$ defines where approximately 99% of the ΔE_{strain} values lie for a given number of rotatable bonds.

Figure 6.7

region.

The assessment of whether a non-covalent intramolecular interaction can be formed is based on chemical intuition and therefore a valid assumption can be made as to whether the ΔE_{strain} could lie outside of this bound. The bound can hence be shifted depending on this outcome. Therefore it can be assumed that the level of flexibility of a molecule, specifically the number of rotatable bonds, is a more sensible measure of the ΔE_{strain} as opposed to the number of atoms in a given molecule. The upper bound of the ΔE_{strain} can be defined and increases with the number of rotatable bonds in the molecule but the location of where the value will lie remains ambiguous. The use of chemical intuition can afford an educated guess as to where the ΔE_{strain} could lie but a precise value cannot be targeted from these results.

6.4.4 RMSD versus ΔE_{strain}

The magnitude of the ΔE_{strain} value is comprised of either single or multiple variations in the molecular DOFs. However a large ΔE_{strain} does not necessarily imply that the RMSD between the in-crystal and gas phase geometries is also large.

In fact, a molecule could possess a relatively small ΔE_{strain} value but has a high RMSD value as the DOF(s) that are being displaced possess small force constants. The converse is also true; a molecule may only need to be slightly displaced along a DOF with a high force constant to yield a low RMSD but a high ΔE_{strain} .

However, to address this matter, Figure 6.8 must first be appreciated. This shows the distribution of the RMSD values between the in-crystal and gas phase geometries ranked by the number of rotatable bonds in the system.

There is a general increase in the range of RMSD values as more rotatable bonds in the molecule are present. This is because torsional angles possess lower force constants and hence are more easily distorted. The magnitude of this distortion will be larger than an analogous amount for bond angles and bond lengths.

A similar feature of this data exists as it did in the discussion pertaining to predicting ΔE_{strain} in Section 6.4.3, whereby the upper bound of the RMSD can be accurately predicted for a given rotatable bond. However, the value of the RMSD within this bound is more unpredictable although chemical intuition could lead to an educated guess as to where in the region the RMSD value may lie.

The presence of any non-covalent intramolecular interactions tends to lead to greater RMSD values as all of these molecules lie above the median value for a set of molecules

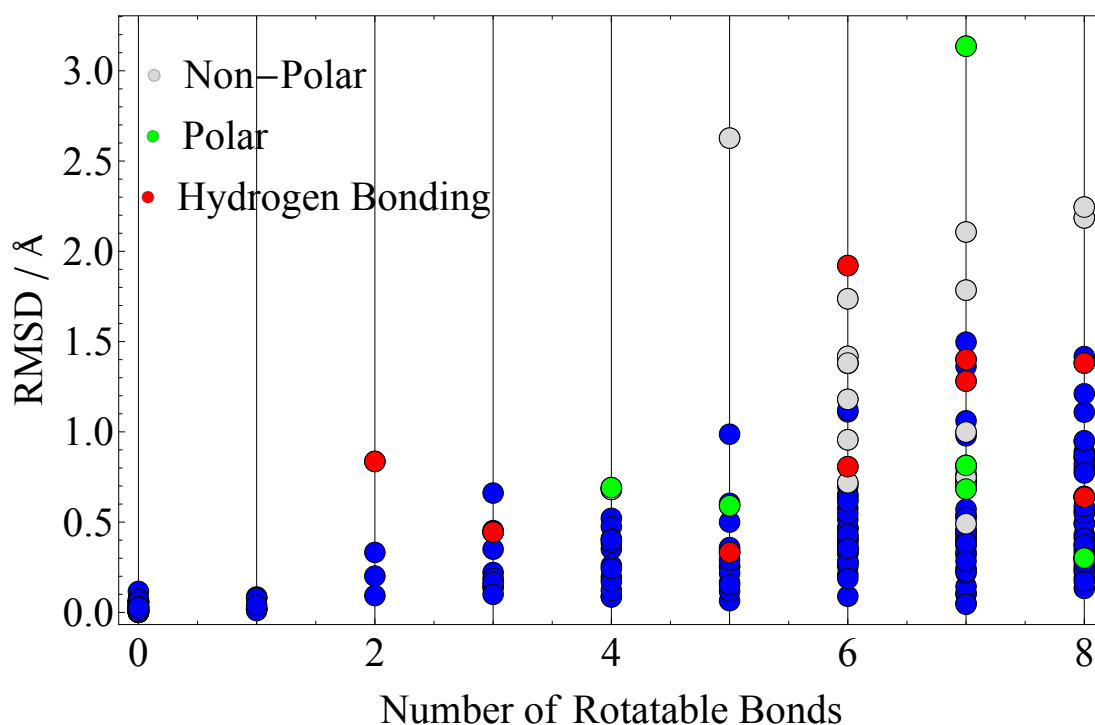


Figure 6.8: RMSD values between the gas and in-crystal geometries for the 224 molecule test set partitioned by the number of rotatable bonds in the molecule.

for a given number of rotatable bonds. The reason for this is due to the additional energy that is required to strain the molecule further away from its in-crystal geometry is offset by the energy released upon the formation of the non-covalent intramolecular interaction(s).

Nonetheless, it has now been established that both the potential magnitude of the ΔE_{strain} values, Section 6.4.3, and the RMSD values are proportional to the number of rotatable bonds in the molecule.

Figure 6.9 shows that a general correlation between these two statistics does exist. Generally, as the RMSD values increases so does the ΔE_{strain} which shows that the further the in-crystal geometry is away from the gas phase, the larger the ΔE_{strain} value. Hence the converse is also true. In addition, molecules that possess non-covalent intramolecular interactions occurred at larger ΔE_{strain} and RMSD values so naturally they occur at larger values in Figure 6.9.

Although this is the general case, there are exceptions. The first one is the aforementioned KIMSOO molecule that possesses the largest ΔE_{strain} value of $36.61 \text{ kJ mol}^{-1}$ but a relatively small RMSD value of 0.64 \AA . This is due to the additional stabilisation caused from the formation of 2 intramolecular hydrogen bonds.

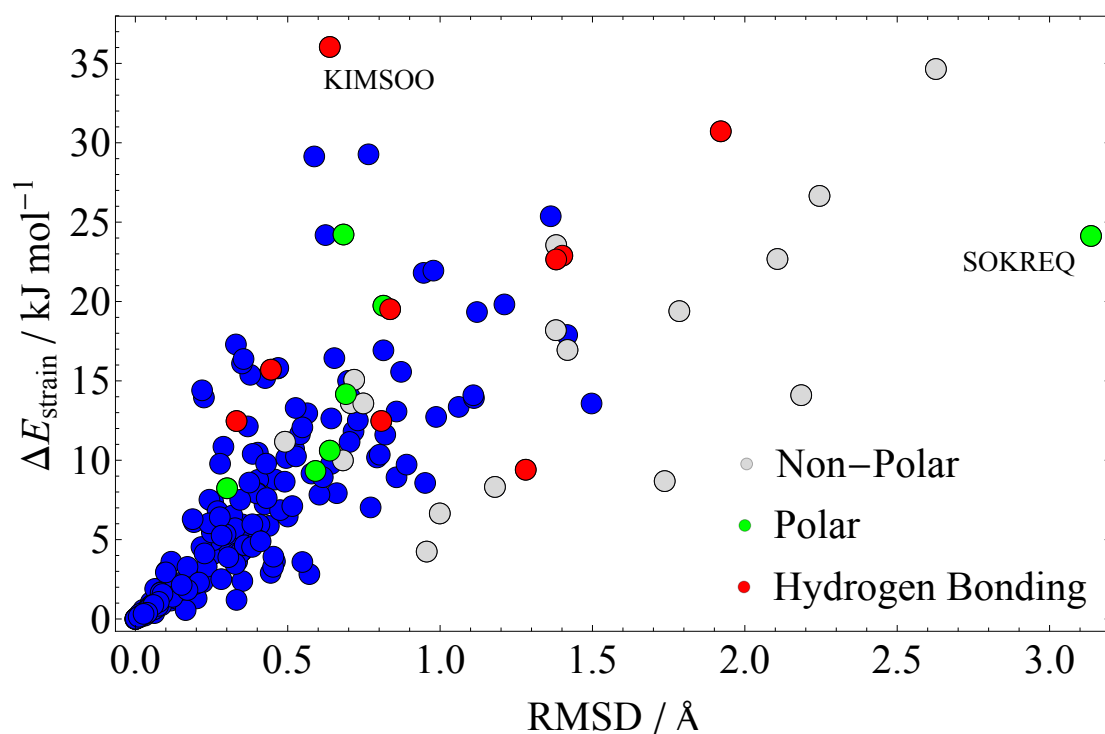


Figure 6.9: The ΔE_{strain} values as a function of the RMSD between the in-crystal and gas phase geometries for the 224 molecule test set. SOKREQ and KIMSOO are labelled as they possess the largest RMSD and ΔE_{strain} values, respectively.

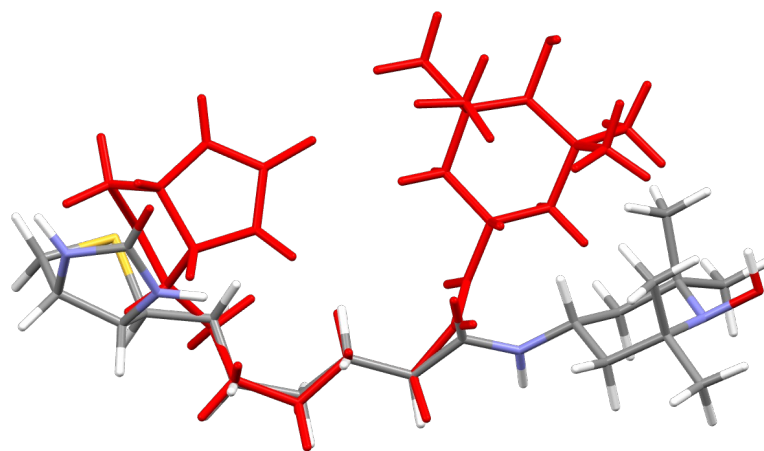


Figure 6.10: The in-crystal geometry (coloured by element) overlaid with that of the gas phase (red) geometry (RMSD = 3.136 Å), for the SOKREQ molecule.

Another case is the SOKREQ molecule. This molecule possesses the largest RMSD value, 3.136 Å, and a relatively large ΔE_{strain} value of 24.13 kJ mol^{-1} . The geometrical difference of SOKREQ is visualised in Figure 6.10. This molecule possesses a non-covalent polar intramolecular interaction between the two termini of the molecule whose rotatable bond force constants are low enough to allow this large geometrical distortion to take place.

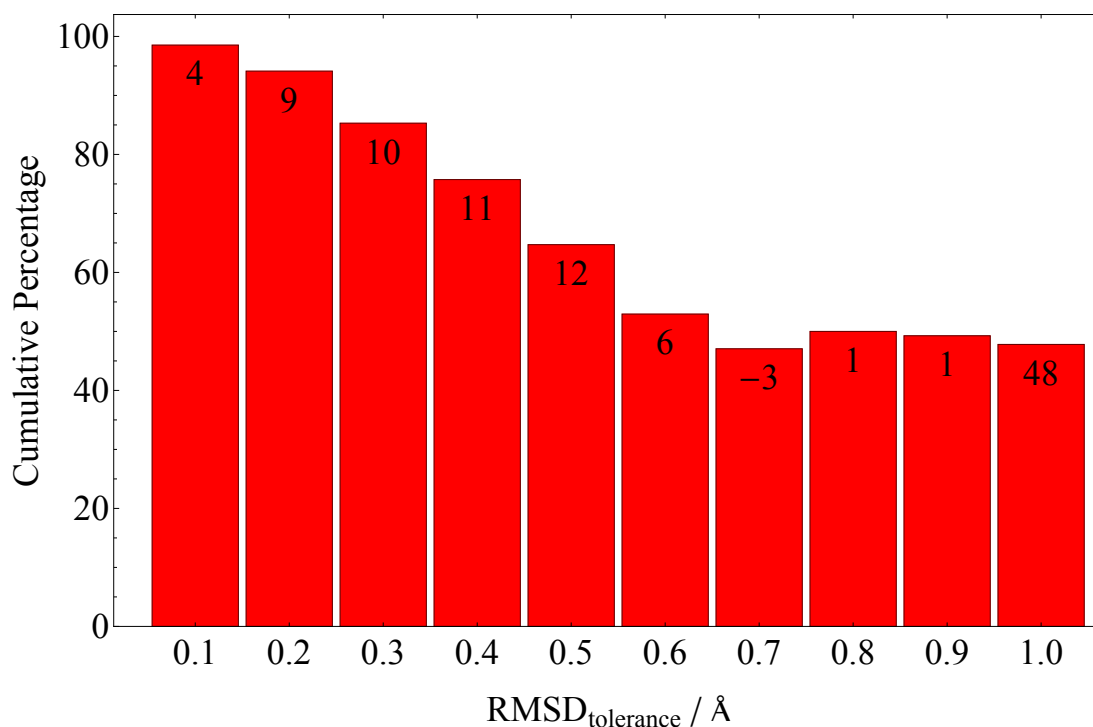


Figure 6.11: The number of approximated crystal structures that gave matches to their original counterparts at a given RMSD tolerance. The number at the top of each column shows the percentage change, rounded to the nearest whole number, as the RMSD tolerance is reduced to the columns value from the RMSD tolerance of the immediate right column.

6.4.5 Reproducing the Observed Crystal Structure

Using the procedure outlined in Figure 6.3, a test of how approximate the displaced gas phase conformer can be to the in-crystal conformation can commence. The accuracy of this approximation allows the minimal number of principal displacements to be used in the CSP methodology outlined in Chapter 4. The first and most pertinent piece of analysis purely relates to how many of the approximated crystal structures yielded matches to their original counterparts. This information is visualised in Figure 6.11.

Perhaps surprisingly, approximately 48% of crystal structures yield matches when the in-crystal geometry is approximated to an $RMSD_{tol}$ of only 1.0 Å. From here, the number of structural matches do not vary by a significant amount until the $RMSD_{tol}$ reaches 0.6 Å where a noticeable increase begins and proceeds until the $RMSD_{tol} = 0.1$ Å. This increase matches the expectation that a more accurate approximation to the actual in-crystal conformation increases the probability of yielding a match to the observed crystal structure.

It is important to realise that an $RMSD_{tol}$ of 0.1 Å matches all crystal structures but one, IXETIO. This is due to the closeness of the packing for this crystal structure. At

an $RMSD_{tol}$ of 0.1 Å, when the approximated in-crystal geometry is reinserted into the crystal, a chemically invalid structure is created; these structures possess overlap of the atoms in adjacent molecules. This is an anomalous case but it can be noted that using an $RMSD_{tol}$ of 0.01 Å does yield a match between the approximated and actual crystal structure. This confirms that this is not a flaw in the methodology, but an unusually sensitive case.

An encouraging result is that approximately 95% of crystal structures yield matches at an $RMSD_{tol}$ of 0.2 Å. Therefore this is a sensible value for the tolerance that is needed in the general case for the majority of crystal structures.

Another interesting feature of this data is that not all reductions in the $RMSD_{tol}$ values yield more crystal structure matches. The reduction in the $RMSD_{tol}$ from 0.8 Å to 0.7 Å observes a reduction in the number of crystal structure matches. Although this is only slight, it seems surprising that a tighter tolerance led to a less favourable result.

However, one must attempt to visualise the variation in the PES that exists for each approximated crystal structure. This approximated PES will steadily converge on the exact PES with each reduction in the $RMSD_{tol}$ value. Therefore there will be different minima on the surface of each PES as it is sensitive to changes in the molecular geometry. The reduction in structure matches shows that these minima can form such that the energy minimisation of the crystal structure is hindered. Thus the crystal structures converge to a different minimum that does not match the observed structure and does not exist for higher $RMSD_{tol}$ values. This result is a rare occurrence but nonetheless is something that should be realised when performing this method.

A more detailed analysis of the reductions in $RMSD_{tol}$ can now proceed. Figure 6.12 shows the distribution of the reductions in $RMSD_{30}$ values (between the crystal structures containing the exact and approximate in-crystal geometries) as the $RMSD_{tol}$ value is reduced by 0.1 Å. The $RMSD_{tol}$ value refers to the tolerance of the $RMSD_1$ value between the exact and in-crystal molecular geometries. The average $RMSD_{30}$ reduction when reducing to a given $RMSD_{tol}$ is highlighted in red.

The range of distributions steadily and consistently decrease as the $RMSD_{tol}$ is reduced. This demonstrates that the approximated in-crystal geometry is converging on the observed conformation. The distribution of each $RMSD_{tol}$ is generally uniform across its range with only minimal number of structures yielding slightly larger reductions.

As the distributions are densely populated, the average change in the RMSD error shows a shallow reduction from $RMSD_{tol}$ 1.0 Å to 0.6 Å where it slightly steepens until the lowest $RMSD_{tol}$, 0.1 Å. Nonetheless, the line traced by linking these average values together

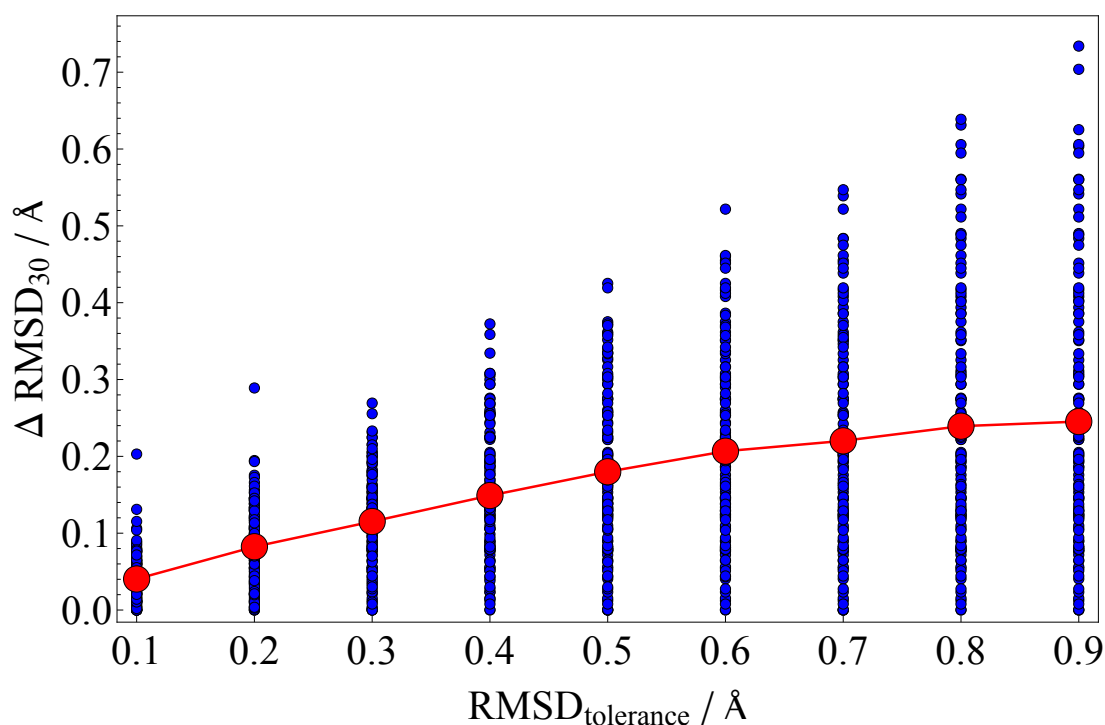


Figure 6.12: The distribution of the changes in the $RMSD_{30}$ values between the crystal structures containing the exact and approximate in-crystal geometries for each $RMSD_{tol}$ value. The average values for each distribution are highlighted in red.

is approximately straight. Therefore it can be afforded that reducing the $RMSD_{tol}$ by 0.1 \AA decreases the quantity that the RMSD is reduced will be approximately 0.05 \AA .

Since it has been observed that closer approximations to the observed in-crystal conformation yield more accurate results, it is now logical to quantify the effect of the distortions on the crystal structure, Figure 6.13.

There exists a clear correlation between these 2 factors in that a given $RMSD_1$ value led to a larger $RMSD_{30}$ value. This result is perhaps obvious as worse approximations to the in-crystal geometry propagate throughout the crystal structure which also yield a worse match between crystal structures. Therefore the crystal structure is more sensitive to variations in molecular geometry than the molecular geometry is to variations in the crystal structure.

The data itself appears ‘funnel-shaped’ in that the points diverge from a single point at $(0.037, 0.083)$ to become more spread out at larger RMSD values. When the $RMSD_1$ and $RMSD_{30}$ values reach approximately 0.25 \AA and 0.6 \AA , respectively, the data resides within 2 bounds. This shows that there is a limit to the extent that the molecular conformation can affect the quality of the crystal structure match. It was observed that variations in $RMSD_1$ values led to larger or smaller than proportional changes in $RMSD_{30}$ values.

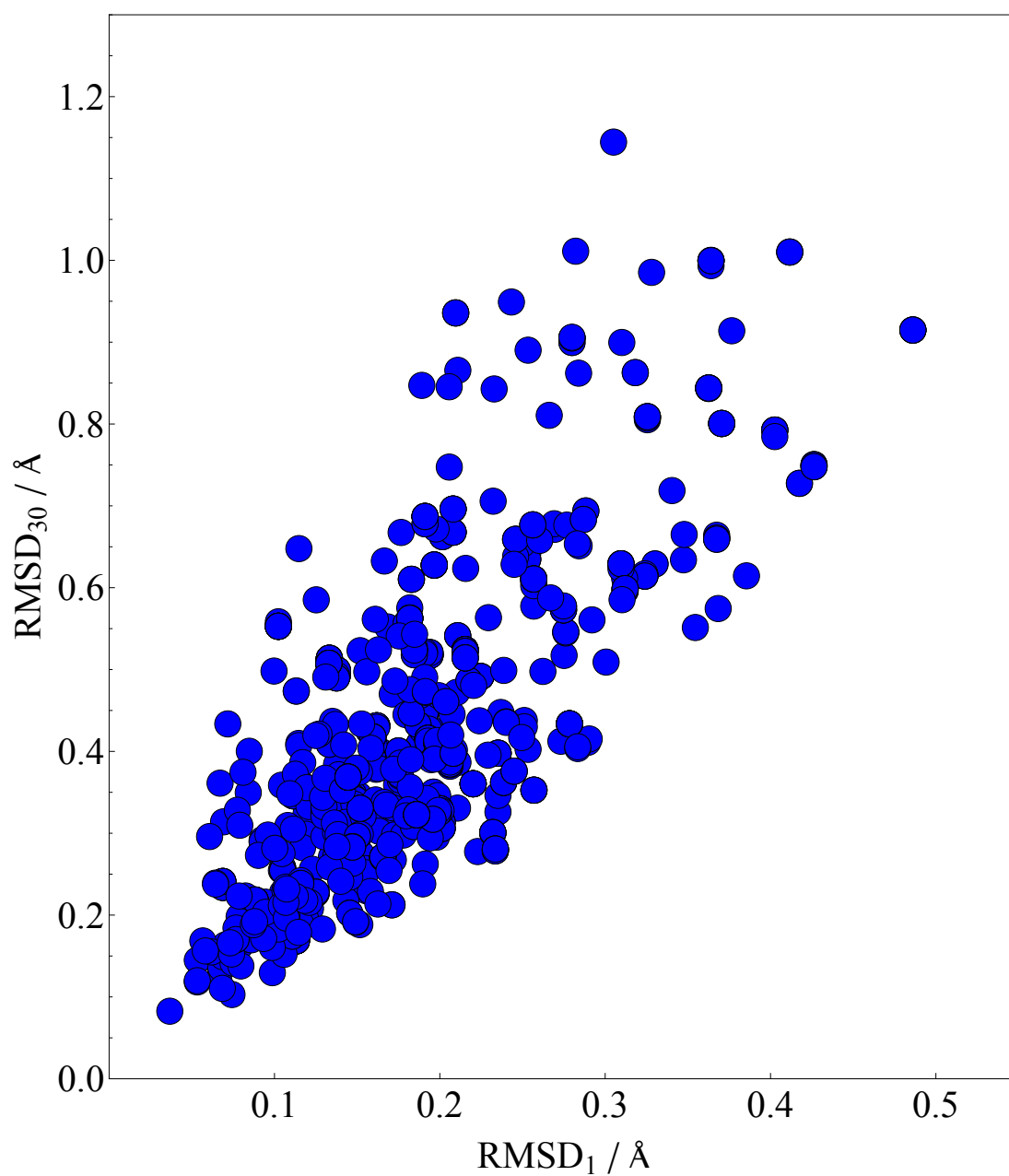


Figure 6.13: RMSD_1 values (between the exact and approximated in-crystal geometries) as a function of the RMSD_{30} values (between the exact and approximated crystal structures).

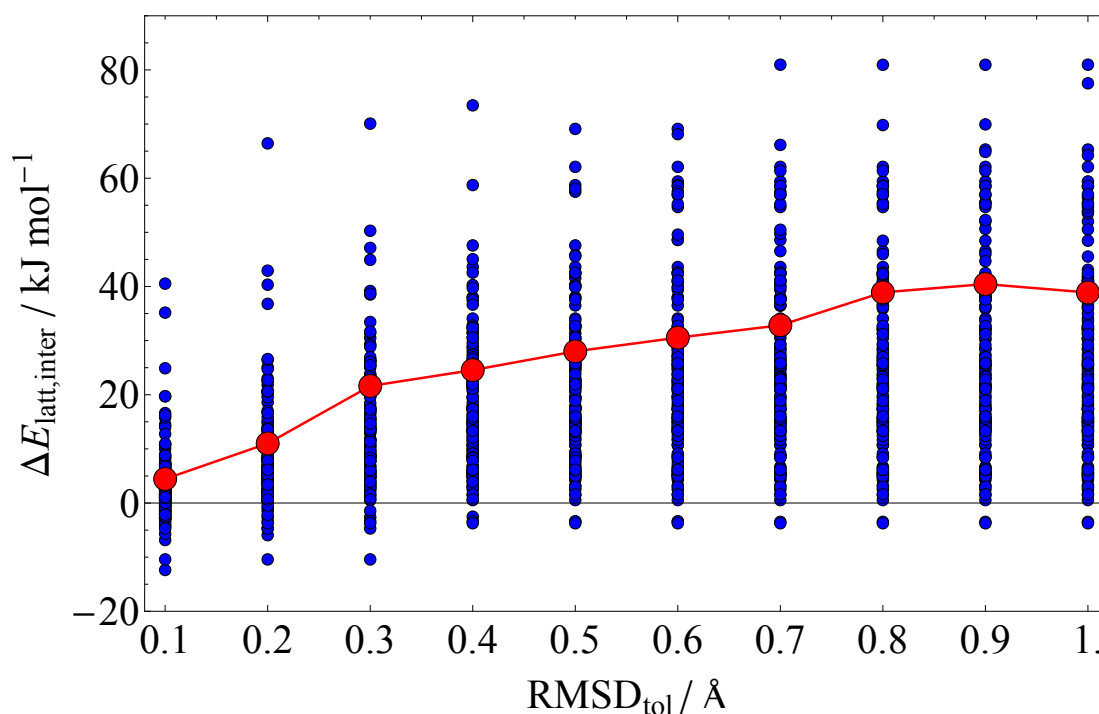


Figure 6.14: Distributions of the intermolecular component of the total lattice energy (relative to the exact lattice energy) for each $RMSD_{tol}$. The red points represent the average lattice energy for each distribution.

Another observation was that the $RMSD_{30}$ values always lie at higher values than its corresponding $RMSD_1$ value. This shows that the smaller changes in the molecular conformation led to larger variations in the crystal structure. This in turn highlights the importance of correctly modelling the internal molecular geometry as it will affect the packing of the molecules in the crystal.

Another interesting assessment that can be made is the observation of the convergence of the intermolecular component of the lattice energy to its exact value as the $RMSD_{tol}$ is reduced. Figure 6.14 shows the distribution of the differences in intermolecular energies between the exact and the approximated crystal structures.

Each distribution at each $RMSD_{tol}$ value possesses a broad range of $\Delta E_{lattice,inter}$ values that stretch over approximately 85 kJ mol^{-1} to 55 kJ mol^{-1} for $RMSD_{tol}$ values of 1.0 \AA and 0.1 \AA , respectively. This reduction in range for the distribution is expected as the accuracy to the approximation of the in-crystal geometry increases as the $RMSD_{tol}$ reduces and hence affords a $\Delta E_{lattice,inter}$ value that is closer to the exact value.

The distribution of $\Delta E_{lattice,inter}$ values over each $RMSD_{tol}$ is also reasonably uniform with the exception of the extreme values at the ends of the distribution. The average $\Delta E_{lattice,inter}$ begins at 38.9 kJ mol^{-1} and gradually reduces to 4.5 kJ mol^{-1} from the $RMSD_{tol}$ values of 1.0 \AA to 0.1 \AA , respectively.

Some molecules possess a negative $\Delta E_{\text{lattice,inter}}$ for which this number of molecules increases as the $RMSD_{\text{tol}}$ is reduced. This is due to the intermolecular energy of the crystal structure that contains the approximated in-crystal geometry being lower than the intermolecular energy of the crystal structure containing the exact in-crystal geometry. However, the molecular energy of the approximated in-crystal geometry in the former crystal structure still remains higher than the corresponding molecular energy of the latter structure. This therefore yields a total lattice energy that is greater for the approximated crystal structure than the exact crystal structure.

6.4.6 Force Constant Analysis

Having gained an understanding of the effect of ΔE_{strain} and RMSD on the crystal structure, the analysis of these two factors can be further decomposed by the principal displacements of a molecule.

Figure 6.15 shows the distribution of force constants for each set of principal displacements for each molecule involved in this study. These are plotted on a logarithmic scale as the values span 7 orders of magnitude. The molecules are separated by the number of rotatable bonds and each distribution contains a large point whose colour is determined by the type, if any, of the non-covalent intramolecular interaction that forms upon the geometry optimisation of the in-crystal geometry.

The position of this point in each force constant distribution shows the minimum force constant value of the principal displacement required to strain the gas phase conformer away from its equilibrium geometry and into an approximated in-crystal geometry such that its crystal structure yields a match to its observed counterpart. More specifically, these values relate to the discussion in Section 6.4.5 in that the majority of crystal structures are found at successive $RMSD_{\text{tol}}$ values.

Since all but 1 molecule yields a structural match at an $RMSD_{\text{tol}}$ of 0.1 Å, the force constant value is taken from the highest successive $RMSD_{\text{tol}}$ match. For instance, if a molecule provides a match at $RMSD_{\text{tol}}$ values of 0.1 Å, 0.2 Å, 0.3 Å, 0.4 Å, 0.5 Å and also 0.7 Å (not 0.6 Å, 0.8 Å, 0.9 Å or 1.0 Å), then the force constant will be taken from the 0.5 Å value; not at the 0.7 Å. The procedure is performed in this way because it is important to know that if a crystal structure match is yielded, it can be consistently reproduced and not just being hit by chance. This will increase the force constant value (the position of the coloured point) but gives an absolute maximum value required to reliably reproduce the observed crystal structure.

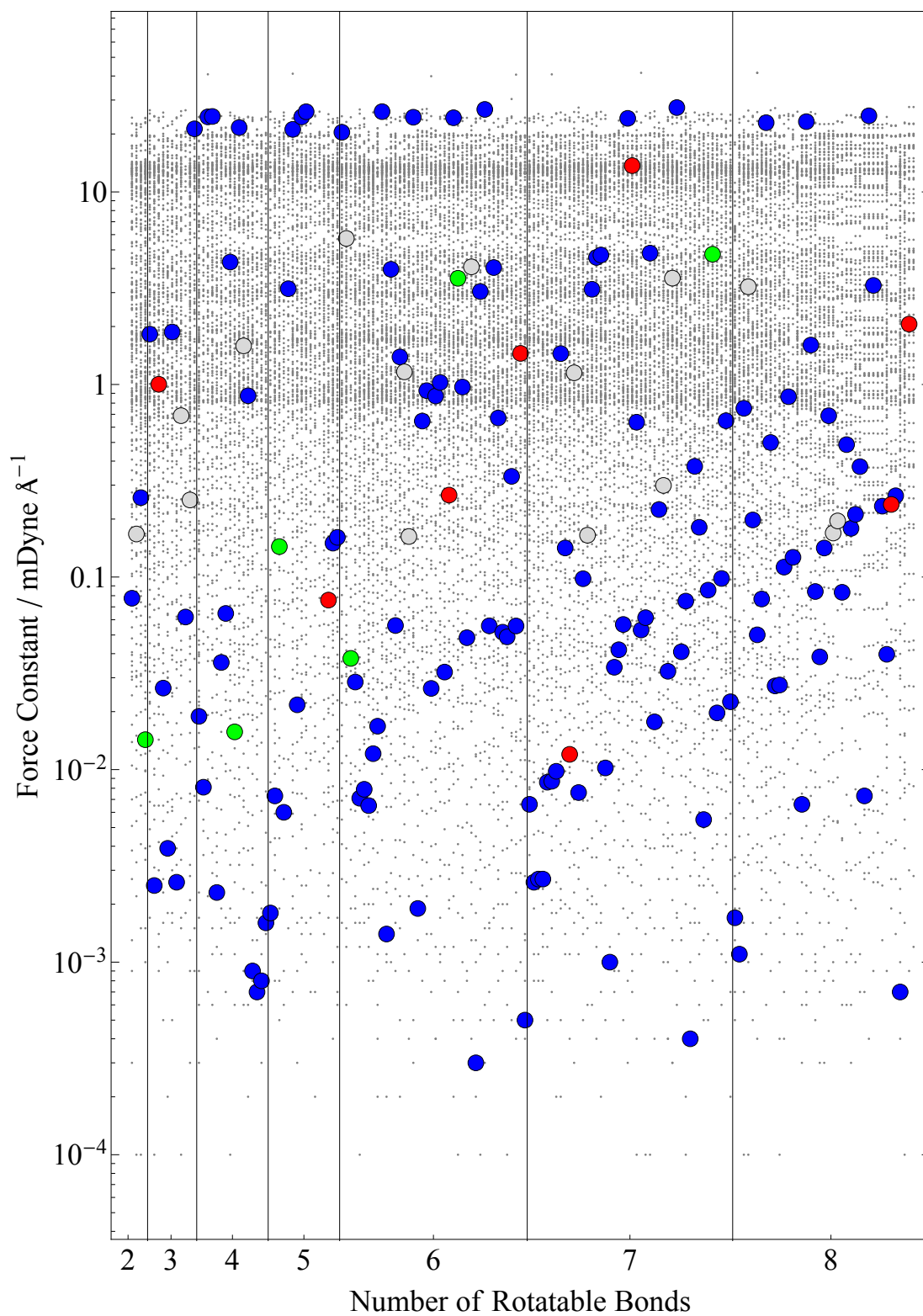


Figure 6.15: The distribution of force constants by molecule (dark grey) partitioned by the number of rotatable bonds. The larger, coloured points show the maximum force constant required to strain the molecule away from its gas phase geometry such that a match is yielded against the observed crystal structure. The blue, green, light grey and red points correspond to molecules that do not form, form polar, non-polar or hydrogen bonding non-covalent intramolecular interactions respectively.

Upon first observation of Figure 6.15, the distributions of the force constants do not vary with the number of rotatable bonds in the molecule. Instead the force constant values vary on a molecule by molecule basis. The distribution of the coloured points also appears almost random and there is no correlation between the number of rotatable bonds in the molecule and the force constant value. There is a correlation between the molecules that possess non-covalent intramolecular interactions in that they reside in higher regions of the distributions as they require stronger forces to break these bonds and strain the molecule away from its energetically stable geometry.

There are some interesting features of this data in that the force constants are sparser up to approximately $0.8 \text{ mDyne } \text{\AA}^{-1}$. From here, the values become more densely populated. This is due to these principal displacements including bond stretching motions that possess high force constants.

The overlap between the force constants whose principal displacements are dominated by torsion angle twists and bond angle bends is more ambiguous. There is no obvious region where a transition takes place although it could be argued that the force constants become more dense at $0.01 \text{ mDyne } \text{\AA}^{-1}$.

The complexity of this representation invites a more specific and simpler visualisation to the data. From the discussion in Section 6.4.5, it was shown that approximately 90% of crystal structures yielded matches to the corresponding crystal structure that possesses an approximated in-crystal conformer to within 0.2 \AA . Figure 6.16a shows the force constant required to reduce the RMSD to below 0.2 \AA against the ΔE_{strain} of the corresponding molecule.

There is an evident proportionality between the force constant required to perform this geometry interconversion and the ΔE_{strain} value. Larger ΔE_{strain} values are energetically strained further from the in-crystal geometry and hence required higher energy molecular motions to perform the conversion. These motions are only accessible when using principal displacements with larger force constants.

In addition, molecules that possess non-covalent intramolecular interactions occur at, not only higher ΔE_{strain} values, but also higher force constant values. Again, this is due to the motions required to break the non-covalent intramolecular interaction(s) that require larger force constants.

Figure 6.16b shows a similar relationship but replacing the ΔE_{strain} by the $RMSD_1$ values between the in-crystal and approximated in-crystal geometry. This shows another proportional relationship where the larger $RMSD_1$ requires a larger force constant value to perform the geometry interconversion.

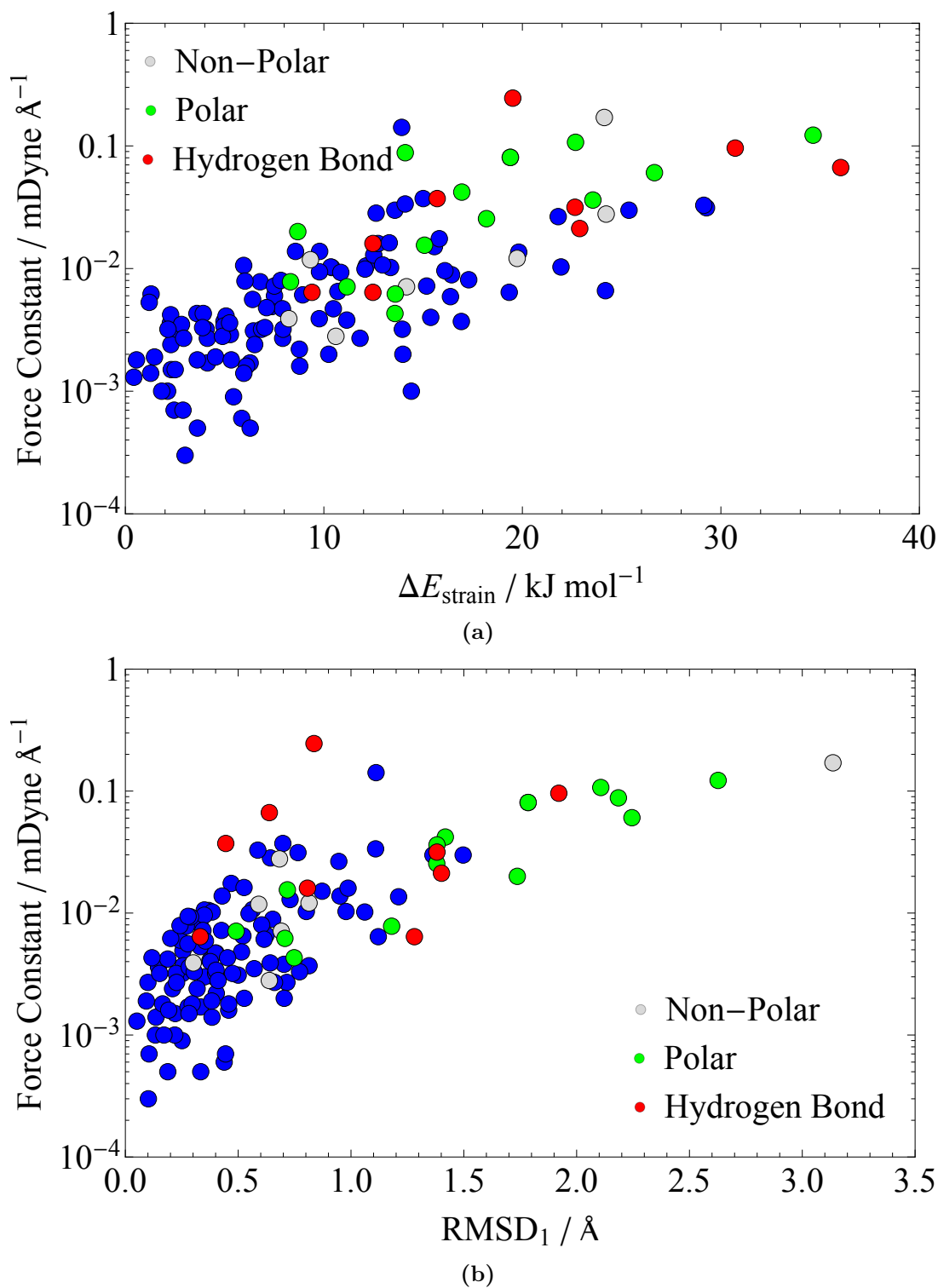


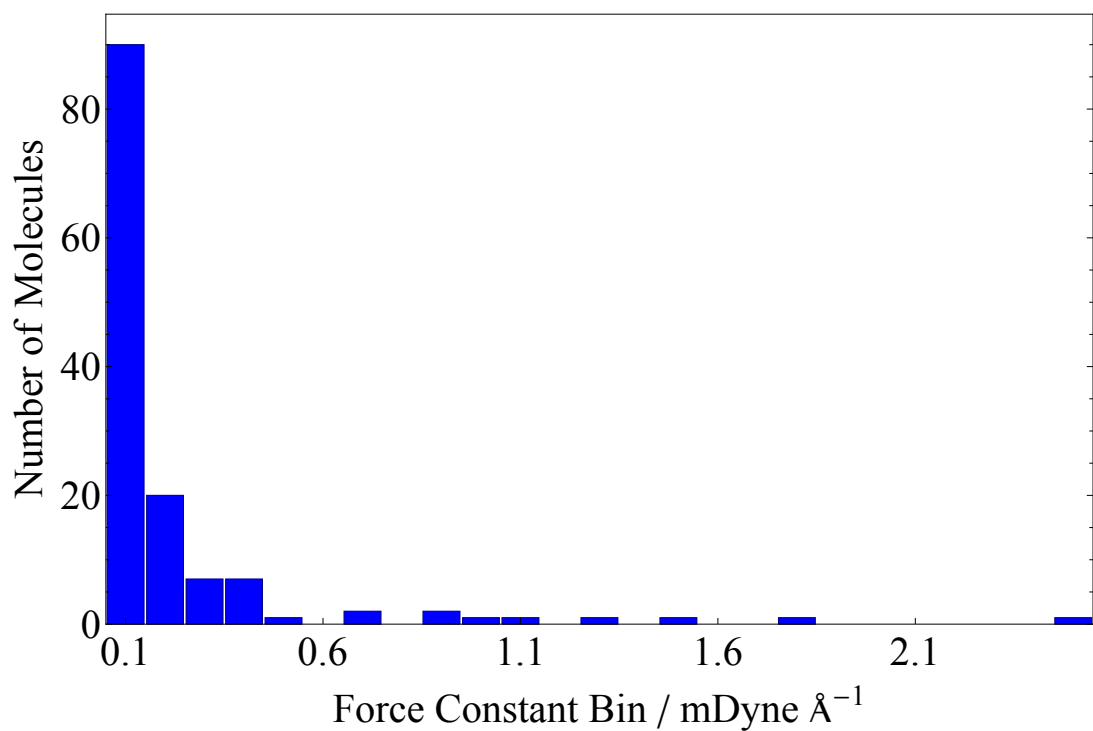
Figure 6.16: The force constants required to bring the RMSD between the in-crystal and approximated in-crystal geometry to within 0.2 Å against the ΔE_{strain} (a) and RMSD values (b). Both of these graphs are cropped at 1 mDyne Å⁻¹ to show the spread of the 99% of the force constant values.

However, in contrast to the Figure 6.16a, the distribution begins to decay and converge on a force constant value as the $RMSD_1$ value increases above 2.5 \AA . All of these molecules possess non-covalent intramolecular interactions that further distort the molecule away from the in-crystal geometry but it can be observed from Figure 6.16b that the force constant required to break these non-covalent polar and non-polar intramolecular interactions are approximately $0.01 \text{ mDyne \AA}^{-1}$ to $0.1 \text{ mDyne \AA}^{-1}$ respectively. This is further reinforced by the examples of the molecules that form intramolecular hydrogen bonds in that these values retain their proportionality as these bonds require larger force constants to break them. Therefore, if more molecules that form intramolecular hydrogen bonds were included in this study, it would be expected that they too would converge on a force constant value that is higher than what exists for the polar and non-polar non-covalent intramolecular interaction set.

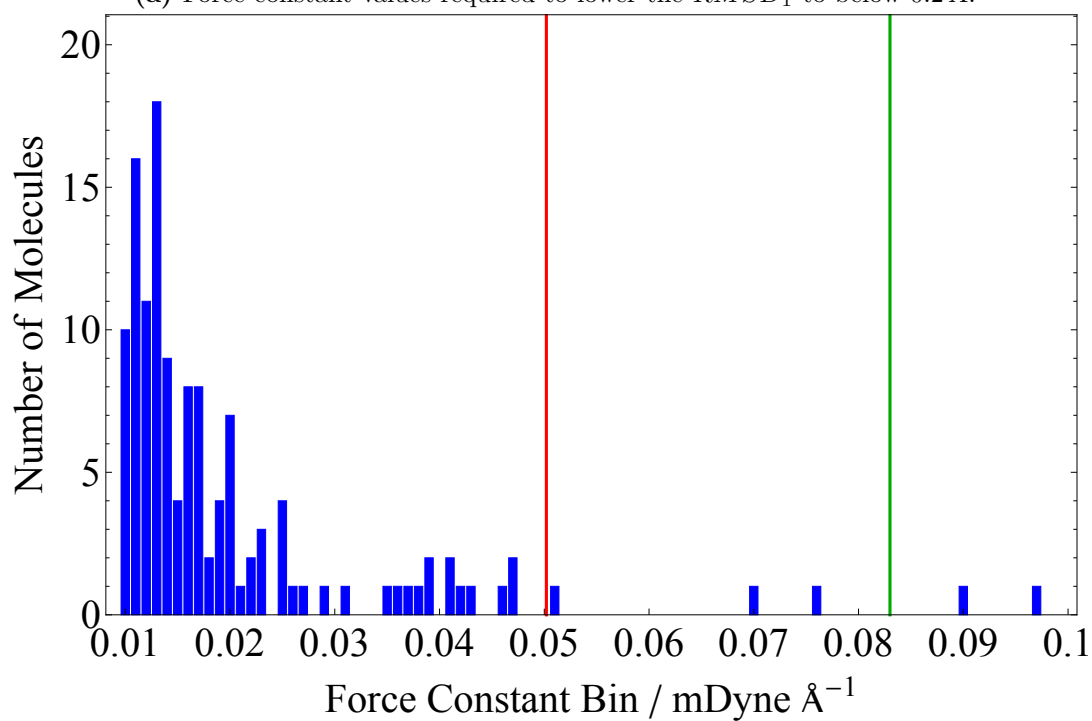
Figure 6.17a shows the molecular force constant value required to bring the exact and approximated in-crystal geometries to within the recommended RMSD of 0.2 \AA of one another. The range of these values are mostly populated in the $0.1 \text{ mDyne \AA}^{-1}$ bin but span to as large as $2.5 \text{ mDyne \AA}^{-1}$. The latter case is the exception of IXETIO that was also singled out as an anomaly in Section 6.4.5. This possesses an exceptionally high force constant that is required to perform the geometry interconversion. The principal displacement with this force constant possesses a specific angle bend that forces the 2 aromatic systems a greater distance away from one another.

In addition, it is observed that the force constant data spans 7 orders of magnitude. Therefore Figure 6.17b shows the most populated bin, $0.1 \text{ mDyne \AA}^{-1}$, in more detail between the values of $0.01 \text{ mDyne \AA}^{-1}$ to $0.1 \text{ mDyne \AA}^{-1}$. This, again, shows a distribution that rapidly decays but allows a finer detail of information to be extracted. The red and green lines show the first and second standard deviations for the whole set of force constants in Figure 6.17a (not Figure 6.17b which shows all values) which occurs at $0.050 \text{ mDyne \AA}^{-1}$ and $0.084 \text{ mDyne \AA}^{-1}$, respectively. This allows predictions of how large a force constant is required to perform the geometry interconversion with a degree of confidence. Therefore taking all of the force constants up to the value of $0.050 \text{ mDyne \AA}^{-1}$ and $0.084 \text{ mDyne \AA}^{-1}$ allows a 68.0% and 95.0% confidence that this set of force constants encompasses all of the necessary principal displacements to perform the geometry interconversion.

In relation to the methodology laid out in Chapter 4, it can then be assumed that including all principal displacements up to $0.084 \text{ mDyne \AA}^{-1}$ will be adequate for 95.0% of molecules.



(a) Force constant values required to lower the $RMSD_1$ to below 0.2 Å.



(b) A zoom of the 0.1 Å bin from Figure 6.17a. The first and second standard deviations of all force constant values are highlighted in red and green, respectively.

Figure 6.17

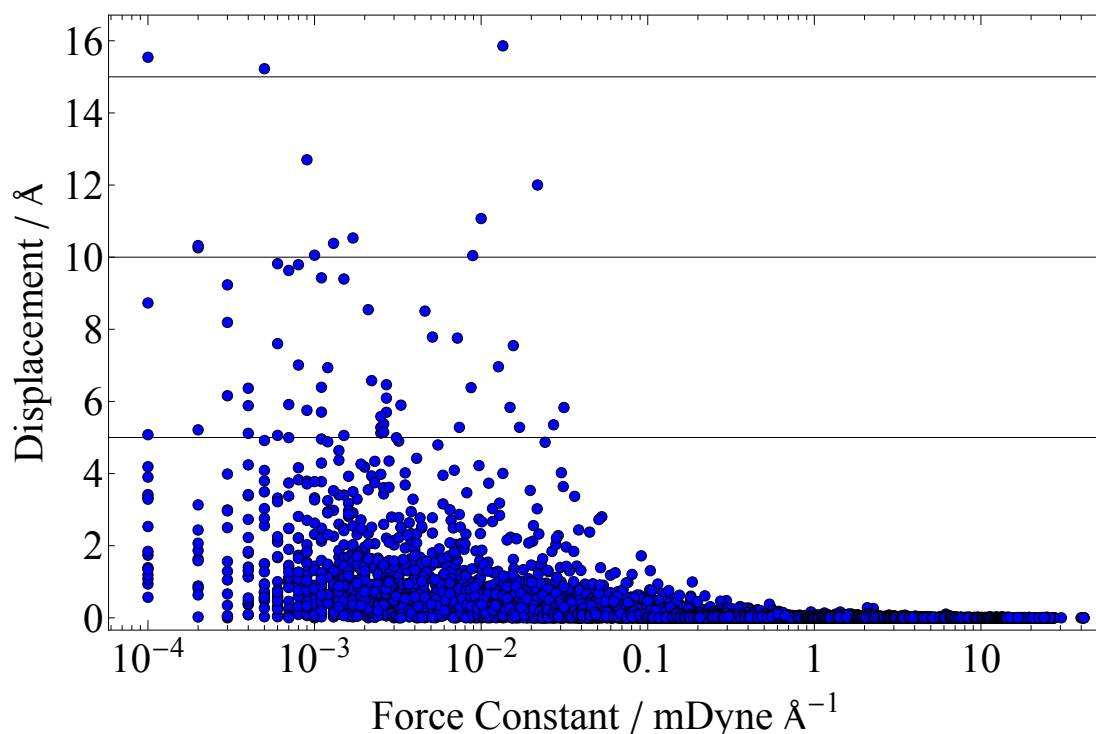


Figure 6.18: The extent of displacement for each principal displacement that possesses a given force constant value for each molecule in Sets 2 and 3.

6.4.7 Displacement Analysis

Additional analysis can be conducted on the extent of displacement for each principal displacement that possesses a given force constant value for each molecule. Figure 6.18 shows these displacement values as a function of the force constant for a given principal displacement for all molecules in the Sets 2 and 3.

The displacement values for each principal displacement span from approximately 0.0 Å to 16.0 Å. There is a strong trend that generally shows that the displacement value decreases as the force constant value increases. The force constants describe the energy required to distort the molecule along a corresponding principal displacement. Therefore higher valued force constants describe higher energy movements that often result in smaller displacement values which led to smaller reductions in the $RMSD_1$ values as they are tweaking bond length values and not performing torsion angle distortions.

The principal displacements that possess lower force constants have a greater range of displacements but these encompass molecular movements that can significantly reduce the $RMSD_1$ values. These therefore can incur larger displacements along these principal displacements.

Displacement values that are greater than 1.0 Å are generally unpredictable until the force constant value reaches approximately 0.1 mDyne Å⁻¹. After this 0.1 mDyne Å⁻¹

tolerance, the displacement values reduce to near-zero values that are negligible in reducing the $RMSD_1$ values. The size of the displacement does not necessarily guarantee a large reduction in the $RMSD_1$ value but since the individual atomic displacement values for each principal displacement are normalised, it is valid to assume that the extent of displacement along a given principal displacement is proportional to a reduction in the $RMSD_1$ value.

6.5 Conclusions

This chapter has presented a large amount of data and provides a significant insight into the decomposition of how the crystalline packing forces affect the molecular conformation.

An interesting feature was the comparison of CRYSTAL09 against CrystalOptimizer for calculating the ΔE_{strain} energies. It is generally observed that the former yields higher ΔE_{strain} values but it remains unclear whether this method predicts higher energy crystal structures of lower energy gas phase conformers. Nonetheless, the general ordering of the ΔE_{strain} energies remained constant which is an encouraging result.

The calculation of the ΔE_{strain} values yields a smooth distribution that ranges from $0.001 \text{ kJ mol}^{-1}$ to $36.610 \text{ kJ mol}^{-1}$. It is also observed that molecules that form a type of non-covalent intramolecular interaction give higher ΔE_{strain} values as these bonds are required to be broken to strain the molecule from the gas phase into the in-crystal geometry.

There is no correlation between the ΔE_{strain} values and the number of atoms in the molecule as the ΔE_{strain} is determined by the amount of flexibility in the molecule; not its size. Therefore there is a strong correlation between the number of rotatable bonds and the ΔE_{strain} energy. For molecules that do not form non-covalent intramolecular interactions, an upper bound for ΔE_{strain} can be calculated using the formula:

$$\Delta E_{\text{strain}}^{\text{max}} = (5 \text{ kJ mol}^{-1}) \cdot (\text{number of rotatable bonds}) + 2.75 \text{ kJ mol}^{-1}. \quad (6.2)$$

Again, the molecules that form non-covalent intramolecular interactions possess larger ΔE_{strain} values that can exceed this limit and are more unpredictable.

The number of rotatable bonds in the molecule is also correlated to the $RMSD_1$ value between the in-crystal and gas phase conformer. This value is also affected by the formation of non-covalent intramolecular interactions which led to large molecular distortions. This $RMSD_1$ value is also correlated to the ΔE_{strain} value where generally a larger

$RMSD_1$ value requires more energy to perform the geometry interconversion between the gas phase and in-crystal geometries. However, this is not always the case as some molecules possess small molecular distortions that afford a high energy penalty. The reverse scenario is a rarer occurrence where larger distortions do not yield low ΔE_{strain} values.

An approximate reproduction of the observed in-crystal geometry showed that an RMSD tolerance of 0.2 Å is required to reproduce approximately 90% of crystal structures. This is important as not all principal displacements are required to perform the geometry interconversion to an accurate standard.

Nonetheless, it was also shown that as the $RMSD_{\text{tol}}$ was lowered the approximated in-crystal geometry converges on the exact conformer. This is not a linear relationship but more of an exponential decay as the RMSD values reduce as the $RMSD_{\text{tol}}$ is decreased. It was also shown that the $RMSD_{30}$ values are affected by the smaller changes in the $RMSD_1$ values.

The distribution of the force constant for all molecules all occur in the same orders of magnitude and are not affected by the size, shape or flexibility of the molecule. However the value of the force constant that reduced the $RMSD_1$ value to below 0.2 Å increased with greater ΔE_{strain} values and molecular distortions. Upon visualisation of these force constants, over 95.0% of molecules only required force constant values up to 0.084 mDyne Å⁻¹ to yield an approximate in-crystal geometry that would lead to a match between this approximated and the observed crystal structure.

The quantity of the displacement is inversely proportional to the value of the force constant. This is due to the principal displacements that possess larger force constants performing higher energy molecular motions. Larger reductions in the $RMSD_1$ value occur within principal displacements with lower force constant values. These principal displacements are those that distort a molecule along the low energy DOFs.

The analysis in this chapter, coupled with the methodologies presented in Chapters 4 and 5, now completes the methodology outlined in Section 4.5. For clarity, an updated version of this methodology is outlined, thus:

1. Perform a geometry optimisation of the molecular conformation to obtain an equilibrium geometry.
2. Perform a principal displacement calculation on the equilibrium conformer.
3. Observe the number of rotatable bonds in the molecule and apply of Equation 6.1 to determine the energy bounds for each principal displacement

4. Apply the $0.084 \text{ mDyne } \text{\AA}^{-1}$ force constant limit to determine the number of principal displacements required.
5. Displace the molecule along each of these principal displacement individually and extract the quantity of displacement required to yield an energy that lies at the energy calculated from step 2 above the equilibrium geometry.
6. Generate a set of Sobol points between 0 and 1 and use each point to calculate a proportion of displacement between the bounds yielded from step 3.
7. Create a list of all possible combinations of the displacements from the Sobol numbers generated in step 6 and displace the molecule along its principal displacements by those quantities.
8. Calculate the energy of each of these displaced conformations and extract the point charges from the corresponding molecular conformations.
9. Fit a PES to a proportion of the energy data and another to a proportion of the point charge data.
10. Perform structure generation of each molecular conformation about the potential energy well where the equilibrium conformer resides.
11. Perform flexible-molecule energy minimisation of each crystal structure using CrystalOptimizer.
12. Perform clustering of the crystal structures yielded from step 11.
13. Obtain a final list of crystal structures and structurally compare these to the observed structure(s).

This work encompasses the majority of work for this thesis and now allows this formalised method to be applied in to CSP studies.

Chapter 7

Case Study I: Blind Test Molecule XXVI

7.1 Introduction to the Sixth Blind Test

The sixth blind test followed a similar approach to the previous blind tests (Section 2.8). Each participating group was permitted to submit up to 2 lists of 100 crystal structures for each target molecule and each list would be ranked by a scoring function (for example, lattice energy). Participants were also required to submit the number of central processing unit (CPU) hours required to yield each set of crystal structures in order to identify the computational expense of the methodologies and their respective accuracies.

A summary of the information on the sixth blind test that was sent to all participating groups will now be presented along with any initial thoughts on the CSP methodology that can be gleaned from simply observing a 2D sketch of each molecule.

This blind test was partitioned into 5 sections such that the major areas of CSP were tested (and whose target molecules are illustrated in Figure 7.1):

- **XXII:** The 2 individual cyclic systems could be modelled as rigid moieties as crystal packing forces are not strong enough to distort to geometries by a significant. However, the torsion angles that join the 5- and 6-membered rings together could cause a puckered (rather than planar) conformation and hence are the only parts of these aromatic systems that need to be modelled as flexible. The nitrile groups can also be assumed to be rigid. However, it could be argued that the bond angles that cause the nitrile groups to exist out of the molecular plane need to be modelled as flexible as only small forces applied to the terminal nitrogen atoms in

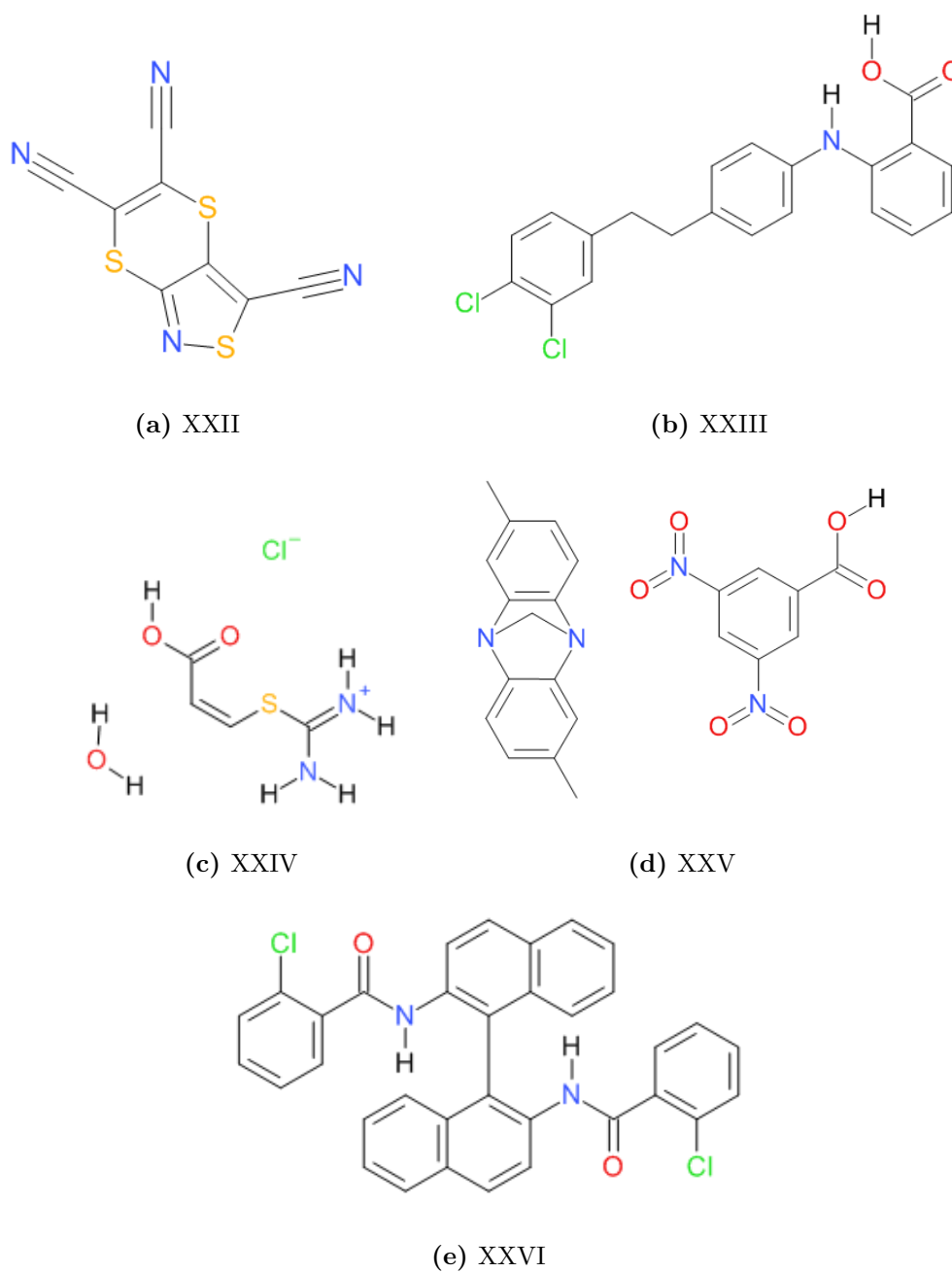


Figure 7.1: The target molecules for the sixth blind test referred to as XXII through XXVI.

the nitrile group will result in a large moment. This could therefore deform the molecule and affect the electrostatic interactions in the crystal structure.

- **XXIII:** This is a partially flexible molecule that possesses 5 known polymorphs in which 3 of these polymorphs are $Z'=1$ and 2 polymorphs were $Z' > 1$. A maximum of 7 torsional angles will need to be included (the 3 bridging C-C dihedrals between the di-chloro and central phenyl groups, the 2 torsional angles either side of the amine group, and the 3 torsional angles associated with the carboxyl group namely H-O-C-C, O=C-O-H and O=C-C-C). One could argue that the carboxyl hydrogen atom could exist on either oxygen atom and therefore 2 molecular conformations for every generated crystal structure should be taken into account. In addition, the hydrogen existing on the amine group forms an intramolecular hydrogen bond with the oxygen of the carboxyl group.
- **XXIV:** This system is a salt and adds another layer of complexity to the calculations as the ionic nature of the chloride and imine functional groups must be modelled effectively. With reference to molecular flexibility, the molecule is largely rigid due to the aromatic character that propagates the whole molecule.
- **XXV:** This system is a co-crystal where both components can be modelled as rigid entities. Unlike molecule XXIII, ambiguity exists about the position of the hydrogen on the carboxyl group.
- **XXVI:** This molecule possesses the greatest number of atoms of any molecule included in this blind test and also possesses 7 torsional angles that need to be modelled to effectively incorporate molecular flexibility into the CSP process. It is commonly known that generally the amide groups exist in a *trans* configuration [151] and therefore denotes a good starting point for these 2 dihedral angles.

Note that information on the performance of the Day Group in the sixth blind test can be found elsewhere [65].

7.2 Author's Contribution to the Sixth Blind Test

The author acknowledges Dr Angeles Pulido for performing the initial conformational search on molecule XXVI.

The author focussed on the CSP study of molecule XXVI. This was due to the flexibility and size of the molecular system.

The CSP study of this molecule was initiated by a conformer search on the target molecule XXVI using an LMCS method [47, 48] within MacroModel [46] where the starting molecular geometry is optimised before being perturbed along random combinations of its calculated normal modes. This implemented the Optimized Potential for Liquid Simulations (OPLS)2005 force-field [49, 50] with updated torsion parameters [157] and was terminated after 50,000 conformers were generated using the maximum and minimum movement distances of 3 Å and 6 Å, respectively. Each conformer that was generated was minimised using the Polak-Ribiere conjugate gradient [158] which was considered converged when the gradients were below 0.05 kJ/mol/Å. Duplicate molecular geometries were removed if the $RMSD_1$ values between any 2 conformers was below 0.02 Å. This yielded a total of 96 unique molecular conformers.

The next stage was a DFT-D (B3LYP-GD3BJ/6-311G(d,p)) geometry optimisation of each of the 96 conformers, using GAUSSIAN09 [67], where the resulting geometries were clustered to remove duplicates using in-house software if the $RMSD_1$ value between any 2 conformers was below 0.5 Å. Of these optimisations, a total of 40 molecular conformers exhibit relative energies, $E_{\text{gas}}^{\text{rel}} < 30 \text{ kJ mol}^{-1}$. The 41st lowest energy conformer was also included, despite its energy falling above the 30 kJ mol⁻¹ cutoff. This was because this conformer was the lowest in energy in which both amide groups were found in a *cis*-conformation, which could offer a good opportunity for intermolecular hydrogen bonding.

A set of distributed multipoles were subsequently generated for each of these 40 unique conformer using a B3LYP/6-311G(d,p) level of theory. The XXVI conformational search and rank 4 multipole generation accounted for 3,745 CPU-hours and is broken down as 5 CPU-hours for the initial OPLS2005 conformational search, 3,709 CPU-hours for the DFT geometry optimisations and the rank 4 multipole generation accounting for 31 CPU-hours.

Owing to the time constraints of the blind test and the flexible molecule CSP methodology outlined in Chapter 4 still under partial development, a rigid molecule crystal structure generation was conducted on the 41 conformers. An initial 2,000 trial crystal structures were generated for each conformer in each of the 6 most common space groups ($P\bar{1}$, $P2_1$, $P2_1/c$, $C2/c$, $P2_12_12_1$, $Pbca$) (12,000 trial crystal structures per conformer) and lattice energy minimised using DMACRYS. An additional 2,000 crystal structures were generated for any combination of conformer and space group that yielded crystal structures within 15 kJ mol⁻¹ of the global lattice energy minimum (relative conformer energy + intermolecular energy). A further 500 additional crystal structures were subsequently generated for any conformers that gave crystal structures within 15 kJ mol⁻¹ of the global lattice energy minimum in each of the next 18 most common space groups

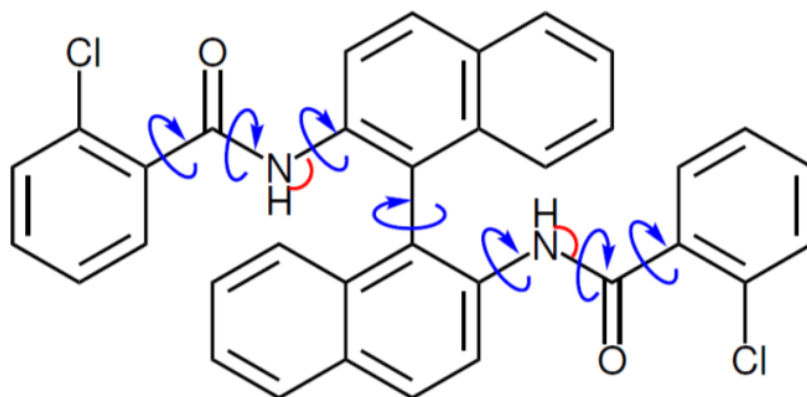


Figure 7.2: Molecule XXVI annotated by the dihedral (blue) and bond (red) angles that are modelled as flexible by CrystalOptimizer.

(*P1*, *C2*, *P1c*, *C1c*, *P2/c*, *P2₁2₁2*, *Pca2₁*, *Pna2₁*, *Fddd2*, *Pccn*, *Pbcn*, *P4₁*, *P4₃*, *I4₁/a*, *P4₁2₁2₁*, *P4₃2₁2₁*, *P3₁*, *P3₂*). All of these additional crystal structures were also lattice energy minimised using DMACRYS. Of these minimised crystal structures, those within 30 kJ mol⁻¹ of the global lattice energy minimum were clustered within each space group using in-house code. This structure generation coupled with the rigid molecule lattice energy minimisation process took 150,419 CPU-hours.

The 397 crystal structures that were within 25 kJ mol⁻¹ of the global lattice energy minimum were re-optimised with CrystalOptimizer to allow for molecular flexibility. This employed the DFT-D (B3LYP-GD3BJ/6-311G(d,p)) level of theory for the intermolecular electrostatic interactions and a PBEPBE-GD3BJ/6-31G(d,p) level of theory for the intramolecular energy model. The latter (which is a lower level of theory) was used for the intramolecular energy model to limit the computational expense due to the size of the molecule and requirement for Hessian calculations in CrystalOptimizer. The flexible DOFs allowed in CrystalOptimizer are illustrated in Figure 7.2 and whose calculations took 22,665 CPU-hours.

In addition, all of the final crystal structures from CrystalOptimizer were re-minimised using the rigid molecule approximation where a PCM was used to generate the multipoles. Two sets of predictions were generated, one with the dielectric constant in the PCM calculation set to 3.0 ϵ_0 and 7.0 ϵ_0 , with ϵ_0 denoting the permittivity of free space. In this case, the higher dielectric constant cannot be justified as a realistic dielectric constant in the crystal, but allows us to examine the dependence of the predictions on the degree of polarisation of the atomic multipoles.

7.3 Results for Molecule XXVI

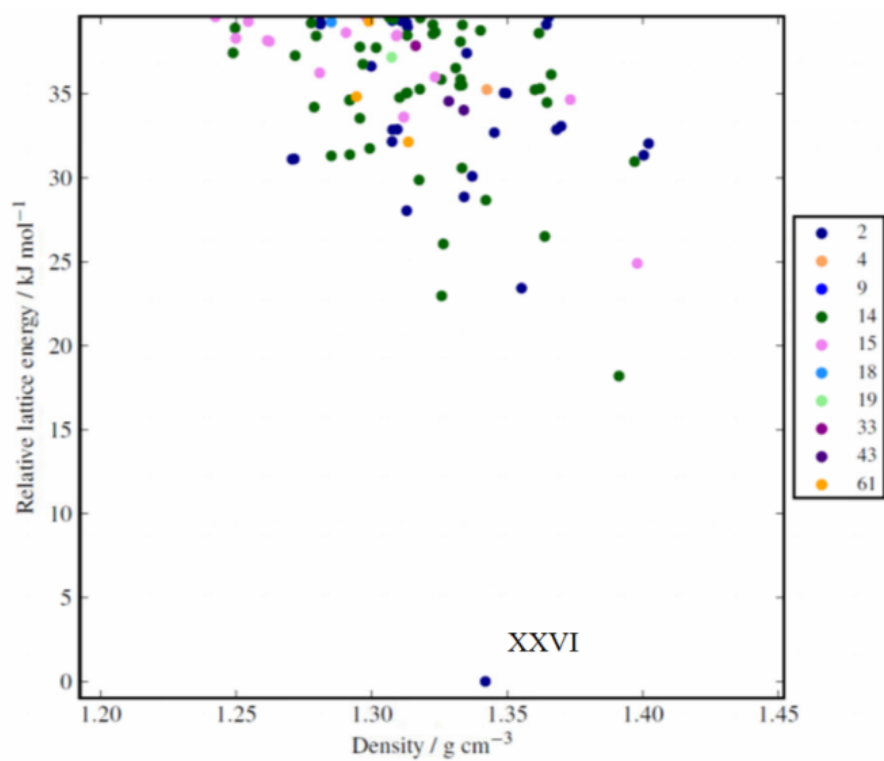
For the 2 sets of crystal structures, each of which was minimised using PCM with a dielectric constant of $3.0\epsilon_0$ and $7.0\epsilon_0$, the relative lattice energies are displayed in Figures 7.3a and 7.3b, respectively.

Once the blind test results were released, the observed crystal structure for molecule XXVI was lattice energy minimised using the same parameters and level of theory as was used in this CSP study and a structural comparison of this optimised, observed structure was performed against the crystal structures in the 2 submitted lists. However, the Day group results were disappointing for molecule XXVI as the observed crystal structure occurred in neither of these 2 lists. The list that used a dielectric constant of $3.0\epsilon_0$ provided crystal structures that were at least approximately 18 kJ mol^{-1} larger in lattice energy than the observed structure. The list that used a dielectric constant of $7.0\epsilon_0$ gave crystal structures whose lattice energies were comparable to the observed structure but still failed to provide a valid match.

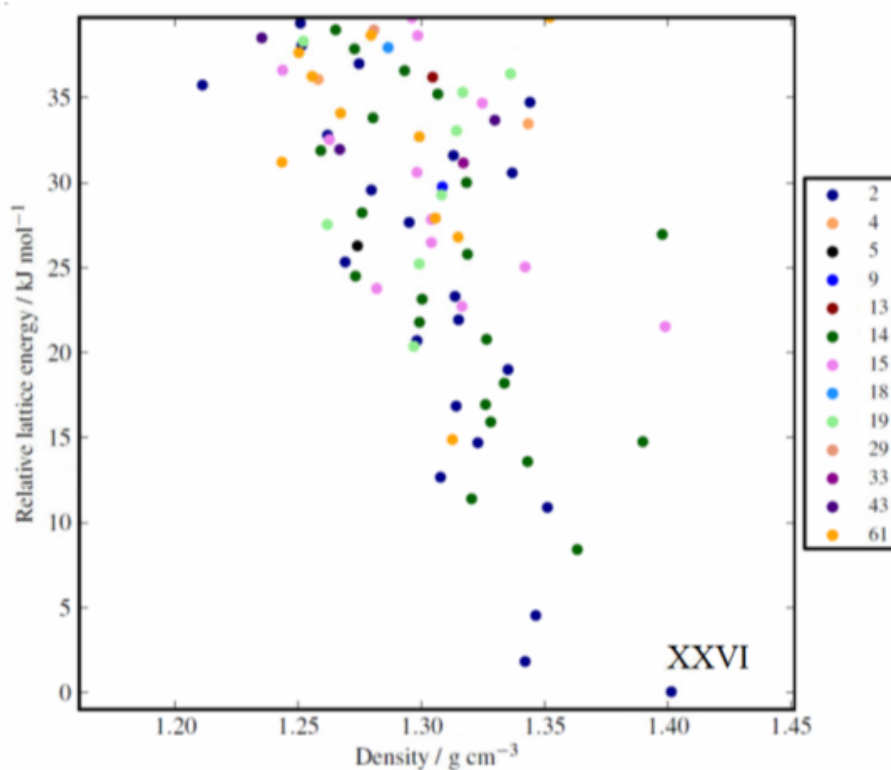
Although the exact cause for the absence of the observed crystal structure in the submitted lists is not entirely clear but an obvious place to start is by investigating the lack of conformational space exploration during the structure generation phase of the CSP process.

Figure 7.4a shows the molecular overlay between the observed molecular conformer and the closest structural match of all of the conformers used in the CSP study. The main geometrical difference arises from a twist of -39.36° about the C-N bond. This therefore significantly effects the packing arrangements such that the observed structure is not found in the final submitted list of crystal structures. Figure 7.4b shows 2 molecules in the observed crystal structure with the highlighted hydrogen bonds. There are combinations of non-covalent inter- and intramolecular bonds between the $\text{NH}\cdots\text{O}$ and $\text{NH}\cdots\text{Cl}$ hydrogen donor-acceptor hydrogen bond pairs, respectively.

Figure 7.5a shows the geometrical difference ($\text{RMSD}=0.729\text{ \AA}$) between the in-crystal geometry and its closest conformer match from the conformational search. This conformer was present within the set of 40 conformers that resided within the 30 kJ mol^{-1} tolerance that was applied to the initial 96 conformers from the conformational search. This overlay generally provides a good conformational match with the exception of the 94.65° torsion twist which causes the C=O bond vector in the carboxyl group to point towards the chloro-phenyl group on the opposing side of the molecule. This is in contrast to the in-crystal geometry where the C=O bond vector points away from the molecule and out into crystalline environment. Whilst this slight variation in molecular geometry is not proof of the failure to produce the observed crystal structure, it certainly played



(a)



(b)

Figure 7.3: XXVI final crystal structures for both sets using a polarisable continuum model with a permittivity of $3.0\epsilon_0$ and $7.0\epsilon_0$ for (a) and (b) respectively. The location of the observed crystal structure is labelled and was not found in either set.

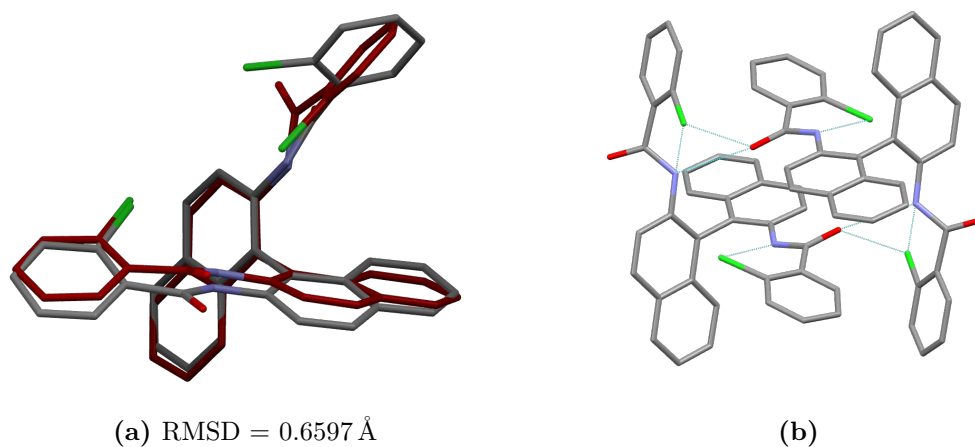


Figure 7.4: (a) shows the molecular overlay of the conformer of XXVI within the observed structure (coloured by element) and the conformer of the best matched structure submitted by the Author (carbons in red). (b) shows the intermolecular hydrogen bonds formed between 2 XXVI molecules in the observed crystal structure. Hydrogen atoms have been removed for clarity.

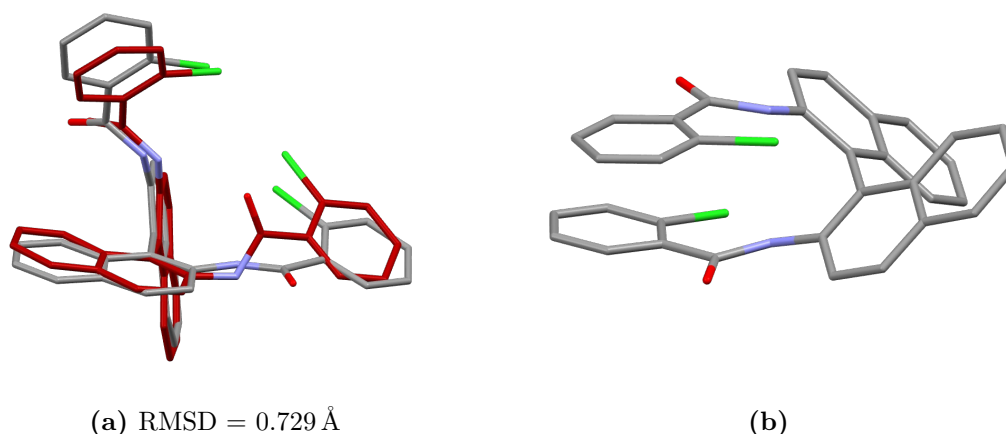


Figure 7.5: (a) shows the molecular overlays between the in-crystal geometry (coloured by element) and the closest conformer match (carbons in red) from a conformational search of molecule XXVI. (b) shows the gas phase geometry of molecule XXVI. Hydrogen atoms have been removed for clarity.

a role and further adds to the importance of correctly modelling molecular flexibility within the CSP process.

Molecule XXVI possesses 5 rotatable bonds and an ΔE_{strain} of $37.93 \text{ kJ mol}^{-1}$. The gas phase geometry of molecule XXVI, Figure 7.5b, shows that this molecule possesses both non-covalent polar and non-polar intramolecular interactions (see Figure 6.1 for specific examples). Therefore it lies outside of the upper bound (22.5 kJ mol^{-1}) predicted by Figure 6.7b in Chapter 6 for 5 rotatable bonds. However, when molecules form any type of non-covalent intramolecular interaction, the ΔE_{strain} becomes unpredictable.

The ΔE_{conf} value for molecule XXVI is 5.30 kJ mol^{-1} which is $32.63 \text{ kJ mol}^{-1}$ greater

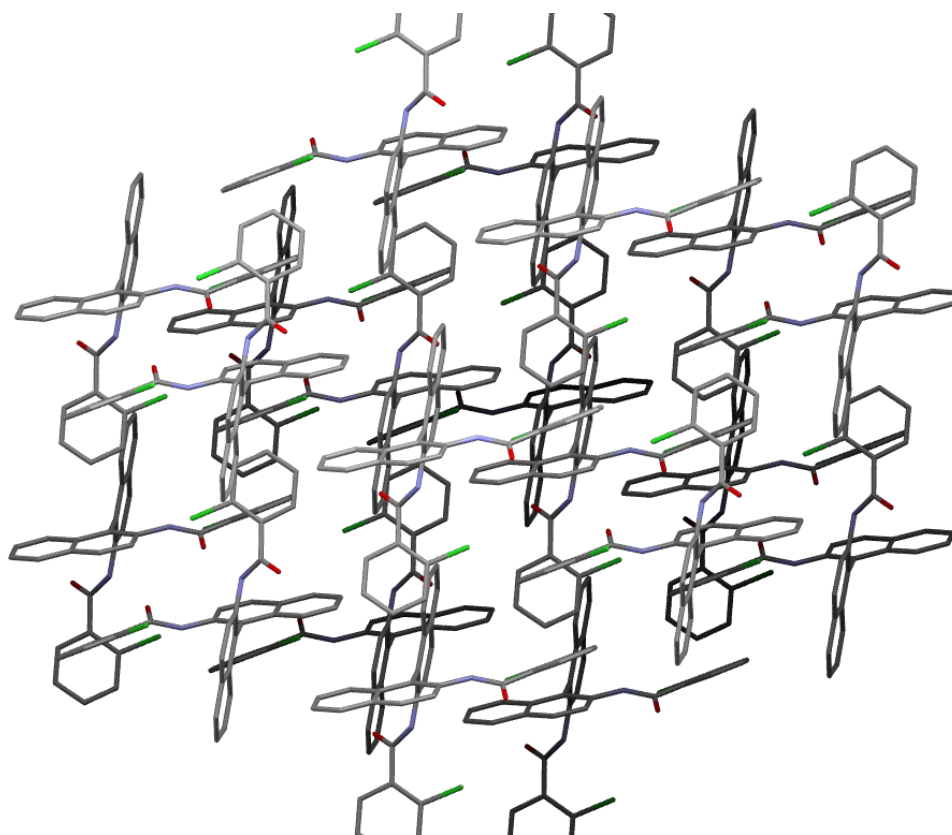


Figure 7.6: The observed crystal structure of molecule XXVI with atoms coloured by element and a depth cue to aid in the visualisation of the 3D crystal structure. Hydrogen atoms have been removed for clarity.

than the ΔE_{strain} . However, it does show that the energy cutoff of 30 kJ mol^{-1} was therefore too low and needed to be higher, perhaps set to 45 kJ mol^{-1} since this molecule forms non-covalent intramolecular interactions. This still would not have included the in-crystal conformation as its absence in the list of 96 conformers was due to a lack of searching the conformational space about each potential energy well where the conformer exists. Although there is no guarantee that this more thorough exploration of the conformational space would have found the observed crystal structure, it certainly would have increased likelihood of the CSP study being successful.

Over all of the participants in the sixth blind test, the observed crystal structure for molecule XXVI did not occur in the lists of submitted crystal structures for 22 of the 25 groups. However, 3 groups were successful in correctly predicting this crystal structure: Price, *et al.*, Elking & Fusti-Molnar and Neumann, Kendrick & Leusen. The packing arrangement of the observed structure for molecule XXVI is visualised in Figure 7.6.

All 3 groups listed the most likely crystal structure from at least 1 of their lists as the observed structure. One list submission from Elking & Fusti-Molnar used an empirical potential that yielded the observed structure to be ranked 8th. The re-ranking of this

list using a DFT (PBE-XDM) level of theory led to this crystal structure occurring at rank 1.

It was reported that many crystal structures in the unsuccessful lists did not possess intermolecular hydrogen bonds or the structure possessed a low packing coefficient. These issues were attributed to the search procedures only employing rigid conformations.

The groups of Price *et al.* and Neumann, Kendrick & Leusen implemented a flexible molecule search procedure which, again, highlights the need for molecular flexibility in all stages of the CSP process. However, Price *et al.* then implemented CrystalOptimizer for the crystal structure energy minimisation with the same level of molecular flexibility as what was presented by the author proceeded by a PCM calculation dielectric constant of $3.0\epsilon_0$ on a final list of crystal structures. This in contrast to Neumann, Kendrick & Leusen whose exact methodology remains elusive due to commercial interest but nonetheless implemented a periodic DFT approach.

7.4 Conclusion

Although a full post-mortem has not been conducted to date on why molecule XXVI was not found in the 2 submitted lists of crystal structures, its failure has been attributed by the Day research group to a lack of molecular flexibility during the structure generation phase. The methodology used to tackle this molecule in this chapter was closest to that utilised by Price *et al.* where the key difference was the treatment of molecular flexibility during the crystal structure generation phase, something that Price *et al.* included. Therefore this does appear to point to the reason for the failure of the prediction of molecule XXVI but an exact cause still remains ambiguous. Nonetheless, this key distinction between methodologies further justifies the need for firstly, the inclusion of molecular flexibility at every stage of the CSP process and, secondly, the need for an accurate and relevant description of this at each of these stages.

In an attempt to minimise the collateral damage that this issue can cause, a methodology introduced by Thompson & Day [120], whereby weighting conformers by their surface area and the ΔE_{conf} value, allows a novel approach for prioritising which molecular conformers are most likely to occur in the crystalline environment. This simple addition to the CSP process implemented would have re-ranked the importance of each conformer and could have led to the observed crystal structure being present in the final list of structures.

Thompson & Day's method could also be coupled with the principal displacement search procedure presented in Chapter 4 to now include molecular flexibility in both the search and lattice energy minimisation procedures. There is, of course, no guarantee that either of these methodologies (in tandem or isolation) would have led to a successful CSP but these certainly present valuable lessons and also provide additional motivation to include them in future CSP studies.

Whilst the methodology outlined in Chapter 4 was partially available to the Day group at the time of the sixth blind test, time constraints did not allow it to be conducted. In addition, there is no guarantee that this methodology would have provided a positive result but it could be assumed that it would have provided a more accurate result than what was submitted. Nonetheless, another molecule presented itself which allowed this state-of-the-art method to be implemented which could help provide reparations for the deficiencies of this chapter.

Chapter 8

Case Study II: Flufenacet

The molecule in this chapter, flufenacet, was proffered for the following CSP study by Professor Ulrich Greisser of the University of Innsbruck, Austria.

8.1 Introduction to Flufenacet

This chapter focuses on the CSP study of a molecule currently undergoing polymorph screening from the Greisser lab, the herbicide molecule flufenacet, Figure 8.1. This molecule was chosen due to the level of flexibility of the molecular system. Figure 8.1 also illustrates the 5 soft torsion angles within the molecule that can contribute to molecular flexibility in the crystalline environment.

In addition, at the time of conducting this research, no crystal structures of flufenacet were present in the CSD and no crystal structures were known by the Griesser group. Therefore this case can be treated as a ‘pseudo-blind test’ that could provide validation of the methods developed in this thesis. This CSP case study allows a test of the ideas

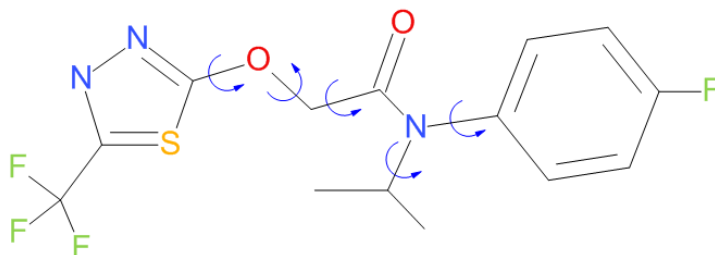


Figure 8.1: The flufenacet molecule annotated with blue, curved arrows that represent the soft torsion angles.

proposed by Thompson & Day [120] (Section 2.2) and the methodology presented in Chapter 4. Therefore the conformer surface area will be taken into account (in addition to the molecular energy) throughout specific stages of the CSP process.

8.2 Crystal Structure Prediction

The CSP methodology was partitioned into rigid and flexible molecule procedures. The comparison of the following state-of-the-art flexible molecule CSP procedure will be benchmarked against the well established methodologies of a rigid molecule CSP process. The procedure outlined in the remainder of this section is relevant to both methodologies. Specific details relevant to each methodology will then be dealt with in the relevant subsections.

The CSP procedure commenced with a 2D sketch of the flufenacet molecule that was used to derive a set of Cartesian coordinates. This 3D molecular representation was then subject to a conformer search using an LMCS method [47, 48] within MacroModel [46] where the starting molecular geometry is optimised before being perturbed along random combinations of its calculated normal modes. This implemented the OPLS2005 force-field [49, 50] with updated torsion parameters [157] and was terminated after 50,000 conformers were generated using the maximum and minimum movement distances of 3 Å and 6 Å, respectively. Each conformer that was generated was minimised using the Polak-Ribiere conjugate gradient [158] which was considered converged when the gradients were below 0.05 kJ/mol/Å. Duplicate molecular geometries were removed if the $RMSD_1$ values between any 2 conformers was below 0.02 Å. This yielded a total of 54 unique molecular conformers.

A geometry optimisation was then performed on each of the 54 conformers using a B3LYP-GD3BJ/6-311G** level of theory, using GAUSSIAN09 [67], where the resulting geometries were clustered to remove duplicates using in-house software if the $RMSD_1$ value between any 2 conformers was below 0.5 Å. These geometry optimised conformers were then ranked by their DFT molecular energies and a 30 kJ mol⁻¹ energy cutoff (above the global minimum) was applied to this list, yielding a set of 22 unique conformers.

In addition, the Connolly surface area (A_{Connolly}) was also calculated for each of these 22 unique conformers using MacroModel [46] and a probe radius size of 1.4 Å. The $E_{\text{conf,bias}}$ values, Equation 2.50, were then calculated for the 22 conformers and subsequently ranked by this new scoring function. Table 8.1 displays the 22 aforementioned conformers ranked by relative $E_{\text{conf,bias}}$.

Conformer ID	E_{rel} kJ mol ⁻¹	$A_{\text{Connolly,rel}}$ Å ²	Relative $E_{\text{conf,bias}}$ kJ mol ⁻¹
1	0.87	242.72	0
2	18.48	258.99	5.42
3	0.00	233.48	6.07
4	10.05	244.81	7.62
5	4.62	235.70	9.03
6	0.96	229.72	9.85
7	24.49	257.47	12.56
8	0.97	225.94	12.69
9	5.31	231.65	12.75
10	20.53	247.35	16.19
11	20.52	245.84	17.32
12	18.08	241.28	18.29
13	25.33	248.84	19.87
14	26.25	248.05	21.39
15	15.86	233.95	21.57
16	20.85	239.49	22.41
17	25.43	243.44	24.03
18	21.56	234.82	26.62
19	21.57	229.97	30.27
20	23.98	232.75	30.59
21	24.01	232.67	30.68
22	29.34	232.82	35.89

Table 8.1: The set of 22 unique conformers of flufenacet within 30 kJ mol⁻¹ of the global energy minimum ordered by their relative $E_{\text{conf,bias}}$ values.

As an aside, the largest A_{Connolly} of these 22 conformers is 258.99 Å² from conformer 2. It is the general rule in CSP that the lower the molecular energy, the more favourable that a conformation is to form low energy crystal structures. However, the opposite is also considered true when applied to the molecular surface area. The greater A_{Connolly} a conformation possesses, the more of the molecule that is available to form non-covalent intermolecular interactions with its neighbouring molecules when inserted into the crystal.

From here, the rigid and flexible molecule CSP methodologies diverge and will be described in their respective sections.

8.2.1 Rigid Molecule CSP Methodology

For each of the 22 conformers, 6000 crystal structures were generated within each of the 6 most common space groups ($P\bar{1}$, $P2_1$, $P2_1/c$, $C2/c$, $P2_12_12_1$, $Pbca$). For any conformer that produced any crystal structures than were within 20 kJ mol^{-1} of the global energy minimum crystal structure, an additional 500 crystal structures were generated for the 18 next most common space groups ($P1$, $C2$, $P1c$, $C1c$, $P2/c$, $P2_12_12$, $Pca2_1$, $Pna2_1$, $Fddd2$, $Pccn$, $Pbcn$, $P4_1$, $P4_3$, $I4_1/a$, $P4_12_12_1$, $P4_32_12_1$, $P3_1$, $P3_2$).

These crystal structures were then subject to rigid molecule lattice energy minimisation using DMACRYS with the W99 potential with updated hydrogen bonding terms [118] and a Van der Waals cutoff radius of 15 \AA . This was followed by a clustering procedure using in-house software and a 1 kJ mol^{-1} energy window which yielded a total of 2854 unique crystal structures.

Crystal structures that were yielded from conformers 1, 2 and 3 and resided within 40 kJ mol^{-1} of the global minimum crystal structure were subject to a flexible molecule lattice energy minimisation using the respective intramolecular energy models that were fitted to these conformers. The resulting crystal structures were lattice energy minimised once more but now using multipoles, instead of fixed point charges, calculated at the B3LYP/6-311G** level of theory. This final step then allows a fair comparison between the search procedures of the rigid and flexible molecule CSP processes as the electrostatic component of the lattice energy minimisation stages are, as will be explained in the following section, now identical.

8.2.2 Flexible Molecule CSP and Analysis

The state-of-the-art flexible molecule CSP methodology, outlined in Chapter 4, was applied to model flufenacet. Since this method can incur a considerable expense, the top three conformers from Table 8.1 were selected for these calculations.

The top three ranked conformers 1, 2 and 3 are the conformers with the lowest $E_{\text{conf,bias}}$, the highest A_{Connolly} and the lowest molecular energy, respectively. These 3 molecular geometries are shown in Figure 8.2.

Conformer 2, Figure 8.2b, possesses the largest surface area in that it is the most ‘open’ geometry of the 3 conformers which gives it the greatest chance of forming intermolecular interactions with its neighbouring molecules when it is placed within the crystal structure.

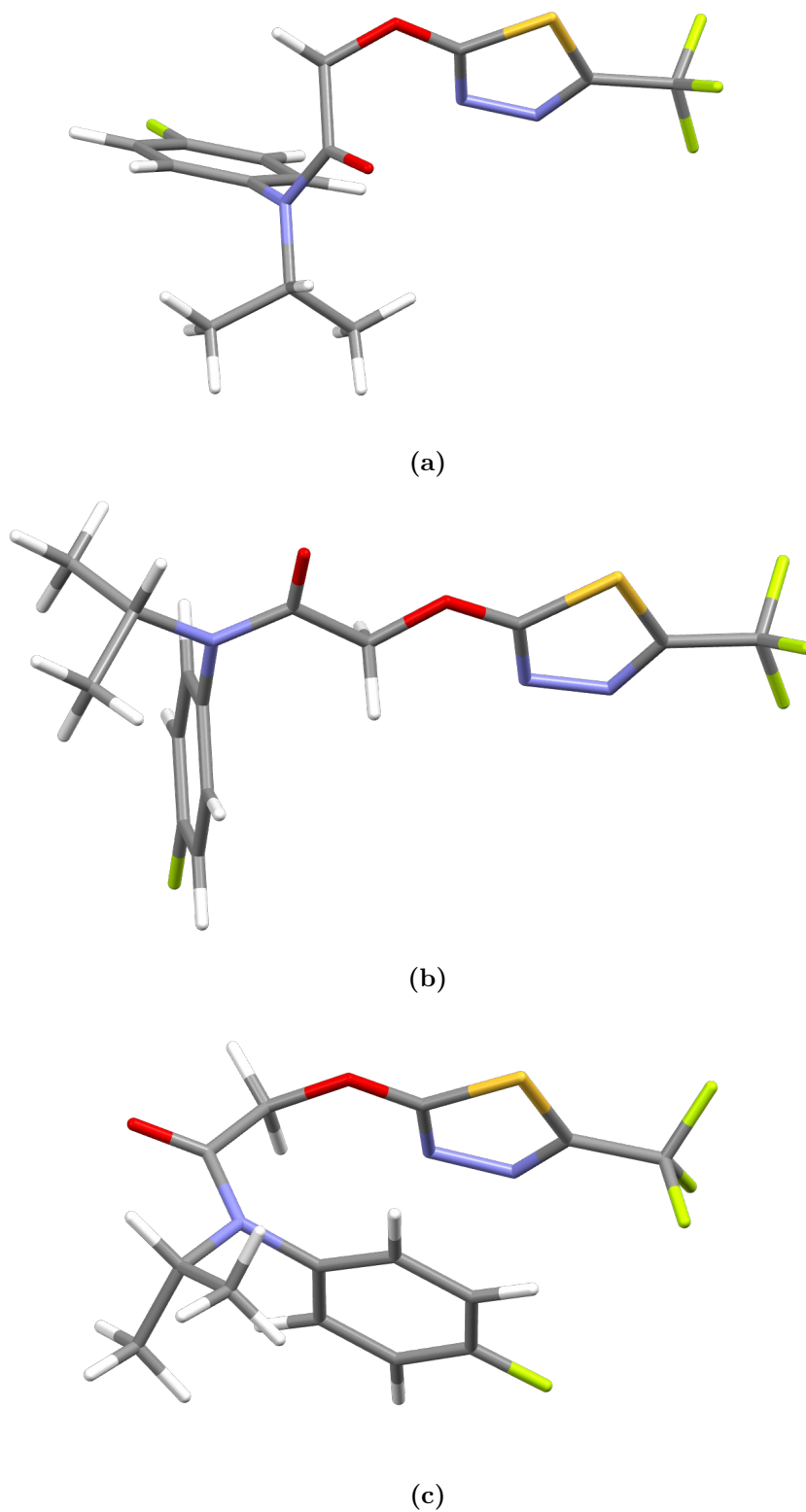


Figure 8.2: (a), (b) and (c) showing conformers 1, 2 and 3 of the flufenacet molecule that possess the lowest $E_{\text{conf,bias}}$, the largest surface area and the lowest molecular energy respectively.

Conformer 3, Figure 8.2c, possesses the lowest molecular energy and surface area of the three conformers. This conformer is exploiting its molecular flexibility to ‘curl-up’ on itself to form intramolecular interactions and lower its molecular energy. This conformer is more spherical than conformers 1 and 2 and hence possesses a lower surface area.

Conformer 1, Figure 8.2a, possesses the lowest $E_{\text{conf,bias}}$ and hence provides a compromise between the two competing factors of molecular energy and surface area.

Each of the 3 conformers was subjected to a geometry optimisation followed by a principal displacement calculation using the numerical gradients molecular energy where both calculations implemented the B3LYP-GD3BJ/6-311G** level of theory using GAUSSIAN09 [67]. The number of principal displacements required for the flexible search procedure can be calculated from a more detailed analysis of that performed in Chapter 6.

8.2.2.1 Selecting the Number of Principal Displacements

Chapter 6 demonstrated that a force constant value of $0.0840 \text{ mDyne } \text{\AA}^{-1}$ was required to encompass all of the necessary principal displacements for 95% of molecules. For flufenacet this would require the first 12 principal displacements (ordered by force constant) for all conformers 1, 2 and 3. The flexible search methodology has only been conducted on a maximum of 3 principal displacements (see Chapter 4) and so this dramatically increases the number of dimensions to be sampled.

Owing to a lack of rigorous testing of this methodology, it is safer and more prudent to experiment with a lower number of dimensions (principal displacements) that is closer to the original value of 3. More specifically, the analysis conducted in Chapter 6 was on all of the molecules in the test set that possessed between 3 and 8 rotatable bonds. Flufenacet possesses 5 rotatable bonds and so this subset will now be analysed in isolation.

The force constant values required to reduce the $RMSD_1$ values between the displaced gas phase and in-crystal geometries below the recommended tolerance of 0.2 \AA for these molecules are presented in Table 8.2. From this data the maximum number of principal displacements required to perform the geometry interconversion between the set of in-crystal and gas phase conformers is 8 from the WEWPOD01 system. This 8th principal displacement possesses a force constant value of $0.0160 \text{ mDyne } \text{\AA}^{-1}$ that is the highest in this set of molecules.

Transferring this limit to flufenacet, conformers 1, 2 and 3 possess 5, 6 and 5 principal displacements below this force constant tolerance, respectively. To be consistent, it

CSD Refcode	Maximum Principal Number of Principal Displacements	Minimum Force Constant
APTSPN	3	0.0049
BAMDIC	3	0.0031
COKQEZ	5	0.0118
DAZYUZ	3	0.0037
ICOQIB	1	0.0043
WENTAL	6	0.0080
WERVIY	6	0.0093
WEWPOD01	8	0.0160
YIOBAP	1	0.0032
ZUHWAA	4	0.0064
	Maximum	0.0160

Table 8.2: Maximum number of principal displacements, and their corresponding force constants, $\text{mDyne } \text{\AA}^{-1}$, required to reduce the $RMSD_1$ value between the displaced gas phase and in-crystal geometries to below 0.2 \AA for all 10 molecules possessing 5 rotatable bonds. The molecule in bold text represents the largest minimum force constant value in this set of molecules.

would be ideal to allow molecular flexibility along 6 principal displacements for each conformer.

However, it is desirable to keep the computational cost to a minimum for this procedure and so the number of principal displacements will be rounded to 5 for all conformers. This now defines the number of dimensions that are needed to be sampled during the flexible search procedure. This is a more manageable number but an advantage of this method is that the complexity of this methodology can be gradually increased.

8.2.2.2 Calculating the Principal Displacement Bounds

A 30 kJ mol^{-1} energy bound above the equilibrium geometry for each conformer for each direction of each principal displacement was chosen. This was to allow the limits of the conformational space that will be sampled, and fitted at a later stage, to be sufficiently larger than the region that will be used during the lattice energy minimisation phase (22.5 kJ mol^{-1}). These boundaries are identical with what was selected in Chapter 4. The summary of the limits for each bound is presented in Table 8.3.

For conformers 1 and 3, the displacement limits about the equilibrium geometry are generally not symmetric. Although the degree of symmetry becomes more apparent as the principal displacements with higher force constants are introduced. This asymmetry

Conformer ID	Principal Displacement Index	Force Constant mDyne Å ⁻¹	Lower Bound Å	Upper Bound Å	Displacement Range Å
1	1	0.0010	-10.55	6.86	17.41
	2	0.0011	-4.50	8.34	12.84
	3	0.0015	-8.23	6.86	15.09
	4	0.0041	-5.62	2.70	8.32
	5	0.0077	-2.97	2.53	5.50
2	1	0.0012	-11.19	11.50	22.69
	2	0.0023	-5.37	5.39	10.76
	3	0.0045	-4.34	4.29	8.63
	4	0.0061	-5.71	5.19	10.90
	5	0.0156	-4.16	4.54	8.70
3	1	0.0012	-5.41	2.01	7.42
	2	0.0022	-7.47	4.89	12.36
	3	0.0029	-3.67	7.05	10.72
	4	0.0043	-5.50	2.18	7.68
	5	0.0054	-2.57	2.11	4.68

Table 8.3: Upper and lower bounds for each principal displacement for the flufenacet conformers 1, 2, and 3.

shows that the principal displacement is anharmonic. The anharmonicity is due to the molecular conformations in that one direction across a given principal displacement incur higher energy costs than the opposite direction and hence the 30 kJ mol⁻¹ limit is reached with smaller displacement values.

Conformer 2 possesses displacement bounds that are the most harmonic out of the 3 conformers. This can be justified from Figures 8.1 and 8.2b where the soft torsion angles identified in the former possesses minimal steric hindrance and are ‘free’ to rotate by equal distances about the equilibrium geometry. This is not true for conformers 1 and 3 which led to the anharmonicity observed in the displacement values.

The range of the principal displacements generally decreases as the force constant value increases. This is because higher force constants are linked with principal displacements that describe molecular motions that incur higher energy costs. This is not true for all principal displacements but is generally the case.

8.2.2.3 Fitting the Energy Model

The next stage was to perform the fitting procedure to the intramolecular energy. 10,000 Sobol points between 0 and 1 were generated over the 5 dimensions. For each conformer, these points were used to create a set of 10,000 displaced conformations (30,000 conformations in total). Each of these displaced conformations were subject to a single point energy calculation using the B3LYP-GD3BJ/6-311G** level of theory.

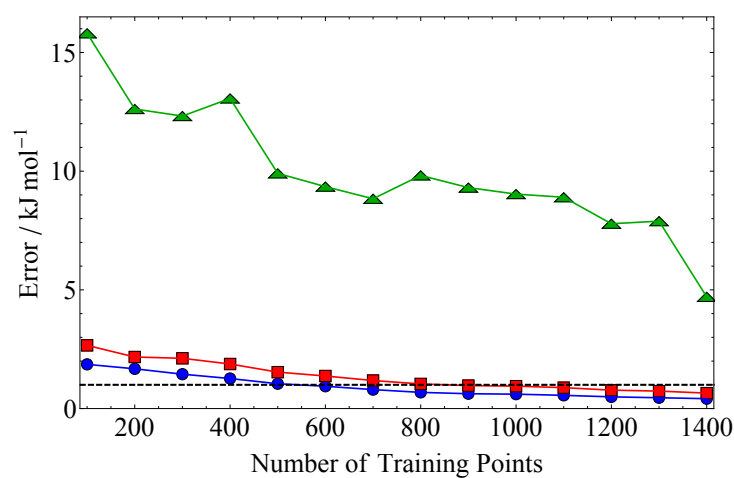
A separate Gaussian Process model was fitted to these displaced conformation energies for conformers 1, 2 and 3. The model was only required to be fitted to energies that resided within 30 kJ mol^{-1} of the energy of the equilibrium geometry. Applying this cutoff reduced the number of energy points from 10,000 to approximately 2400, 1800 and 2200 for conformers 1, 2 and 3 respectively. Thompson & Day [120] observed that the ΔE_{strain} value for a given molecule will not exceed the value of 22.5 kJ mol^{-1} . This is also in exact agreement with the ΔE_{strain} value observed in the Chapter 6 for molecules with 5 rotatable bonds, Equation 6.1. Therefore the fitting was performed up to 30 kJ mol^{-1} to ensure the most accurate fit possible up to the energy limit of Thompson & Day's observation.

From the cropped points per conformer, the first 1000 points that possessed an energy within 22.5 kJ mol^{-1} were used for the test set. The fitting procedure was then performed multiple times using the remaining points as a training set at 100 point intervals.

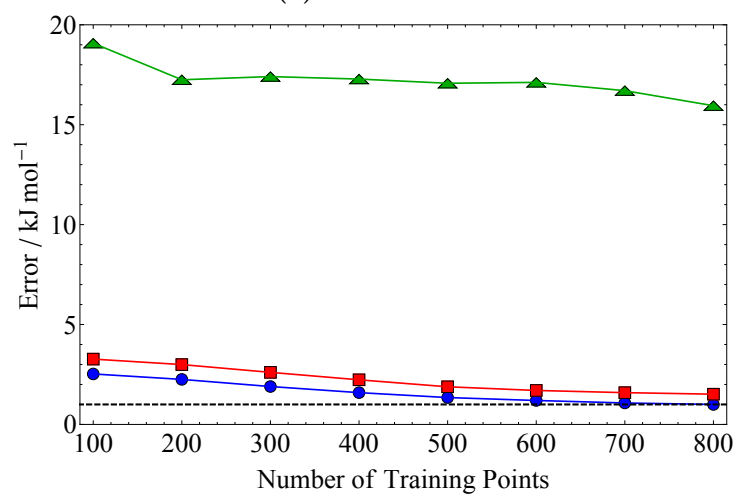
The mean unsigned errors, accompanied by the standard deviations and maximum absolute errors for each conformer at each 100 training point interval are displayed in Figure 8.3.

The number of training points available vary from 800 to 1400. The maximum absolute error also varies between conformers and there is no consistent value per number of training points. This value steadily decreases as more training points are added but led to unpredictable movements between these intervals. The maximum absolute error clearly requires a significantly larger amount of training points to reduce it to below the desired 1 kJ mol^{-1} tolerance over the 5 dimensions.

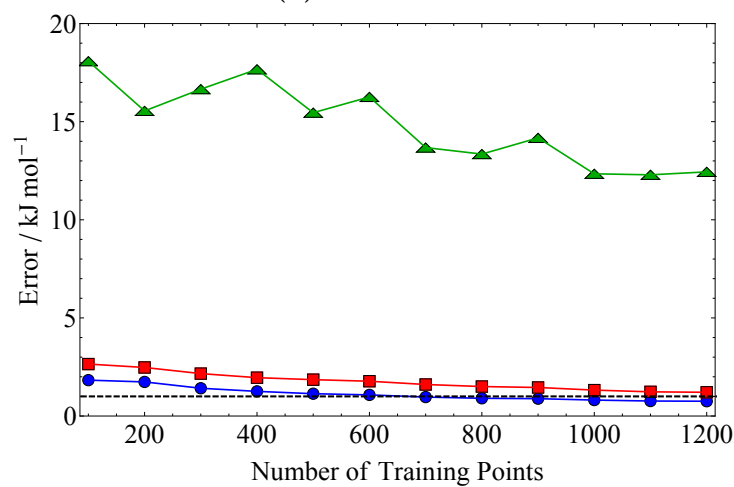
The unsigned mean errors begin at approximately 2 kJ mol^{-1} to 3 kJ mol^{-1} for all 3 conformers. This error gradually decreases to below 1 kJ mol^{-1} at 700 training points for conformer 1 and converges to approximately 1 kJ mol^{-1} for conformers 2 and 3 at the maximum number of training points, 800 and 1200 respectively. In contrast to the maximum absolute error, the unsigned mean error shows that the overall fit of the surface is good with only specific regions causing issues.



(a) Conformer 1



(b) Conformer 2



(c) Conformer 3

Figure 8.3: Shows the mean unsigned error (red), standard deviation (blue) and maximum absolute error (green) for each number of training points used for the flufenacet conformers 1, 2 and 3. The dashed lines are plotted at 1 kJ mol^{-1} to aid in the reading of the figures.

The standard deviation follows the same trend as the unsigned mean error in that it steadily decreases as more training points are added. This value drops below 1 kJ mol^{-1} at 900, 1200 training points for conformers 1 and 3 respectively. The standard deviation for conformer 2 does not drop below 1 kJ mol^{-1} before the maximum 800 training points that were available for the fitting procedure.

8.2.2.4 Flexible Molecule Crystal Structure Generation

For each conformer, 10,000 trial crystal structures were generated in the 6 most common space groups ($P\bar{1}$, $P2_1$, $P2_1/c$, $C2/c$, $P2_12_12_1$, $Pbca$) (60,000 structures per conformer). The structures all possessed $Z'=1$ to minimise the computational expense of this stage. This procedure implemented an extension of the original structure generator [94] whereby the intramolecular DOFs can simultaneously be sampled with the lattice parameters.

In this case, the intramolecular DOFs consist of the first 5 principal displacements of the given conformer. This led to an 18 dimensional space to be explored (lattice parameters: \mathbf{a} , \mathbf{b} , \mathbf{c} , α , β , γ ; 3 molecular positions and 3 molecular orientations within the unit cell; and the 5 principal displacements).

The intramolecular energy model that possessed the highest number of training points was used. This was to reduce errors (unsigned mean and absolute maximum) to a minimum to afford the most accurate results as possible.

8.2.2.5 Flexible Molecule Lattice Energy Minimisation

The crystal structures were subject to flexible molecule lattice energy minimisation. This intermolecular lattice energy minimisation employed DMACRYS with the W99 potential with updated hydrogen bonding terms [118], a Van der Waals cutoff radius of 15 \AA and conformer dependent, fixed point charges. The minimisation of the lattice energy employed a Simplex algorithm [72] that allowed the molecular geometry to flex along combinations of its principal displacements. More specifically, DMACRYS was used for the rigid molecule lattice energy minimisations whereas in-house code was used to perform the molecular distortions along the chosen set of principal displacements.

Which principal displacements the molecular geometry is displaced along is user selected. The quantity of displacement for each principal displacement forms the input parameters for the Simplex minimiser. The distorted molecular geometry is reinserted into the crystal structure and then fed back into DMACRYS. This process iterates between the two stages until a total lattice energy minimum is found.

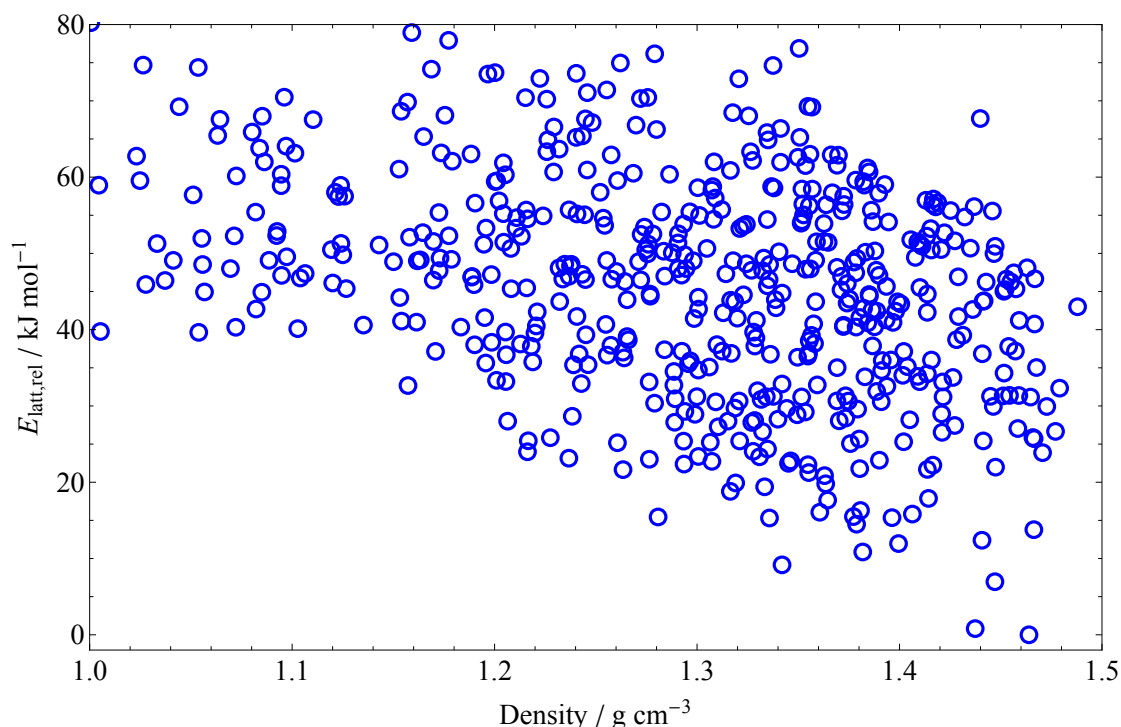


Figure 8.4: CSP landscape for the rigid molecule CSP of flufenacet.

These minimised crystal structures were then clustered using COMPACK and a 1 kJ mol^{-1} energy window. Only crystal structures whose lattice energy was within 20 kJ mol^{-1} of the global minimum were subject to a rigid molecule lattice energy minimisation using DMACRYS with multipoles calculated using the B3LYP-GD3BJ/6-311G** level of theory yielded from GAUSSIAN09 [67].

8.3 Rigid Molecule CSP Results

The CSP landscape for the rigid molecule of flufenacet is plotted in Figure 8.4. The unique crystal structures densely populate the region from approximately 10 kJ mol^{-1} to 70 kJ mol^{-1} above the global lattice energy minimum with only 7 crystal structures lying within a 10 kJ mol^{-1} window of this global minimum.

The global lattice energy minimum crystal structure is visualised in Figure 8.5. This crystal structure occurs in the $P\bar{1}$ space group for conformer 1 and is ranked with the most favourable $E_{\text{conf,bias}}$ in Table 8.1.

Since particular emphasis was placed on conformers 1, 2 and 3, in Section 8.2.2, Figure 8.6 shows the CSP landscape broken down for these three conformers only. Table 8.4 shows the total relative energy decomposition for the lowest energy crystal structure of each conformer.

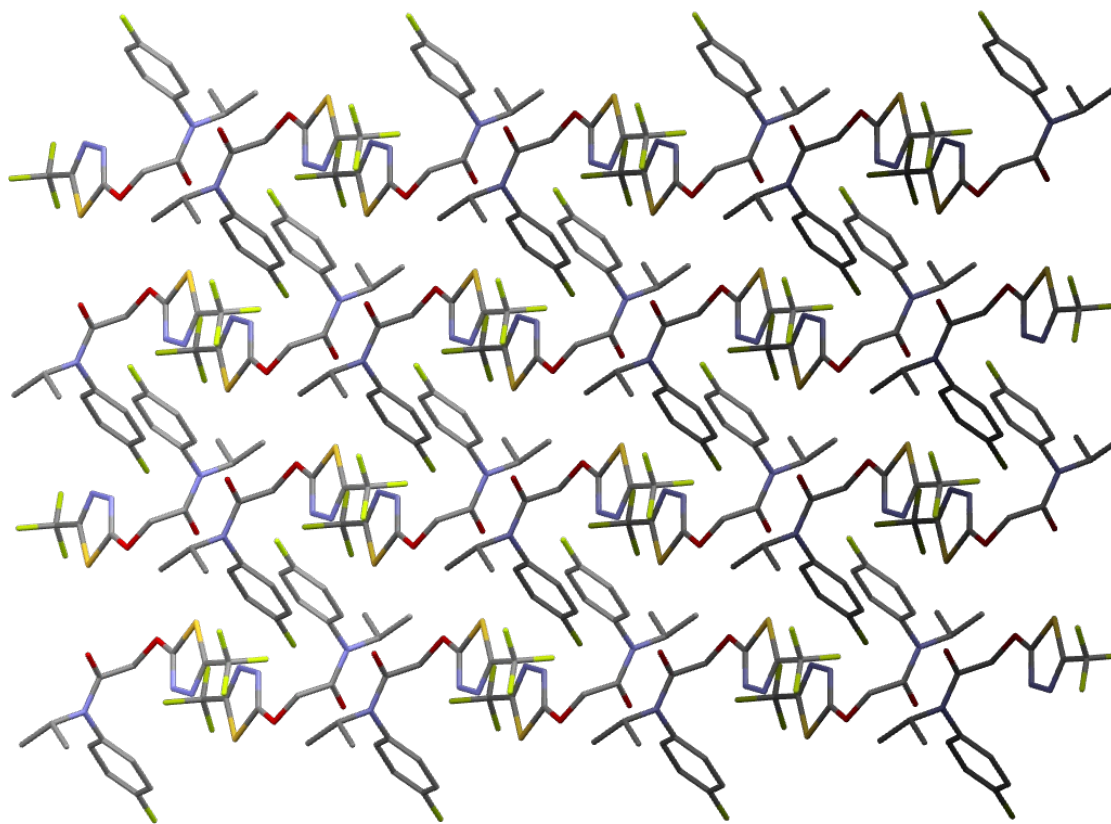


Figure 8.5: Global lattice energy minimum crystal structure for the flufenacet molecule in the $P\bar{1}$ space group. Hydrogen atoms have been removed for clarity.

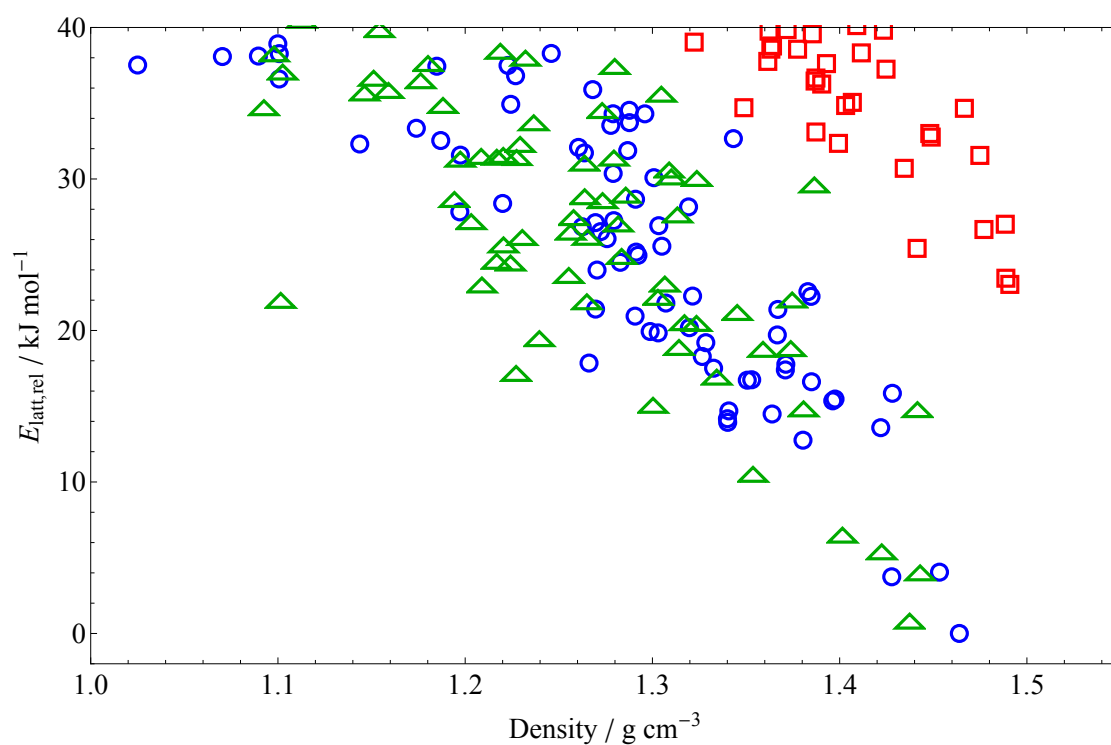


Figure 8.6: CSP landscape for the rigid molecule search and lattice energy minimisation CSP procedure of flufenacet conformers 1 (blue), 2 (red) and 3 (green).

Conformer ID	$E_{\text{rel,intra}}$ kJ mol ⁻¹	E_{inter} kJ mol ⁻¹	$E_{\text{rel,latt}}$ kJ mol ⁻¹
1	0.87	-109.02	-108.15
2	18.48	-114.78	-96.30
3	0.00	-107.42	-107.42

Table 8.4: The relative intramolecular energy, $E_{\text{rel,intra}}$, intermolecular energy, E_{inter} , and the relative total energy, $E_{\text{rel,latt}}$, for the lowest energy crystal structures from flufenacet conformers 1, 2 and 3.

Conformer 1 yields the lowest lattice energy crystal structure. This is in contrast to the ‘traditional’ CSP hypothesis that the conformer with the lowest molecular energy, conformer 3 in this case, should yield the lowest lattice energy. Furthermore, the lowest energy crystal structure yielded from conformer 3 possesses an $E_{\text{rel,tot}}$ that is 0.73 kJ mol⁻¹ greater than the lowest energy crystal structure for conformer 1. This is also observed in Figure 8.6 between the lowest energy crystal structures yielded from conformers 1 and 3. The inclusion of this $E_{\text{conf,bias}}$ term prioritises more of the relevant areas of the CSP landscape that are, in this case, lower in total lattice energy.

The molecular energy of conformer 2, which possesses the largest A_{Connolly} , lies at an energy that is 18.48 kJ mol⁻¹ greater than the global energy minimum afforded from conformer 3. However the lowest energy crystal structure yielded from conformer 2 possesses the lowest E_{inter} as the molecule possesses the most available surface area for forming intermolecular interactions. Observing Figure 8.6, this difference is shown when comparing the positions of the lowest energy crystal structures yielded from conformers 1 and 2.

However, this reduction in E_{inter} does not compensate the energy penalty of forming this more strained conformer. This therefore shows that it is the balance between the molecular energy and the A_{Connolly} that provides the most energetically stable crystal structures.

8.3.1 Intermolecular Lattice Energy versus Molecular Surface Area

Figure 8.7 shows the correlation between the molecular surface area of the 22 conformers presented in Table 8.1 against the crystal structure that yielded the lowest intermolecular component of the lattice energy from that conformer.

A correlation exists in that molecular conformers with larger surface areas yield lower intermolecular lattice energies. This is because as the molecule ‘opens-up’ and the molecular surface area increases, more of the molecular functional groups are able to

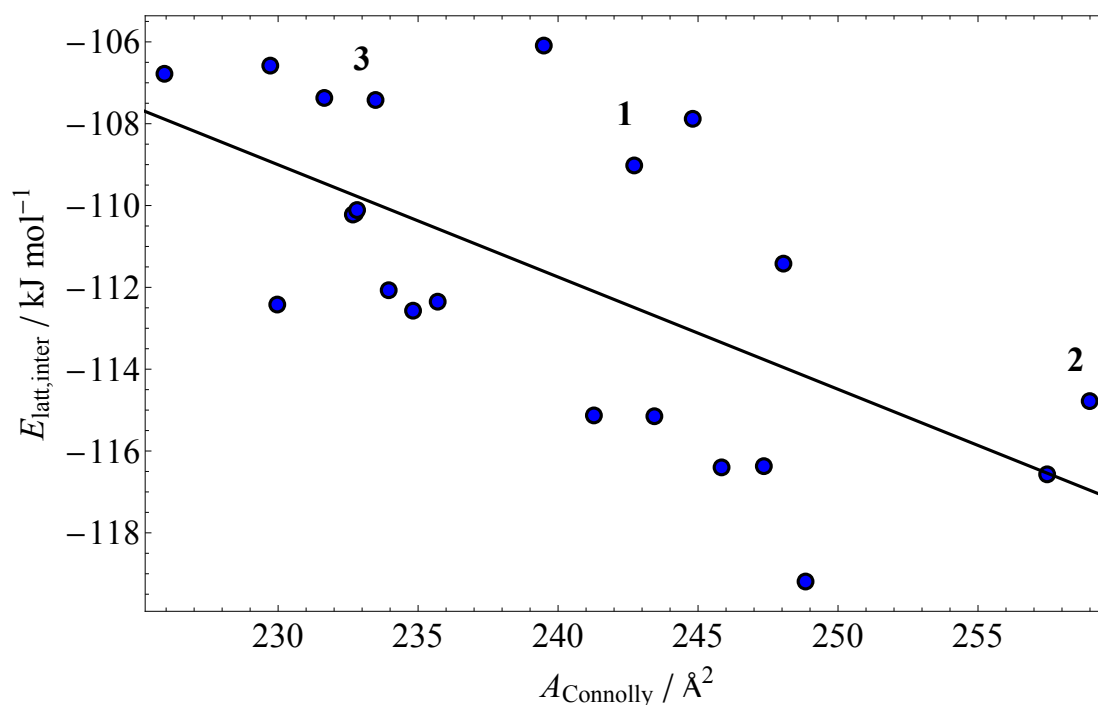


Figure 8.7: The correlation between the molecular surface area, A_{Connolly} , and the intermolecular lattice energy, $E_{\text{latt,inter}}$, from the lowest energy crystal structures for each of the 22 conformers of flufenacet. Conformers 1, 2 and 3 are highlighted. The line represents a line of best fit to the data which possesses the equation $E_{\text{latt,inter}} = -0.27(\text{\AA}^{-2}) \cdot A_{\text{Connolly}}(\text{\AA}^2) - 45.89(\text{ kJ mol}^{-1})$.

participate in forming intermolecular bonding interactions with neighbouring molecules. Therefore the intermolecular lattice energy is lowered. The line of best fit is shown in Figure 8.7 that possesses the equation:

$$E_{\text{latt,inter}} = -0.27(\text{\AA}^{-2}) \cdot A_{\text{Connolly}}(\text{\AA}^2) - 45.89(\text{ kJ mol}^{-1}). \quad (8.1)$$

In addition, conformers 1, 2 and 3 have been labelled in this figure also since these conformers are used in the flexible molecule CSP process whose results will be described momentarily. From these labels, it can be seen that conformer 3, that possesses the lowest molecular energy, yields the highest intermolecular lattice energy value of the 3 conformers. This is in contrast to conformer 1, which possesses the lowest $E_{\text{conf,bias}}$, and yields a lower intermolecular lattice energy than conformer 3. Conformer 2 possesses the largest surface area and hence yields the lowest intermolecular energy of these 3 conformers. However, other molecular conformers possess a smaller molecular surface areas and lower intermolecular lattice energies.

Therefore it is a sound assumption to make in that generally molecular conformers that possess larger surface areas yield crystal structures that possess lower intermolecular lattice energy values. Although there are exceptions to this rule, it is generally the case.

Space Group	Conformer		
	1	2	3
$P\bar{1}$	4991	2832	4308
$P2_1$	5317	5403	4500
$P2_1/c$	7072	5283	5647
$P2/c$	2591	3548	5311
$P2_12_12_1$	4136	3918	5679
$Pbca$	2981	2257	3659

Table 8.5: The number of valid crystal structure minimisations for flufenacet conformers 1, 2 and 3.

8.4 Flexible Molecule CSP Results

The number of valid lattice energy minimisations for each conformer in each space group are displayed in Table 8.5. The number of valid minimisations vary by both space groups and conformers.

Generally, space group $Pbca$ gives the lowest valid number of minimisations with no conformer achieving >40% of valid minimisations. From the opposite perspective, space group $P2_1/c$ yields the largest number of valid minimisations with all conformers achieving >52% of valid minimisations.

The CSP landscape for the flexible search is displayed in Figure 8.8. The first observation is that there are a larger amount of crystal structures than for the rigid CSP landscape (Figure 8.6). This is due to the greater dimensionality of the potential energy surface when using the flexible CSP method whereby more stable minima exist.

Conformer 1 still yields the global energy minimum crystal structure. The comparison of the global energy minimum in-crystal conformation for the flexible and rigid CSP methods are displayed in Figure 8.9. These 2 geometries are similar (RMSD=0.375 Å) where the only differences are a twisting of the chloro-phenyl group (21.29°), a rotation of the tri-fluoro group (27.92°) and the twisting of isopropyl group (15.43°).

Figure 8.8 shows the location of many crystal structures on the CSP landscape leading to the figure becoming cluttered. Therefore Figure 8.11 shows the individual CSP landscapes for conformers 1, 2 and 3 and the accompanying in-crystal molecular geometry that afforded the lowest total energy crystal structure for that particular conformer.

The distributions for each conformer can now be observed more clearly. Conformer 2 now shows that there are no crystal structures that exist within 10 kJ mol⁻¹ of the global total energy minimum. This highlights, as it did with the rigid molecule CSP,

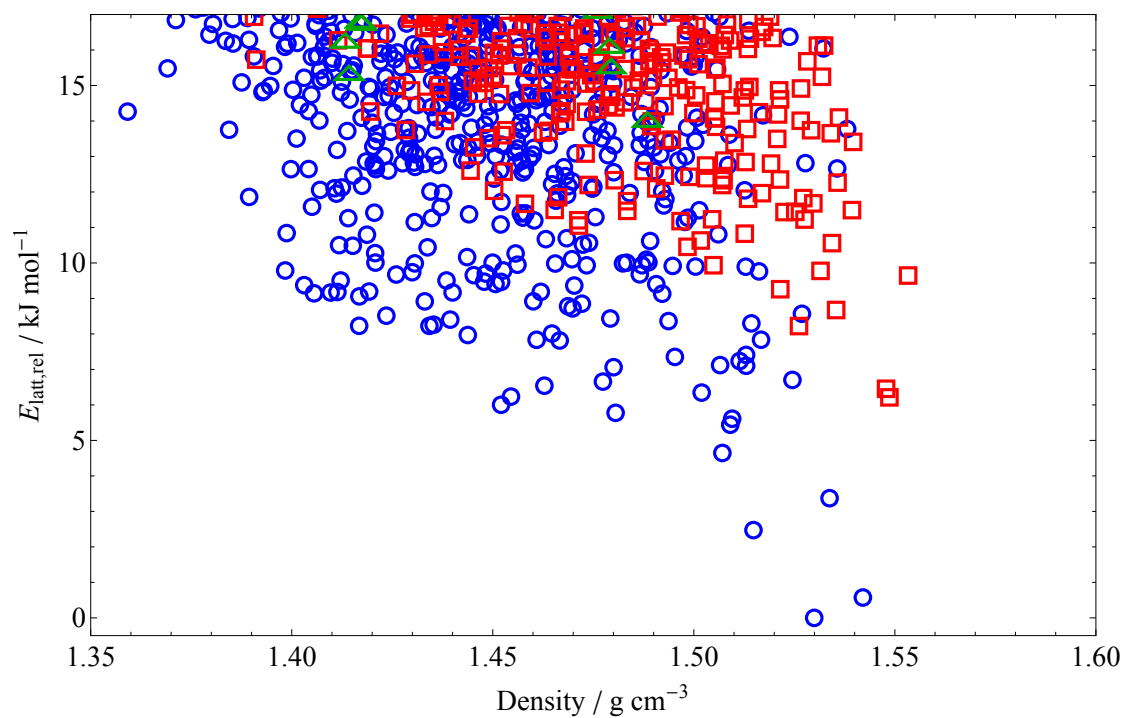


Figure 8.8: CSP landscape for flufenacet conformers 1 (blue), 2 (green) and 3 (red).

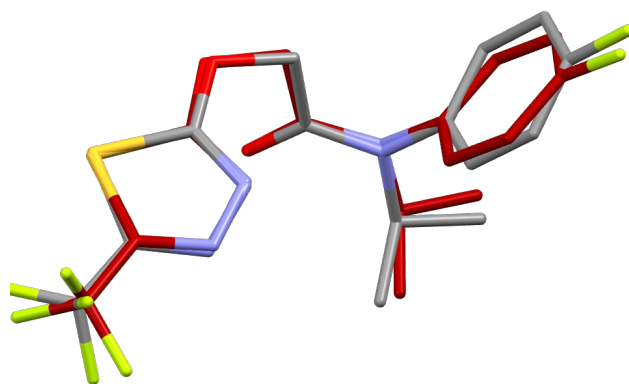


Figure 8.9: Shows the overlay of the molecular geometries ($\text{RMSD}=0.375 \text{ \AA}$) of the global energy minimum crystal structures for flufenacet from the rigid (carbons in red) and flexible (coloured by element) CSP processes. Hydrogen atoms have been removed for clarity.

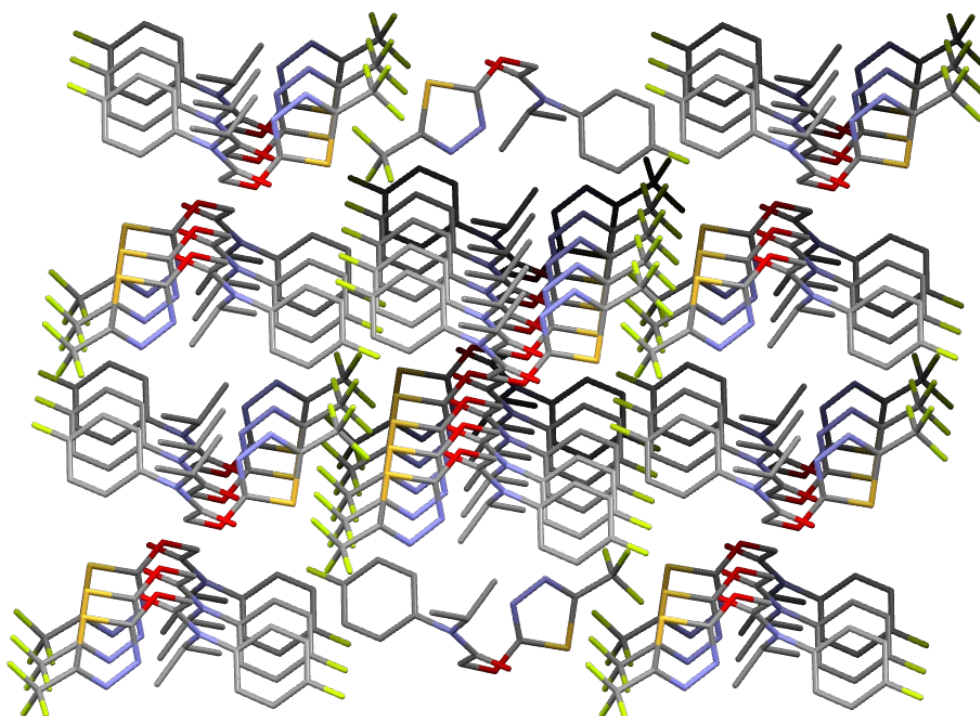


Figure 8.10: The global energy minimum crystal structure yielded from the flexible molecule CSP process. Hydrogen atoms have been removed for clarity.

that the molecular conformation with the largest surface area does not yield the lowest energy crystal structure. Completely prioritising the molecular energy, conformer 3, leads to these low energy crystal structures but it is, again, the compromise between the molecular surface area and molecule energy, conformer 1, that yields the lowest energies.

The ΔE_{strain} values from the lowest energy crystal structures for conformers 1, 2 and 3 are 4.34 kJ mol^{-1} , 0.60 kJ mol^{-1} and 2.94 kJ mol^{-1} , respectively. In relation to the data presented in Chapter 6, these ΔE_{strain} values are well within the maximum tolerance of 22.5 kJ mol^{-1} calculated by Equation 6.1.

8.4.1 Evaluation of the Sampling Procedure

Figure 8.12 shows the relative lattice energy of the crystal structures found against the number of valid lattice energy minimisations for each of the 6 space groups for conformer 1. The analogous figures for conformers 2 and 3 are presented in Appendix D.

Regardless of the space group it can be observed that low energy crystal structures (relative to each space group) are still being generated even at a higher number of valid minimisations. This shows poor convergence of each space group and that, due to the

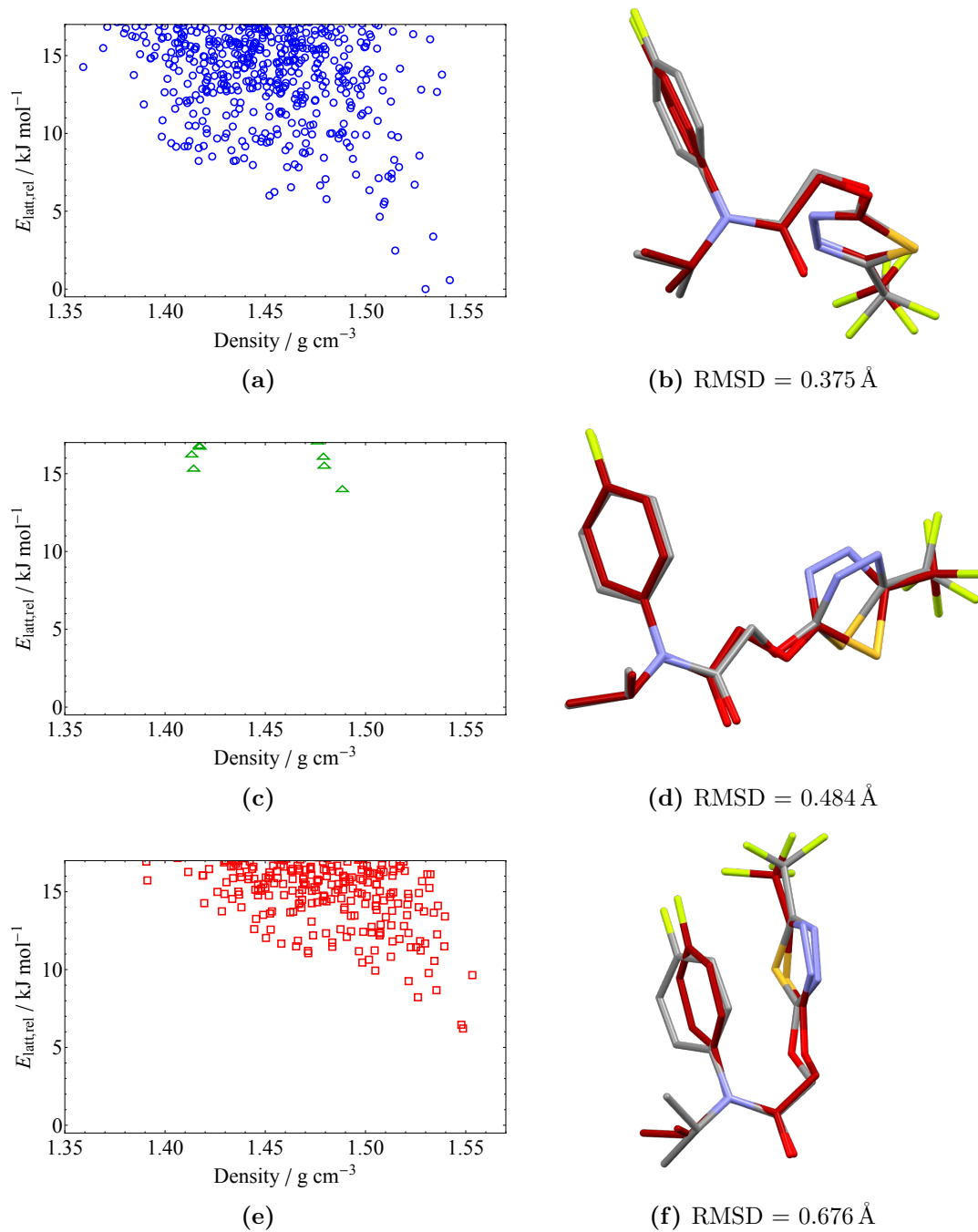


Figure 8.11: CSP landscapes for conformers 1, 2 and 3 in (a), (c) and (e), respectively. (b), (d) and (f) shows the in-crystal molecular geometry that afforded the lowest total energy minimum crystal structure (coloured by element) against the original geometry of that conformer (carbons in red) for conformers 1, 2 and 3 respectively. Hydrogen atoms have been removed for clarity.

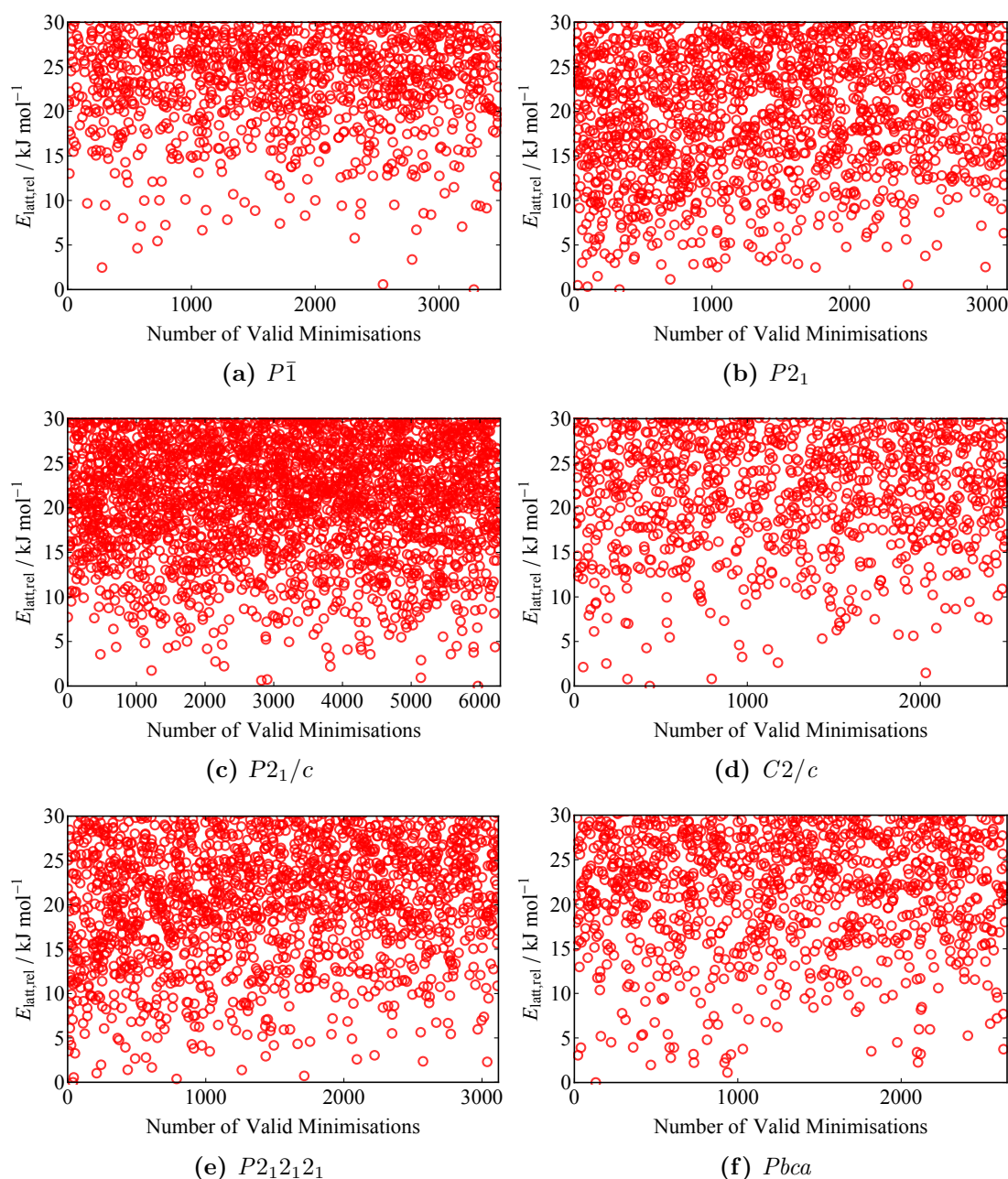


Figure 8.12: All 6 plots show the number of unique crystal structures being generated for conformer 1 of flufenacet for a given total energy against the number of valid minimisations. The figure captions indicate the space group.

dimensionality of the PES, more crystal structures are required to be generated to ensure all areas of the PES have been adequately sampled.

In particular, the $P\bar{1}$ space group was considered converged after the generation of 6000 crystal structures for FUQLIM in Chapter 4. This is not the case the $P\bar{1}$ space group when the flexible molecule structure generation using 5 principal displacements (as opposed to 3 principal displacements for FUQLIM).

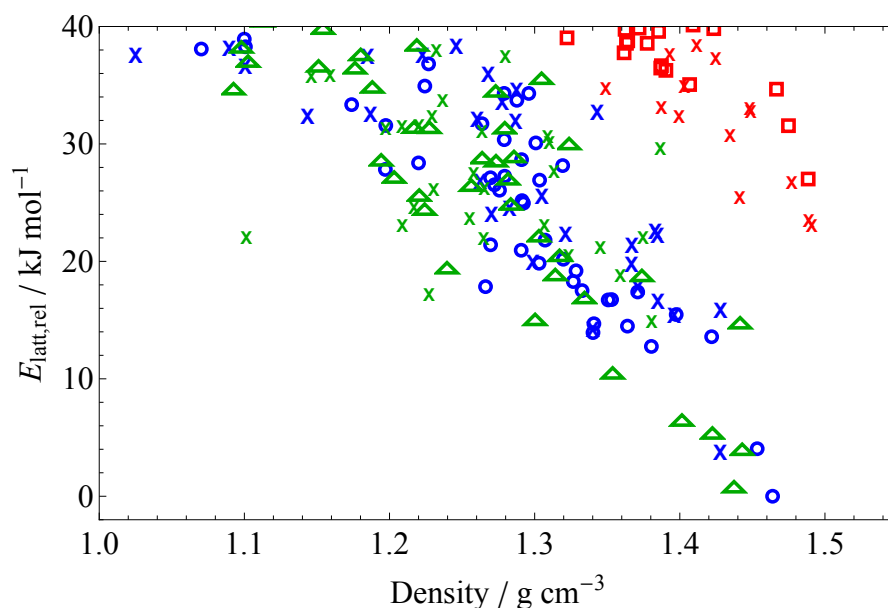


Figure 8.13: CSP landscape for flufenacet conformers 1 (blue circles), 2 (green triangles) and 3 (red squares) under the rigid molecule approximation. An 'X' represents the location of crystal structures that were present in the rigid search set but not the flexible search set whereas the other shapes show the location of crystal structures that were present in both sets.

In addition, the molecules from the rigid molecule CSP process that were then re-lattice energy minimised using the energy models for conformers 1, 2 and 3 were compared against the completely flexible molecule set, Figure 8.13. In both sets, the global energy minimum crystal structure was found but only 59.8% of crystal structures from the rigid search set existed in the flexible search set. This shows that the number of crystal structures or the sampling technique when generating the crystal structures was inadequate.

8.5 Conclusions

Although no empirical crystal structures were present at the time of completion of this research, a variety of lessons can still be gleaned on the behaviour of the flexible molecule CSP process.

Two differing CSP approaches were conducted on the herbicide molecule, flufenacet. The aim was to contrast the results yielded from using a 'traditional' CSP approach treating the molecule as rigid after the conformational search against using the state-of-the-art method for treating molecular flexibility in CSP that was outlined in Chapter 4. However, since there are no observed structures available, it is impossible to say which method is more accurate or even if molecular flexibility was needed for this

study. Nonetheless, this was the first implementation of the principal displacement search methodology and it was shown that the method is formalised and is ready to be used for any other future CSP studies.

The use of molecular surface area to weight the conformers based on both their molecular surface area and the intramolecular energy reordered the priority of the conformers and led the top 3 conformers being those that possessed the best $E_{\text{conf,bias}}$, the lowest molecular energy and the highest molecular surface area. The former, in both rigid and flexible CSP methodologies, yielded the global energy minimum crystal structure. This was always up to several kJ mol^{-1} lower in energy than the lowest energy crystal structure yielded from the conformer that possessed the lowest molecular energy.

Interestingly, this research also showed that the conformer with the largest molecular surface area yields the lowest intermolecular lattice energy but this is not offset by the additional ΔE_{strain} that is required to ‘open-up’ the molecule. Therefore, again, it is the balance of this molecular surface area and molecular energy that is required to yield the lowest energy crystal structures.

The implementation of the principal displacement methodology provided a slight insight into the behaviour of the method. This was using 5 principal displacements for each flufenacet conformer. Using 10,000 points (crystal structures) to sample this 5 dimensional space was not adequate for this study as only approximately 60% of crystal structures from the rigid search procedure were found in the flexible search. Therefore it would be recommended for future CSP studies that more points would be used. Exactly how many points is still unknown and would require rigorous testing of the method.

With respect to the intramolecular energy model, 10,000 training points was too low a number as many of the generated points were discarded as they yielded molecular conformations that laid above the 30 kJ mol^{-1} energy cutoff. This was determined by observing the maximum absolute error in the fitting procedure. Conformers 2 and 3 possessed approximately a 15 kJ mol^{-1} error with the maximum number of training points available. Conformer 1 still possessed approximately a 5 kJ mol^{-1} error which is significantly lower when compared against conformers 2 and 3 but is still higher than a 1 kJ mol^{-1} tolerance. Nonetheless, the unsigned mean errors were at approximately this 1 kJ mol^{-1} tolerance which shows that the fitting was generally good but did not provide a good fit in certain areas of the PES.

These fitting regimes may have yielded bad fits due to the vast amount of energies of the displaced molecular conformations that existed higher than the 30 kJ mol^{-1} cutoff. This vastly reduced the amount of the training points that were available to perform

the fitting. The solution would be to continue generating these displaced molecular conformations until a target number that existed under the energy cutoff was present.

Rigorous testing of this method still needs to be commenced but this chapter has provided an insight into its behaviour and serves as a benchmark to future testing and development to proceed.

Chapter 9

Conclusions & Future Work

Although this thesis possesses 6 chapters of novel research, it can be partitioned into 3 distinct sections when observing the thesis as a whole.

The first section is the research involved in Chapter 3. This quantifies the effect of incorporating current methods to treat molecular flexibility in CSP for small organic molecules that possess limited flexibility. This section also tests the ability for the revised W99 potential to accurately describe hydrogen bonding interactions in molecular crystals.

The second section of research concerns the research conducted in Chapters 4 to 6. This section commences with the creation of a novel method to treat molecular flexibility at the trial crystal structure generation phase by implementing the principal displacements of the gas phase conformer of a molecule to perform geometrical distortions. This section then aimed to establish a set of rules to determine which principal displacements, how many and how far to traverse along each one are required to feed into this flexible molecule crystal structure generation method. This was achieved by firstly presenting another novel method that found the optimal solution to these questions posed in the previous sentence before implementing this on a large test set of molecules.

The final section then focussed on the application of CSP methodologies to 2 specific case studies. The first was an unsuccessful CSP that only implemented current CSP methodologies. The second implemented the novel flexible structure generation method. This was the first test case for this method since its inception and allowed stringent analysis on its behaviour to be tested throughout.

A more detailed discussion about the further work for each of these sections will now commence.

9.1 Molecular Flexibility & Williams99 Force-Field Testing

The research in this thesis commenced with the testing of the revised W99 potential that includes a more accurate description of the hydrogen bonding interactions in molecular crystals. This chapter also quantified the effect of adding molecular flexibility into the CSP calculations during the lattice energy minimisation phase. This research was generally successful but highlighted areas where the CSP process can breakdown. This is especially prevalent with respect to the crystal structures of glycine where the formal, electrostatic charges present on the termini of the molecule, are greatly affected and effected by the neighbouring molecules.

These charges from the neighbouring molecules must therefore be modelled accurately if the correct packing arrangement of glycine is to be calculated. This can be performed by using a PCM model on the already optimised crystal structures yielded from CrystalOptimizer and the revised W99 potential. These are the most accurate results that include descriptions of molecular flexibility and the intermolecular hydrogen bonding interactions. The addition of the PCM model allows a more accurate description of the charge-charge interactions present in crystal and also allows variation of the molecular position and orientation from effects from the crystalline environment. Therefore this should also improve the unit cell packing and yield more accurate results (hence lowering the N_{Lower} and ΔE values). This theoretical model has not been discussed in this report as it is not directly relevant to the research but more information can be found within other sources.[159]

Another area of further work for this section would be to conduct calculations on the set of 54 molecules using an ‘OrigRigid’ method where the original W99 potential and a rigid body approximation would be implemented. These results would quantify the effect of incorporating both modifications (molecular flexibility and W99rev) into the CSP calculations. This would answer the question, ‘if the addition of flexibility and the W99rev incur accuracy increases of 5.5% and 6.3%, respectively, does incorporating both modifications incur an 11.8% increase in accuracy?’. This question is not as simple as may first appear. Although individually, these modifications allows increases in accuracy, using both may induce a synergic effect that gives an additional improvement.

9.2 Flexible Molecule Structure Generation

Chapter 4 presents a novel method that implements molecular principal displacements to search the conformational space of a molecule. This can be coupled with the typical structure generation that already exists within CSP to simultaneously search the inter- and intramolecular environments. The FUQLIM molecule provided an ideal test case for this new methodology as a previous CSP attempt had previously failed to predict the existence of 2 of the 3 known polymorphs due to deficiencies in the treatment of molecular flexibility in the CSP process. This new methodology successfully predicted the existence of all the 3 known polymorphs of FUQLIM and hence provided a proof-of-concept case for the procedure.

However, this methodology still contained some inherent questions that required answering. For FUQLIM, the crystal structures of each of the 3 polymorphs were known in advance of any CSP calculations being performed and so the in-crystal geometries were also known. Therefore the shape and size of the search space (how many principal displacements and how far to traverse along each one) were already defined and so the procedure could be tailored to ensure that the relevant in-crystal molecular geometries were sampled. Therefore the questions that remained were: how many principal displacements were required?, which principal displacements were required?, and how far does each one need traversing along?

The answer to these questions required a whole other chapter of research, Chapter 5, that presented several methodologies for performing the interconversion between the gas phase and in-crystal geometries and observing which principal displacements were required. The simplest method was the utilisation of a Cartesian approach however this neglected the subtlety of curvilinear space and was not deemed scientifically robust.

Another method was to minimise the RMSD between the 2 conformations by successively introducing each principal displacement in the order of increasing force constant. Whilst this was scientifically robust, the procedure was relatively computationally expensive and was not prone to finding the ideal solution that possessed the global minimum RMSD value between the 2 conformations.

The third and most efficient method was solve the equation that described the molecular distortion required to minimise the RMSD between the gas phase and in-crystal geometries directly. This was achieved by reformulating the equation into a least squares format. This method was extremely computationally efficient and found the global RMSD minimum, hence this method was chosen to proceed with the research.

The final chapter of this section of research, Chapter 6, implemented this least squares methodology on a large test set of molecules with varying levels of flexibility to observe which principal displacements were required to convert the gas phase into the in-crystal geometry. It was found that a force constant value up to $0.084 \text{ mDyne } \text{\AA}^{-1}$ was required to perform this conversion to within a measured RMSD tolerance of 0.2 \AA . Therefore including all principal displacements up to this force constant value will, in 95% of cases, include all of the necessary molecular motions to reduce the RMSD between the gas phase and in-crystal conformations to below this 0.2 \AA tolerance. This now answers the questions of how many and which principal displacements are required for the methodology outlined in Chapter 4. However, more definition is required on these rules.

Further research would commence by performing more of these calculations on molecules with 2 to 5 rotational bonds. This will complement the test set of molecules that already exist and allow a fairer number of molecules to be present for each rotatable bond bin.

9.3 CSP Case Studies

The final section of research in this thesis was performing two CSP studies on specific molecules. This first of these focused on molecule XXVI as part of the sixth blind test. A rigid molecule crystal structure generation process, followed by the a flexible molecule lattice energy minimisation procedure was performed that failed to predict the observed crystal structure. The reason heavily points to the lack of modelling of molecular flexibility during the structure generation phase of the CSP process. Since the latter is highly relevant to the research conducted in this thesis, a second CSP case study was performed with the goal of redeeming the failures of this blind test molecule.

The research conducted in Chapters 4 to 6 was implemented for the CSP study of flufenacet (a ‘standard’ rigid molecule CSP procedure was also performed to measure the differences using this novel method). This was the first instance, with the exception of the FUQLIM molecule, that this novel methodology had been implemented. Therefore the input parameters used were analysed during the procedure in attempt to learn more about its behaviour. This also allowed the method to be formalised for future CSP studies.

The results showed that ranking conformers by their $E_{\text{conf,bias}}$ value (a weighting of molecular surface area and molecular energy) as opposed to solely the molecular energy yielded crystal structures that were lower in lattice energy. Also, weighting conformers purely by molecular surface area yielded the lowest intermolecular lattice energies.

However, the additional ΔE_{strain} required to ‘open-up’ the molecule was not recouped by this reduced intermolecular lattice energy.

The number of training points required to fit the PES was needed to be increased as the unsigned mean error generally gave an approximate 1 kJ mol^{-1} value at the maximum number of available training points but in the majority of cases gave errors that were up to ten times larger for the maximum absolute error. Therefore rigorous testing is required to find an optimal number of training points that will provide an accurate fit to the intramolecular energy whilst keeping the computational costs to a reasonable figure.

The number of crystal structures generated needed to be increased as it was apparent that the flexible molecule search procedure was incomplete. This was due to a large number of crystal structures that were present in the rigid molecule search not being present in the flexible search. This is due to added dimensionality of the intra- and intermolecular space that is required to be explored. Therefore another area of further research would be to determine how many crystal structures are required to effectively sample a given space group when sampling a multitude of dimensions of the molecular conformational space coupled with the ‘standard’ dimensions of the crystal (unit cell parameters, molecular position and orientation within the unit cell).

Appendix A

54 Small, Organic Molecules

This appendix illustrates the skeletal formulas of the 54 molecules used in Chapter 3. All diagrams are accompanied by a six character CSD reference code. The 4 molecules that were added for this study possess their CSD reference codes in **bold** font.



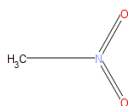
ETHLEN



QQQCIV



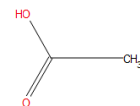
METAMI



NTROMA



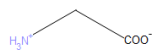
FORMAM



ACETAC



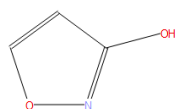
EDAWIP



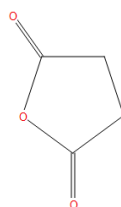
FEPNAP



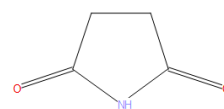
TRAZOL



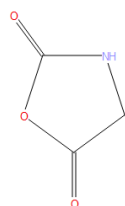
NEZMUA



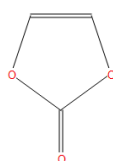
SUCANH



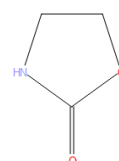
SUCCIN



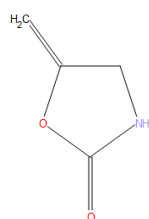
OXAZDO



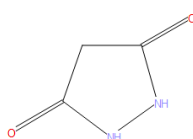
VINLYC



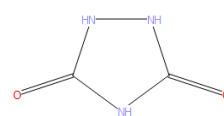
OXAZIL



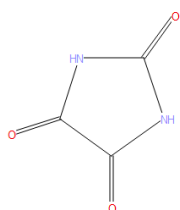
YOBQAH



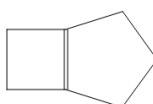
DUNVEN



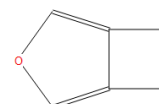
KOXRIY



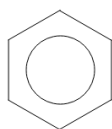
PARBAC



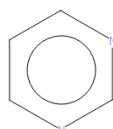
JIZREP



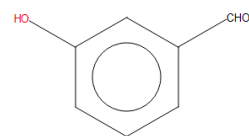
XULDUD



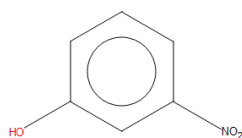
BENZEN



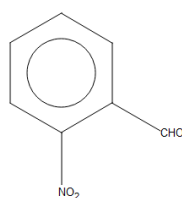
PRMDIN



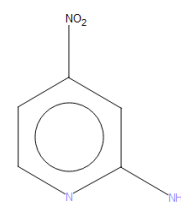
XAYCIJ



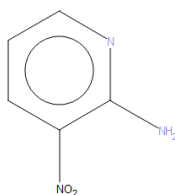
MNP HOL



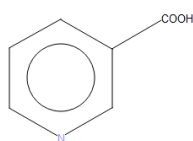
NIBZAL



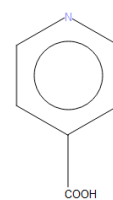
SEGRUR



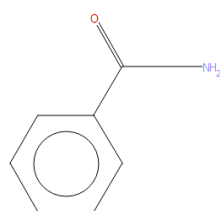
AMNTPY



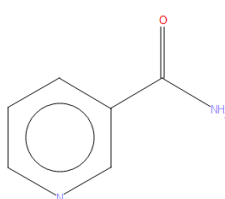
NICOAC



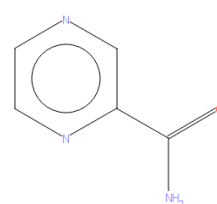
ISNICA



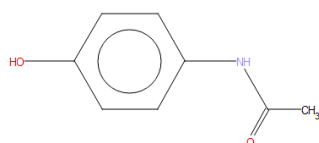
BZAMID



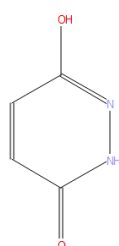
NICOAM



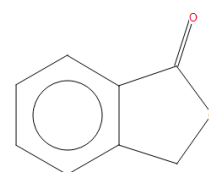
PYRZIN



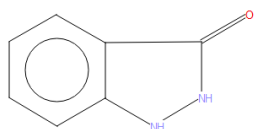
HXACAN



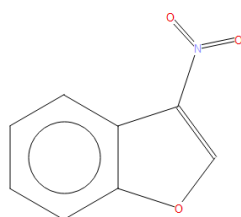
MALEHY



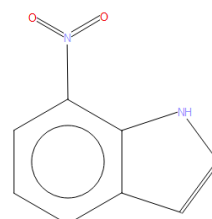
HEZQUY



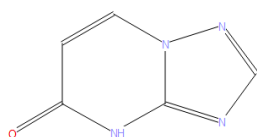
FADMIG



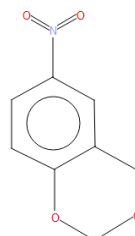
NIVBUP



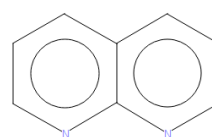
FULKUS



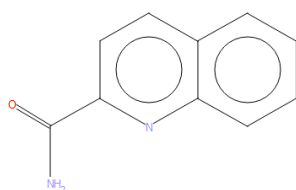
QAJYIJ



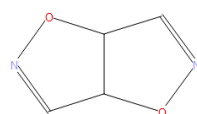
GEYWIQ



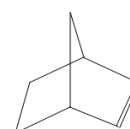
NAPTyr



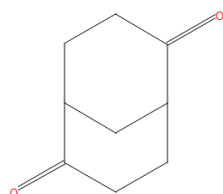
QUINCB



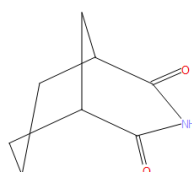
DIHIXL



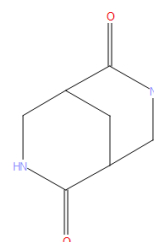
HOBOP



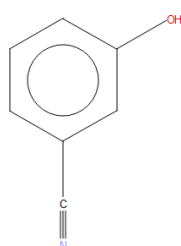
HEBBEV



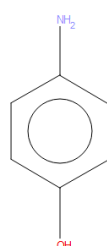
BOQQUT



DOGTIC



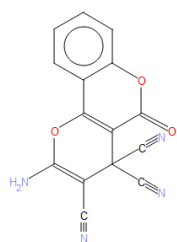
ABEGEU



AMPHOL

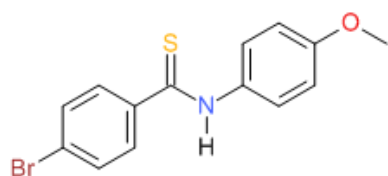


FORAMO

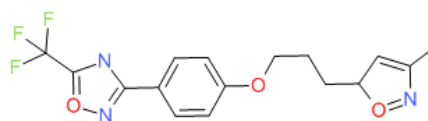
**FOYBOL**

Appendix B

Molecular Geometry Interconversion Using Principal Displacements

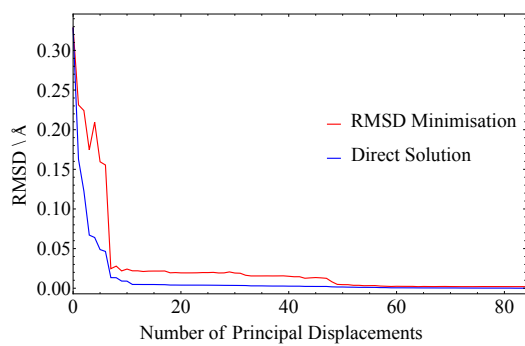


(a) HIBGUV

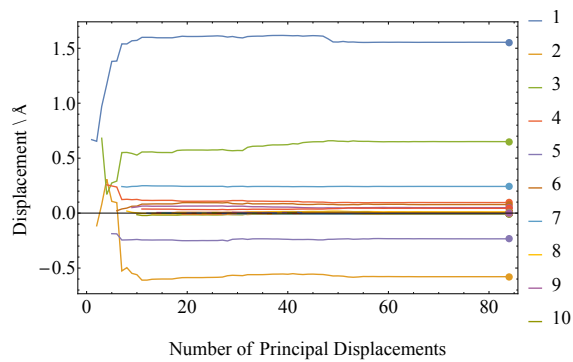


(b) HAJYUN

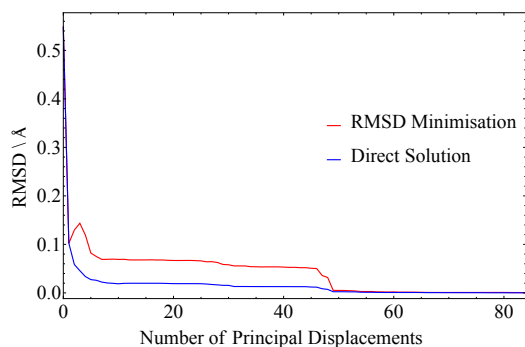
Figure B.1: The two molecules, HIBGUV and HAJYUN, from the test set of Thompson that were excluded from the study in Chapter 5 due to the presence of halogen atoms.



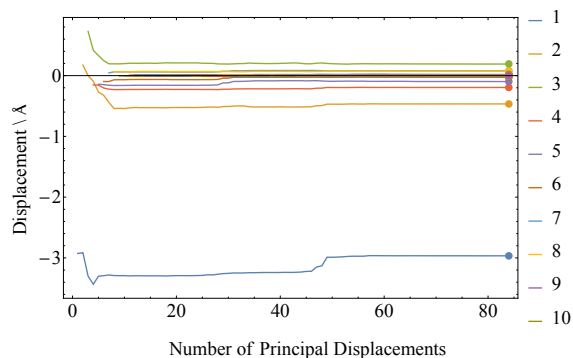
VEMTOW



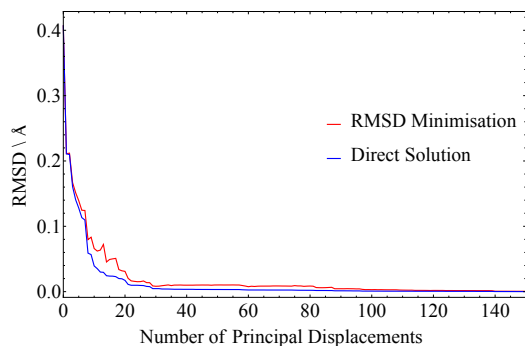
VEMTOW Displacements



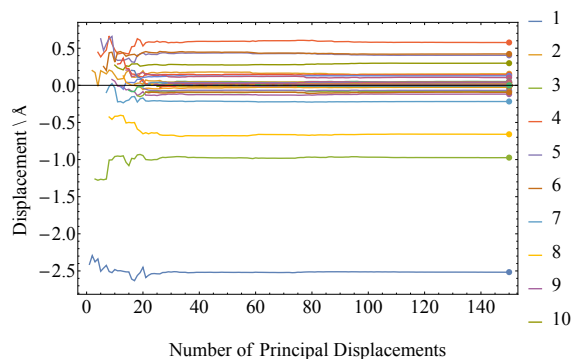
VEMTOW01



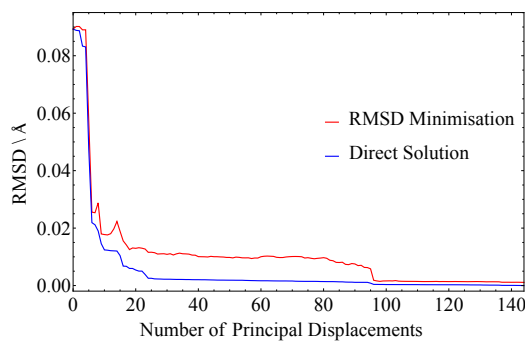
VEMTOW01 Displacements



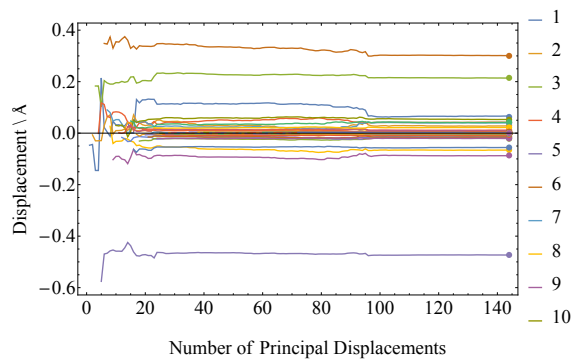
SIKRIN



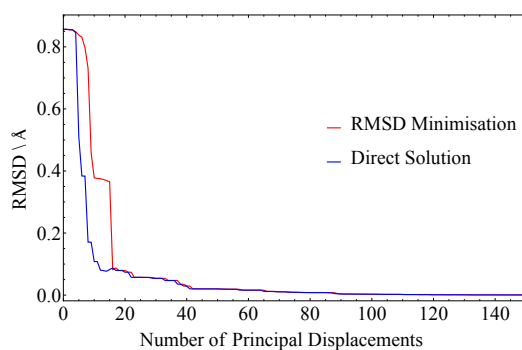
SIKRIN Displacements



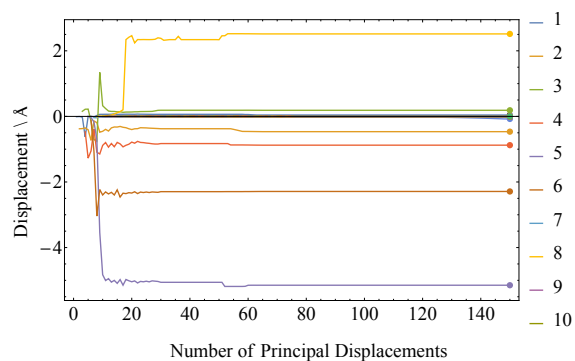
NEQNIG



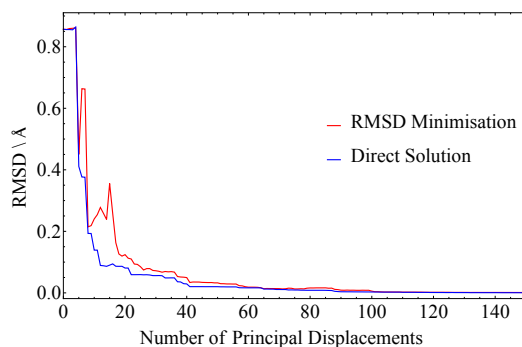
NEQNIG Displacements



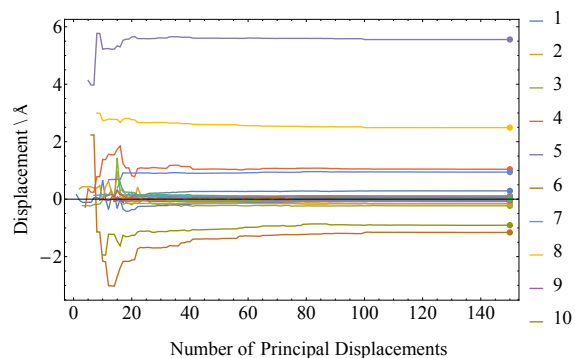
CELHIL



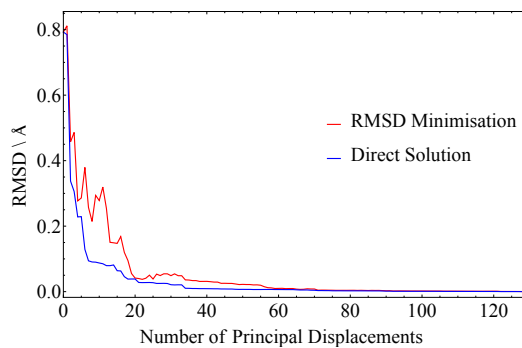
CELHIL Displacements



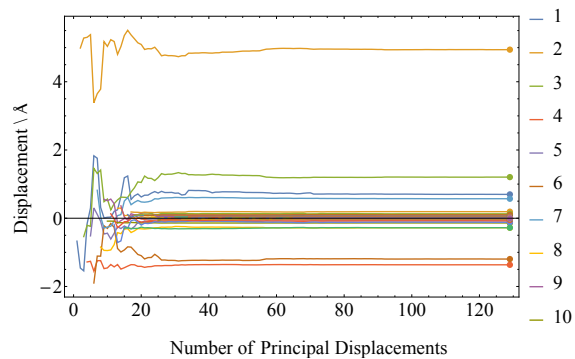
CELHIL01



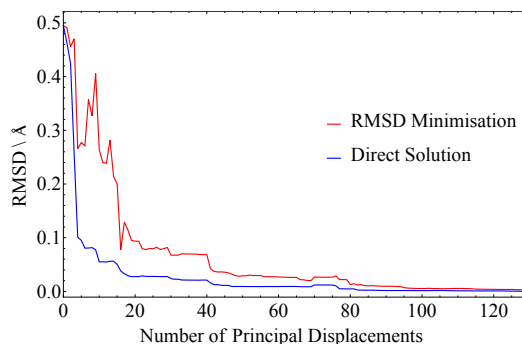
CELHIL01 Displacements



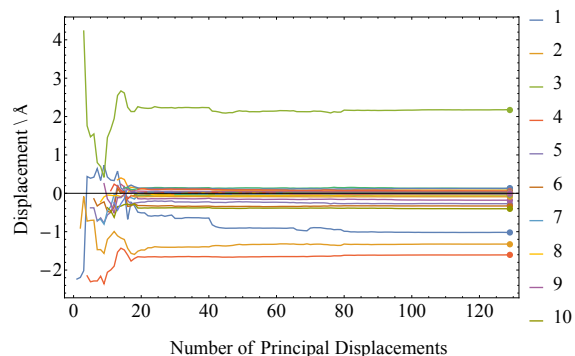
DANQEP



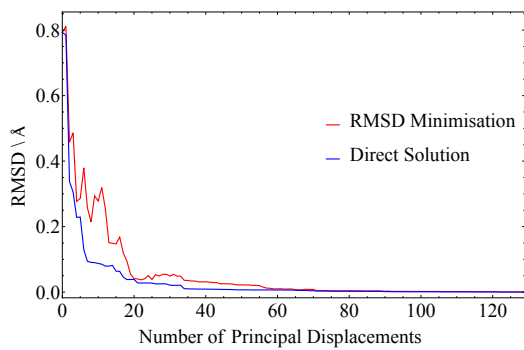
DANQEP Displacements



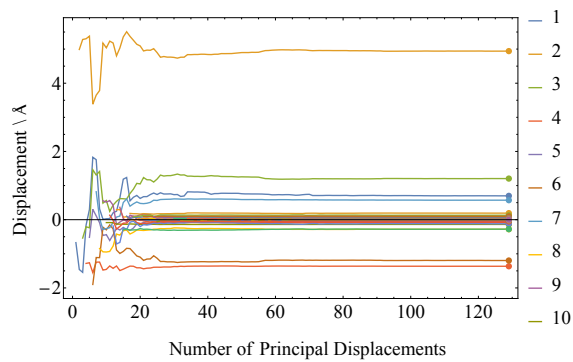
DANQEP01



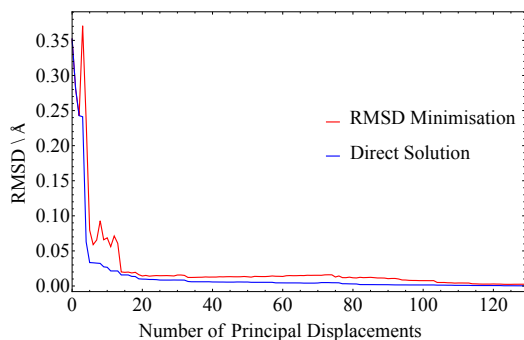
DANQEP01 Displacements



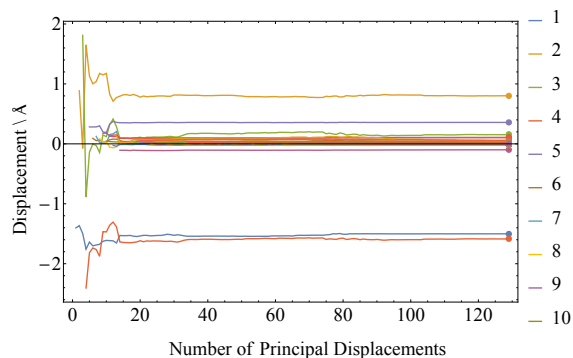
DANQEP



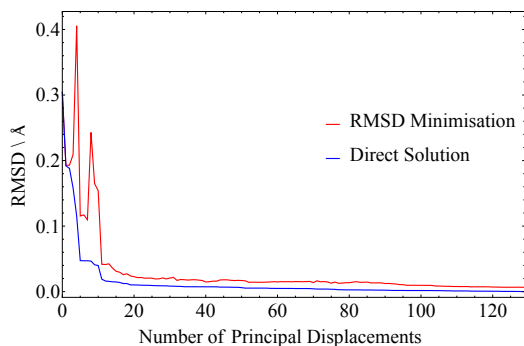
DANQEP Displacements



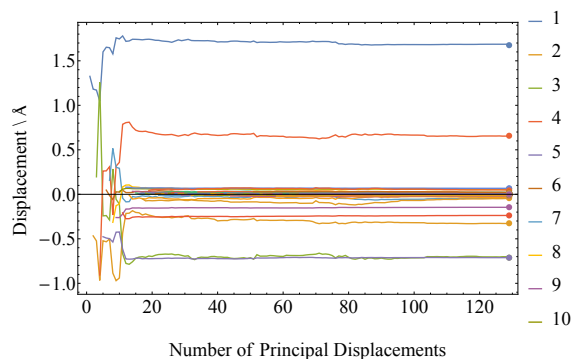
DANQEP02 (a)



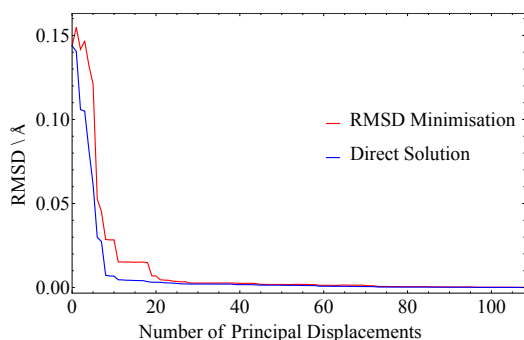
DANQEP02 (a) Displacements



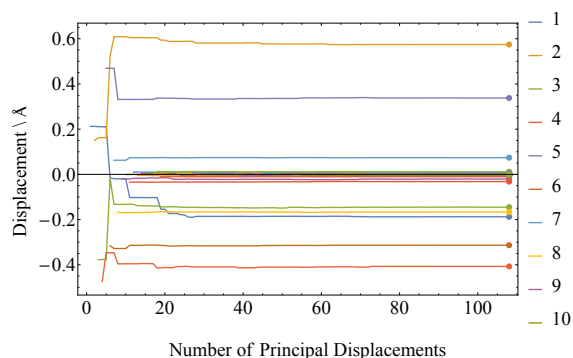
DANQEP02 (b)



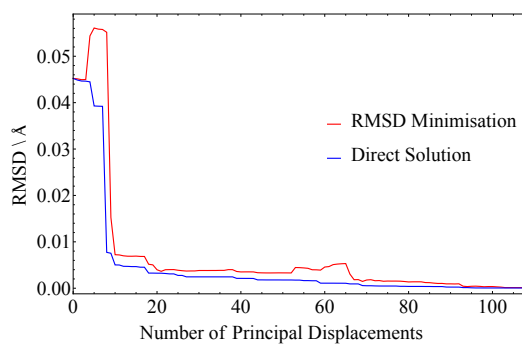
DANQEP02 (b) Displacements



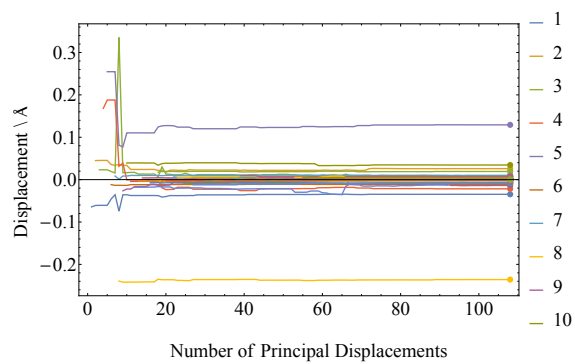
GALCAX



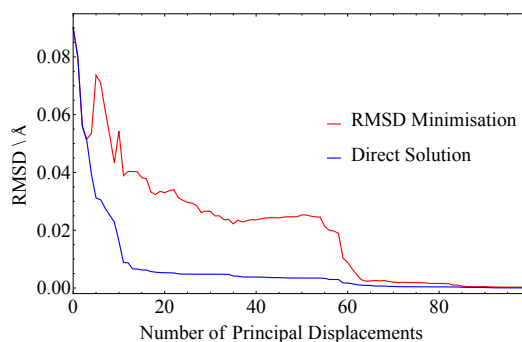
GALCAX Displacements



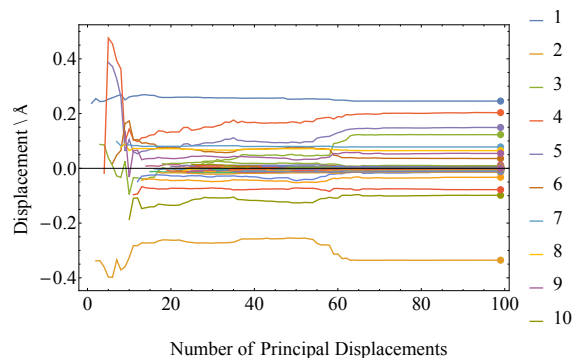
GALCAX01



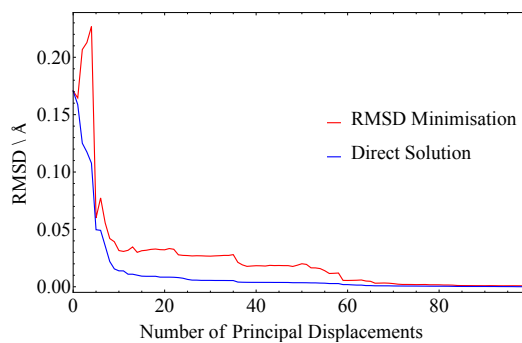
GALCAX01 Displacements



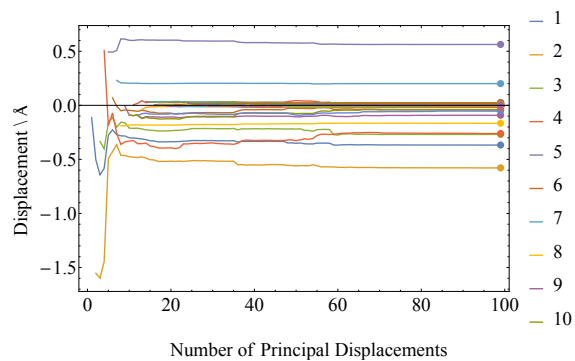
MABZNA (a)



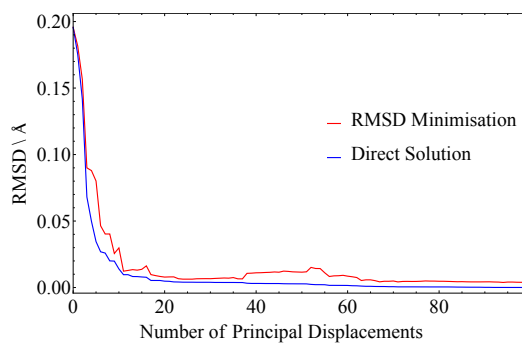
MABZNA (a) Displacements



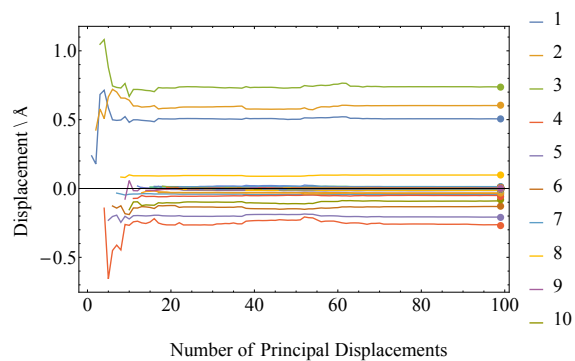
MABZNA (b)



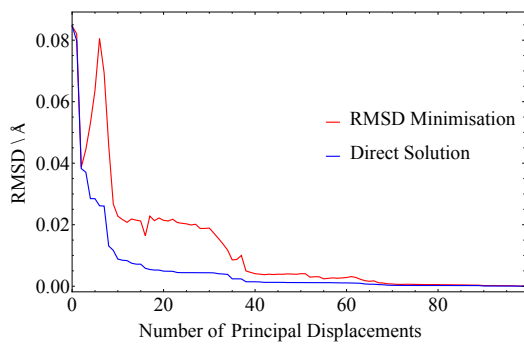
MABZNA (b) Displacements



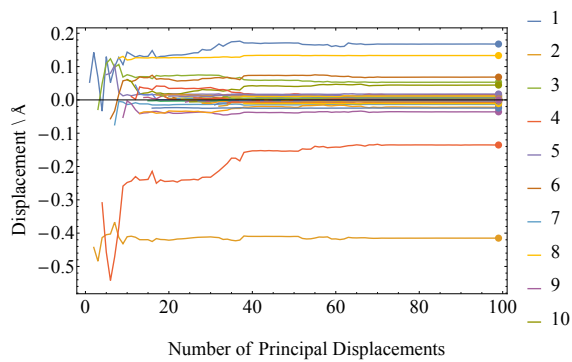
MABZNA01 (a)



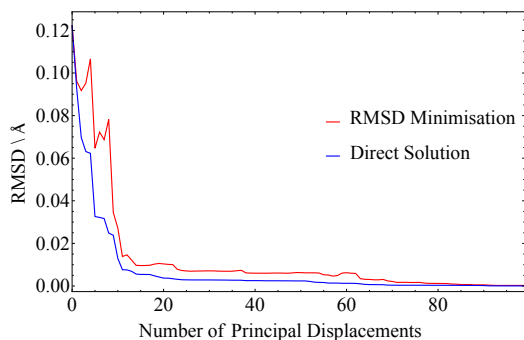
MABZNA01 (a) Displacements



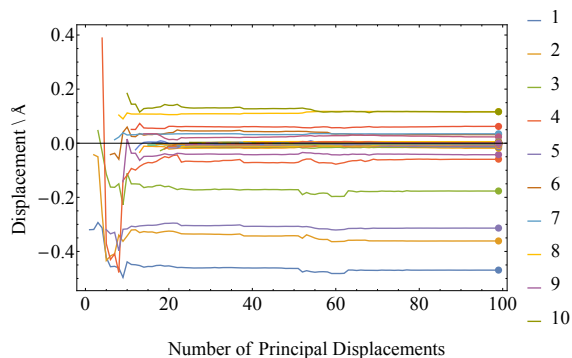
MABZNA01 (b)



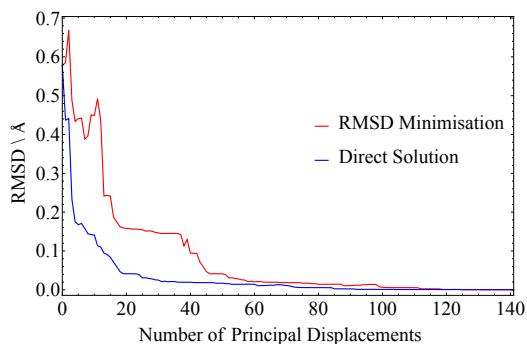
MABZNA01 (b) Displacements



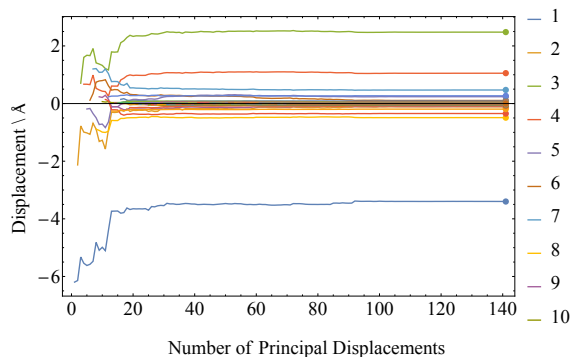
MABZNA02



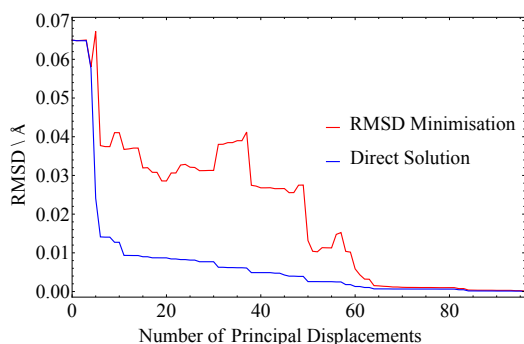
MABZNA02 Displacements



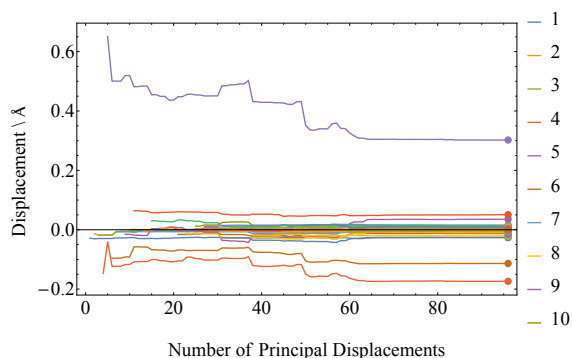
DADNUR



DADNUR Displacements



FAHNOR



FAHNOR Displacements

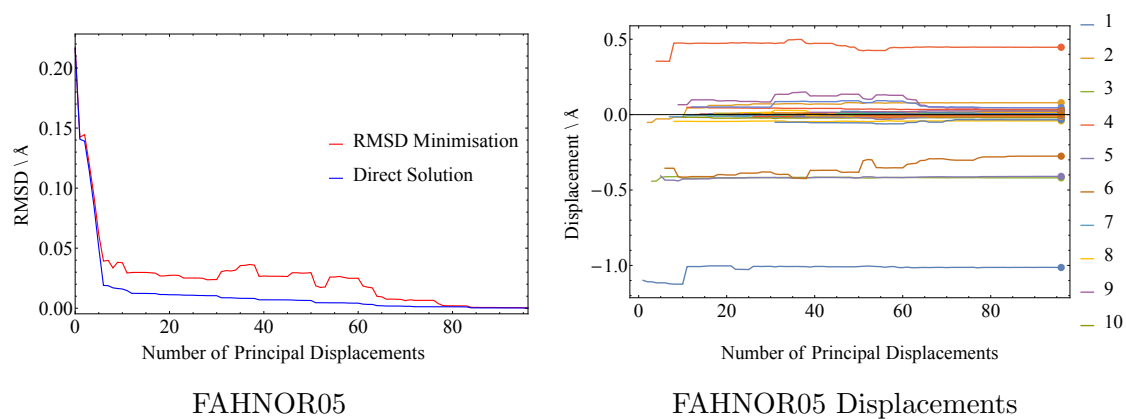


Figure B.2: The data for each molecule used in the test set. Each molecule possesses two graphs: the tracking of the RMSD when using the *RMSD Minimisation* and *Direct Solution* methodologies (left) and the displacement values derived for each principal displacement for both methodologies (right).

Appendix C

Decomposition of Molecular Strain in Crystals

This appendix contains the CSD reference codes along with the number of atoms and number of rotatable bonds for the 175 molecules used in the ΔE_{strain} distribution study in Chapter 6.

Table C.1: The 175 molecules listed in Chapter 6, along with the number of atoms and the number of rotatable bonds for each molecule. The molecules highlighted in red, green and grey are those that form non-covalent intramolecular hydrogen, polar and non-polar bonding interactions upon the geometry optimisation of the in-crystal geometry.

CSD RefCode	Number of Atoms	Number of Rotatable Bonds
ADUMAO	7	62
AFEWIR	7	53
AHUGAL	4	66
BIZXUE	7	42
BOKZUY	3	50
BOMCOV	6	45
BOMJIY	3	42
CELHIL	8	52
CELHIL01	8	52
CEWDEO	7	54
CIHDAY	3	50
COCAIN	5	43
Continued on the next page...		

CSD RefCode	Number of Atoms	Number of Rotatable Bonds
COFWID	6	56
COFXIF	7	52
COJXEE	6	61
COKQEZ	5	44
CUNTUA	7	52
DABMIC	8	58
DABZIQ	7	50
DADNUR	6	49
DANQEP	8	45
DANQEP01	8	45
DANQEP02 _a	8	45
DANQEP02 _b	8	45
DAZYUZ	5	51
DIWNAY	7	51
EJEXAT	3	55
FAHNOR	5	34
FAHNOR05	5	34
FIBKUW01	6	42
FIBKUW02	6	42
FIDLAF	6	52
FOLBEO	7	65
FOVMIN	3	59
FOWREP	2	58
FULPIL	7	68
FUWPAO	8	68
GADBOC	6	67
GADCAP	6	51
GALCAX	7	38
GALCAX01	7	38
GEJLEM	4	69
GENRAT	3	53
GEPQOI	3	65
GEZSOT	8	59
GISSUW	8	67
Continued on the next page...		

CSD RefCode	Number of Atoms	Number of Rotatable Bonds
GOCSUN	8	45
GODVUQ	7	47
GOLWEK	7	53
GUDMUN	6	58
HALSIW	2	55
HEKPUI	8	52
HEVDIW	6	62
HOHLIA	7	59
HOTPUC	7	54
HPSANA	7	56
HUHXIR	8	48
ICOQIB	5	58
IDODIO	7	53
IDUZAI	8	62
IFIJOX	4	63
IHUHIC	6	53
IJOZEM	6	57
IPBPNM	5	47
ISIKAW01	6	48
IXETIO	6	48
JIRDIX	4	52
JIRSIN	6	54
JIXRIS	8	58
KEBVAO	7	55
KEBVOC	4	51
KESYEN	6	49
KIFGUB	7	57
KIMSOO	8	56
LAGBEB	7	53
LAYQUZ	2	50
LOHNIF	7	54
MABZNA0	4	35
MABZNA01 _a	4	35
MABZNA01 _b	4	35
Continued on the next page...		

CSD RefCode	Number of Atoms	Number of Rotatable Bonds
MABZNA02	4	35
MABZNA1	4	35
MEQSAD	6	58
METXEP	8	60
MIYCIH	8	45
MODYEK	8	61
MOWBAB	6	64
MUBLEA	8	49
MUPSIK	7	55
NAJBW	8	54
NEBYAV	8	67
NELQEB	8	50
NEQNIG	6	50
NIQPEJ	7	53
NOKVEP	4	59
ODNPDS02	5	28
ODNPDS11	5	28
PABRAM	4	48
PILCUI	6	49
PUTQIE	6	53
PUYXOW	7	58
QIQTEP	8	58
QIRQAJ	8	65
QOFHID	7	49
QOSLUG	8	48
RAPTUY	4	46
ROQNAM	7	62
RUJTUL	6	51
RUSLOG	7	51
SAVXIW	7	66
SETTEQ	6	56
SEVJAF	7	64
SIBHOA	8	50
SIHHUM	7	50
Continued on the next page...		

CSD RefCode	Number of Atoms	Number of Rotatable Bonds
SIKRIN	4	40
SOKRAM	6	52
SOKREQ	7	50
SOPVUO	4	50
SOYQUT	3	56
SOZCEP	7	57
SURTHI	7	52
TABSEU	8	54
TBPEBA10	8	65
TIXTOK	6	48
TOCGOH	6	57
TOPTAT	7	20
VAHCAI	6	34
VAQXAM	6	31
VATWUK	3	45
VELZOB	6	58
VEMTOW	6	30
VEMTOW01	6	30
VEPJEF	8	39
VEPJOP	8	24
VIMXIY	7	35
VIPQER	7	27
VIXNAS	6	38
VOHYAT	8	36
VOLGOT	8	30
WAHSAA	8	31
WENTAL	5	60
WERVIY	5	56
WEWPOD01	5	54
WIFCUK	6	46
WIPXOK	6	54
WOJGEH	6	34
XAFVEG	6	51
XAPLOQ	7	59
Continued on the next page...		

CSD RefCode	Number of Atoms	Number of Rotatable Bonds
XEPRUG	8	55
XERDED	8	36
XIFWEP	7	40
XIQVOJ	6	31
XOFRIT	8	54
XOMLIV	6	53
XOXQEG	7	48
YAHREF	6	38
YIBXUD	3	47
YIRSAU	6	52
YOWZIT	8	32
YAXBAP	5	32
ZERMOY	2	20
ZUHWAA	5	64

Appendix D

Case Study II: Flufenacet

The convergence graphs for space groups $P\bar{1}$, $P2_1$, $P2_1/c$, $C2/c$, $P2_12_12_1$ and $Pbca$ for flufenacet conformers 2 and 3 in Figures D.1 and D.2, respectively. All are plotted against the global minimum crystal structure yielded from conformer 1.

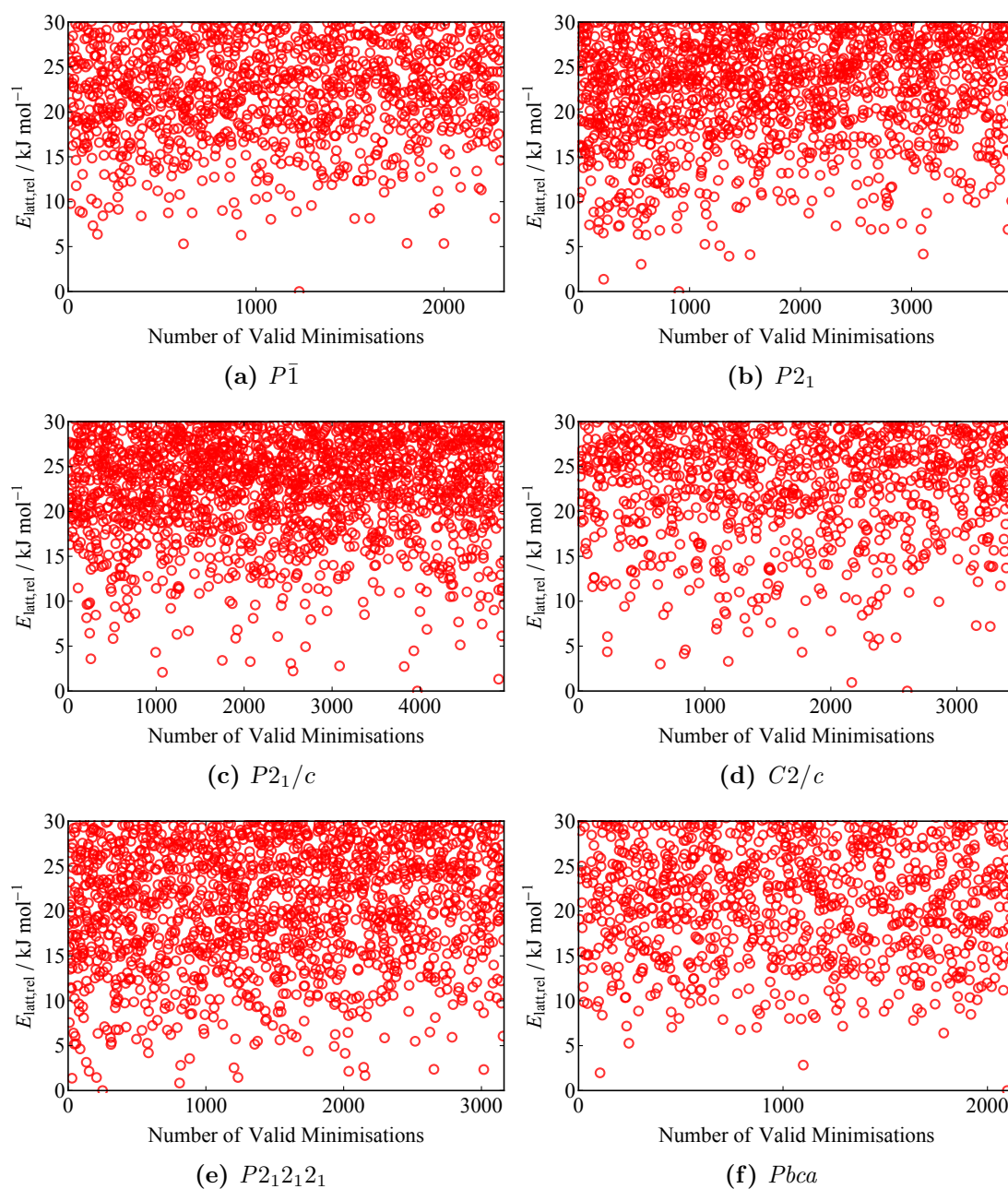


Figure D.1: All 6 plots show the number of unique crystal structures being generated for conformer 2 of flufenacet for a given total energy against the number valid minimisations. The figure captions indicate the space group.

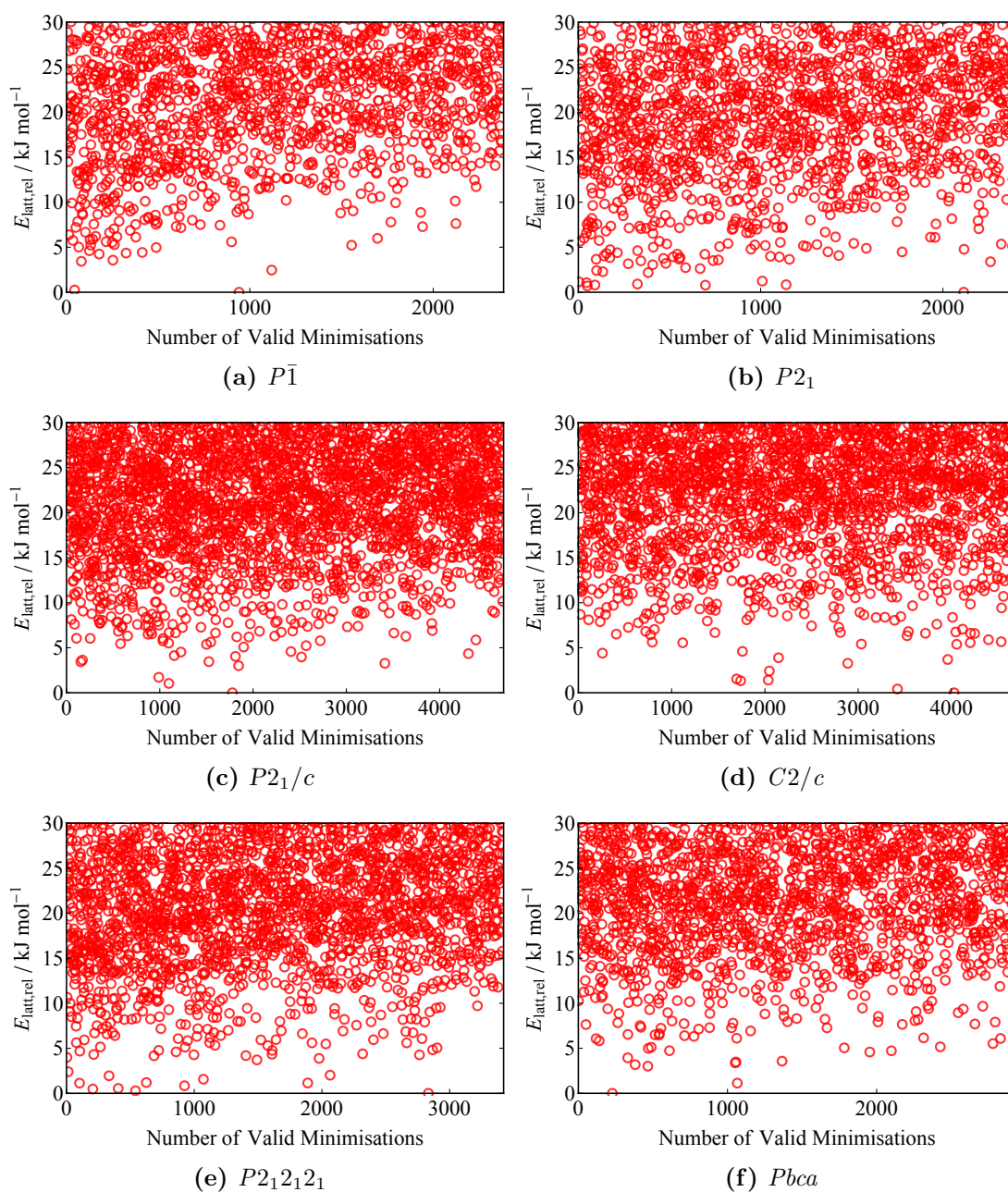


Figure D.2: All 6 plots show the number of unique crystal structures being generated for conformer 3 of flufenacet for a given total energy against the number valid minimisations. The figure captions indicate the space group.

Bibliography

- [1] L. Cano-Cortes, A. Dolfen, J. Merino, and E. Koch. Determination of screened coulomb repulsion energies in organic molecular crystals: A real space approach. *Physica B-Condensed Matter*, 405(11):S185–S187, 2010.
- [2] S. Fleutot, H. Martinez, J. C. Dupin, I. Baraille, C. Forano, G. Renaudin, and D. Gonbeau. Experimental (x-ray photoelectron spectroscopy) and theoretical studies of benzene based organics intercalated into layered double hydroxide. *Solid State Sciences*, 13(9):1676–1686, 2011.
- [3] M. M. Freund and F. T. Freund. Solid solution model for interstellar dust grains and their organics. *Astrophysical Journal*, 639(1):210–226, 2006.
- [4] S. S. Naghavi, M. Fabrizio, T. Qin, and E. Tosatti. Electron-doped organics: Charge-disproportionate insulators and hubbard-frohlich metals. *Physical Review B*, 88(11):8, 2013.
- [5] V. Stehr, R. F. Fink, B. Engels, J. Pflaum, and C. Deibel. Singlet exciton diffusion in organic crystals based on marcus transfer rates. *Journal of Chemical Theory and Computation*, 10(3):1242–1255, 2014.
- [6] J. Maddox. Crystals from first principles. *Nature*, 335(1):201, 1988.
- [7] A. Gavezzotti. Are crystal structures predictable? *Accounts of Chemical Research*, 27(10):309–314, 1994.
- [8] A. Gavezzotti. Calculation of intermolecular interaction energies by direct numerical integration over electron densities. i. electrostatic and polarization energies in molecular crystals. *The Journal of Physical Chemistry B*, 106(16):4145–4154, 2002.
- [9] C. Ouvrad and S. L. Price. Towards crystal structure prediction for conformationally flexible molecules: The headaches illustrated by aspirin. *Crystal Growth Design*, 4(6):1119–1127, 2004.

- [10] M. D. Gourlay, J. Kendrick, and F. J. J. Leusen. Rationalization of racemate resolution: Predicting spontaneous resolution through crystal structure prediction. *Crystal Growth & Design*, 7(1):56–63, 2007.
- [11] F. J. J. Leusen. Crystal structure prediction of diastereomeric salts: A step toward rationalization of racemate resolution. *Crystal Growth & Design*, 3(2):189–192, 2003.
- [12] P. G. Karamertzanis, A. V. Kazantsev, N. Issa, G. W. A. Welch, C. S. Adjiman, C. C. Pantelides, and S. L. Price. Can the formation of pharmaceutical cocrystals be computationally predicted? II. crystal structure prediction. *Journal of Chemical Theory and Computation*, 5(5):1432–1448, 2009.
- [13] A.J. Cruz-Cabeza, G. M. Day, W. D. S. Motherwell, and W. Jones. Prediction and observation of isostructurality induced by solvent incorporation in multicomponent crystals. *Journal of the American Chemical Society*, 128(45):14466–14467, 2006.
- [14] A. T. Hulme and S. L. Price. Toward the prediction of organic hydrate crystal structures. *Journal of Chemical Theory and Computation*, 3(4):1597–1608, 2007.
- [15] A. J. Cruz-Cabeza, S. Karki, L. Fabian, T. Friscic, G. M. Day, and W. Jones. Predicting stoichiometry and structure of solvates. *Chemical Communications*, 46(13):2224–2226, 2010.
- [16] G. M. Day. Current approaches to predicting molecular organic crystal structures. *Crystallography Reviews*, 17(1):3–52, 2011.
- [17] A. Gavezzotti. Ten years of experience in polymorph prediction: what next? *CrystEngComm*, 4(61):343–347, 2002.
- [18] G. M. Day, W. D. S. Motherwell, H. L. Ammon, S. X. M. Boerrigter, R. G. Della Valle, E. Venuti, A. Dzyabchenko, J. D. Dunitz, B. Schweizer, B. P. van Eijck, P. Erk, J. C. Facelli, V. E. Bazterra, M. B. Ferraro, D.W. M. Hofmann, F. J. J. Leusen, C. Liang, C. C. Pantelides, P. G. Karamertzanis, S. L. Price, T. C. Lewis, H. Nowell, A. Torrisi, H. A. Scheraga, Y. A. Arnautova, M. U. Schmidt, and P. Verwer. A third blind test of crystal structure prediction. *Acta Crystallographica Section B*, 61(5):511–527, 2005.
- [19] B. P. van Eijck, W. T. M. Mooij, and J. Kroon. Ab initio crystal structure predictions for flexible hydrogen-bonded molecules: Part II. accurate energy minimisation. *Journal of Computational Chemistry*, 22(8):805–815, 2001.
- [20] P. Vishweshwar, J. A. McMahon, M. Oliveira, M. L. Peterson, and M. J. Zaworotko. The predictably elusive form II of aspirin. *Journal of the American Chemical Society*, 127(48):16802–16803, 2005.

- [21] H. Nowell and S. L. Price. Validation of a search technique for crystal structure prediction of flexible molecules by application to piracetam. *Acta Crystallographica Section B*, B(61):558–568, 2005.
- [22] J. Bauer, S. Spanton, R. Henry, J. Quick, W. Dziki, W. Porter, and J. Morris. Ritonavir: An extraordinary example of conformational polymorphism. *Pharmaceutical Research*, 18(6):859–866, 2001.
- [23] L. Yu, G. A. Stephenson, C. A. Mitchell, C. A. Bunnell, S. V. Snorek, J. J. Bowyer, T. B. Borchardt, J. G. Stowell, and S. R. Byrn. Thermochemistry and conformational polymorphism of a hexamorphic crystal system. *Journal of the American Chemical Society*, 122(4):585–591, 2000.
- [24] C. A. Mitchell, L. Yu, and M. D. Ward. Selective nucleation and discovery of organic polymorphs through epitaxy with single crystal substrates. *Journal of the American Chemical Society*, 123(44):10830–10839, 2001.
- [25] C. Shuang, I. A. Guzei, , and L. Yu. New polymorphs of ROY and new record for coexisting polymorphs of solved structures. *Journal of the American Chemical Society*, 127(27):9881–9885, 2005.
- [26] M. Vasileiadis, A. V. Kazantsev, P. G. Karamertzanis, C. S. Adjiman, and C. C. Pantelides. The polymorphs of ROY: application of a systematic crystal structure prediction technique. *Acta Crystallographica Section B*, 68(6):677–685, 2012.
- [27] A. J. Cruz-Cabeza and J. Bernstein. Conformational polymorphism. *Chemical Reviews*, 114(4):2170–2191, 2013.
- [28] Wolfram Research Inc. Mathematica 10, 2014.
- [29] F. H. Allen. CSD - The Cambridge Structural Database: a quarter of a million crystal structures and rising. 58(3):380–388, 2002.
- [30] F. H. Allen and W. D. S. Motherwell. Applications of the cambridge structural database in organic chemistry and crystal chemistry. *Acta Crystallographica Section B*, 58(1):407–422, 2002.
- [31] I. J. Bruno, J. C. Cole, P. R. Edgington, M. Kessler, C. F. Macrae, P. McCabe, J. Pearson, and R. Taylor. Conquest - new software for searching the Cambridge Structural Database and visualising crystal structures. 58(3):389–397, 2002.
- [32] C. F. Macrae, I. J. Bruno, J. A. Chisholm, P. R. Edgington, P. McCabe, E. Pidcock, L. Rodriguez-Monge, R. Taylor, J. van de Streek, and P. A. Wood. Mercury CSD 2.0 - new features for the visualization and investigation of crystal structures. *Journal of Applied Crystallography*, 41(1):466–470, 2008.

- [33] Pitzer. *Discussions of the Faraday Society*, 107:4519–4529, 1951.
- [34] J. W. Ochterski. *Vibrational analysis in gaussian*. 1999.
- [35] D. J. Willock, S. L. Price, M. Leslie, and C. R. A. Catlow. The relaxation of molecular crystal structures using a distributed multipole electrostatic model. *Journal of Computational Chemistry*, 16(5):628–647, 1995.
- [36] J. Baker, A. Kessi, and B. Delly. The generation and use of delocalized internal coordinates in geometry optimization. *The Journal of Chemical Physics*, 105(1):192–212, 1996.
- [37] P. Pulay and G. Fogarasi. Geometry optimization on redundant internal coordinates. *The Journal of Chemical Physics*, 96(4):2856–2860, 1992.
- [38] J. Reimers. A practical method for the use of curvilinear coordinates in calculations of normal-mode-projected displacements and duschinsky rotation matrices for large molecules. *The Journal of Chemical Physics*, 115(20):9103–9109, 2001.
- [39] E. B. Wilson, J. C. Decius, and P. C. Cross. *Molecular Vibrations*. Dover Books, 1st edition, 1955.
- [40] G. M. Day, W. D. S. Motherwell, and W. Jones. A strategy for predicting the crystal structures of flexible molecules: the polymorphism of phenobarbital. *Physical Chemistry Chemical Physics*, 9(14):1693–1704, 2007.
- [41] T. G. Cooper, W. D. S. Motherwell, and G. M. Day. Database guided conformational selection in crystal structure prediction of alanine. *CrystEngComm*, 9(7):595–602, 2007.
- [42] G. M. Day and T. G. Cooper. Crystal packing predictions of the alpha-amino acids: methods assessment and structural observations. *CrystEngComm*, 12(8):2443–2453, 2010.
- [43] B. Görbitz, B. Dalhus, and G. M. Day. Pseudoracemic amino acid complexes: blind predictions for flexible two-component crystals. *Physical Chemistry Chemical Physics*, 12(30):8466–8477, 2010.
- [44] A. J. Kazantsev, P. G. Karamertzanis, C. S. Adjiman, C. C. Pantelides, S. L. Price, P. T. A. Galek, G. M. Day, and A. J. Cruz-Cabeza. Successful prediction of a model pharmaceutical in the fifth blind test of crystal structure prediction. *International Journal of Pharmaceutics*, 418(2):168–178, 2011.
- [45] R. E. Bellman. *Dynamic Programming*. Princeton University Press, 1st edition, 1957.

- [46] Schrodinger LLC. Macromodel, 2011.
- [47] I. Kolossvary and W. C. Guida. Low mode search. an efficient, automated computational method for conformational analysis: Application to cyclic and acyclic alkanes and cyclic peptides. *Journal of the American Chemical Society*, 118(21):5011–5019, 1996.
- [48] I. Kolossvary and W. C. Guida. Low-mode conformational search elucidated: application to C39H80 and flexible docking of 9-deazaguanine inhibitors into PNP. *Journal of Computational Chemistry*, 20(15):1671–1684, 1999.
- [49] W. L. Jorgensen and J. Tirado-Rives. The OPLS force field for proteins. energy minimizations for crystals of cyclic peptides and crambin. *Journal of the American Chemical Society*, 110(6):1657–1666, 1988.
- [50] W. L. Jorgensen, D. S. Maxwell, and J. Tirado-Rives. Development and testing of the OPLS all-atom force field on conformational energetics and properties of organic liquids. *Journal of the American Chemical Society*, 118(45):11225–11236, 1996.
- [51] W. D. Cornell, P. Cieplak, Bayly C. I., I. R. Gould, K. M. Merz Jr., D. M. Ferguson, D. C. Spellmeyer, T. Fox, J. W. Caldwell, and P. A. Kollman. Second generation force field for the simulation of proteins, nucleic acids, and organic molecules. *Journal of the American Chemical Society*, 117(19):5179–5197, 1995.
- [52] J. N. Israelachvili. *Intermolecular and Surface Forces*. San Diego: Academic Press, 1st edition, 1992.
- [53] T. Beyer and S. L. Price. The errors in lattice energy minimisation studies: Sensitivity to experimental variations in the molecular structure of paracetamol. *CrystEngComm*, 2(34):183–190, 2000.
- [54] J. Patterson and B. Bailey. *Solid-State Physics: Introduction to the Theory*. Springer, 2nd edition, 2010.
- [55] A. D. Becke. Perspective: Fifty years of density-functional theory in chemical physics. *The Journal of Chemical Physics*, 140(18), 2014.
- [56] W. Kohn and L. J. Sham. Self-consistent equations including exchange and correlation effects. *Physics Review*, 140(4A):1133–1138, 1965.
- [57] P. Soderhjelm, G. Karlström, and U. Ryde. Comparison of overlap-based models for approximating the exchange-repulsion energy. *The Journal of Chemical Physics*, 124(24):244101, 2006.

- [58] J. P. Perdew, A. Ruzinszky, J. Tao, V. K. Staroverov, G. E. Scuseria, and G. I. Csonka. Prescription for the design and selection of density functional approximations: More constraint satisfaction with fewer fit. *Journal of Chemical Physics*, 123(6):062201–062209, 2005.
- [59] J. P. Perdew, Chevary J. A., Vosko S. H., K. A. Jackson, M. R. Pederson, D. J. Singh, and C. Fiolhais. Atoms, molecules, solids, and surfaces: Applications of the generalized gradient approximation for exchange and correlation. *Physical Review B*, 46(11):6671–6687, 1992.
- [60] D. C. Langreth and Mehl M. J. Beyond the local-density approximation in calculations of ground-state electronic properties. *Physical Review B*, 28(4):1809–1834, 1983.
- [61] A. J. Cohen, P. Mori-Sánchez, and W. Yang. Challenges for density functional theory. *Chemical Reviews*, 112(1):289–320, 2012.
- [62] N. Mardirossian and M. Head-Gordon. Characterizing and understanding the remarkably slow basis set convergence of several minnesota density functionals for intermolecular interaction energies. *Journal of Chemical Theory and Computation*, 9(1):4453–4461, 2013.
- [63] A. D. Becke. Density-functional exchange-energy approximation with correct asymptotic-behavior. *Physical Review A*, 38(6):3098–3100, 1988.
- [64] C. T. Lee, W. T. Yang, and R. G. Parr. Development of the colle-salvetti correlation-energy formula into a functional of the electron-density. *Physical Review B*, 37(2):785–789, 1988.
- [65] A. M. Reilly, R. I. Cooper, C. S. Adjiman, S. Bhattacharya, A. D. Boese, J. G. Brandenburg, P. J. Bygrave, R. Bylsma, J. E. Campbell, R. Car, D. H. Case, R. Chadha, J. C. Cole, K. Cosburn, H. M. Cuppen, F. Curtis, G. M. Day, R. A. DiStasio Jr, A. Dzyabchenko, B. P. van Eijck, D. M. Elking, J. A. van den Ende, J. C. Facelli, M. B. Ferraro, L. Fusti-Molnar, C.-A. Gatsiou, T. S. Gee, R. de Gelder, L. M. Ghiringhelli, H. Goto, S. Grimme, R. Guo, D. W. M. Hofmann, J. Hoja, R. K. Hylton, L. Iuzzolino, W. Jankiewicz, D. T. de Jong, J. Kendrick, N. J. J. de Klerk, H.-Y. Ko, L. N. Kuleshova, X. Li, S. Lohani, F. J. J. Leusen, A. M. Lund, J. Lv, Y. Ma, N. Marom, A. E. Masunov, P. McCabe, D. P. McMahon, H. Meekes, M. P. Metz, A. J. Misquitta, S. Mohamed, B. Monserrat, R. J. Needs, M. A. Neumann, J. Nyman, S. Obata, H. Oberhofer, A. R. Oganov, A. M. Orendt, G. I. Pagola, C. C. Pantelides, C. J. Pickard, R. Podeszwa, L. S. Price, S. L. Price, A. Pulido, M. G. Read, K. Reuter, E. Schneider, C. Schober, G. P.

- Shields, P. Singh, I. J. Sugden, K. Szalewicz, C. R. Taylor, A. Tkatchenko, M. E. Tuckerman, F. Vacarro, M. Vasileiadis, A. Vazquez-Mayagoitia, L. Vogt, Y. Wang, R. E. Watson, G. A. de Wijs, J. Yang, Q. Zhu, and C. R. Groom. Report on the sixth blind test of organic crystal structure prediction methods. *Acta Crystallographica Section B*, 72(4):439–459, 2016.
- [66] F. Jensen. *Introduction to Computational Chemistry*. John Wiley and Sons Ltd., 2nd edition, 2007.
- [67] M. J. Frisch, G. W. Trucks, H. B. Schlegel, G. E. Scuseria, M. A. Robb, J. R. Cheeseman, G. Scalmani, V. Barone, B. Mennucci, G. A. Petersson, H. Nakatsuji, M. Caricato, X. Li, H. P. Hratchian, A. F. Izmaylov, J. Bloino, G. Zheng, J. L. Sonnenberg, M. Hada, M. Ehara, K. Toyota, R. Fukuda, J. Hasegawa, M. Ishida, T. Nakajima, Y. Honda, O. Kitao, H. Nakai, T. Vreven, J. A. Montgomery Jr., J. E. Peralta, F. Ogliaro, M. Bearpark, J. J. Heyd, E. Brothers, K. N. Kudin, V. N. Staroverov, T. Keith, R. Kobayashi, J. Normand, K. Raghavachari, A. Rendell, J. C. Burant, S. S. Iyengar, J. Tomasi, M. Cossi, N. Rega, J. M. Millam, M. Klene, J. E. Knox, J. B. Cross, V. Bakken, C. Adamo, J. Jaramillo, R. Gomperts, R. E. Stratmann, O. Yazyev, A. J. Austin, R. Cammi, C. Pomelli, J. W. Ochterski, R. L. Martin, K. Morokuma, V. G. Zakrzewski, G. A. Voth, P. Salvador, J. J. Dannenberg, S. Dapprich, A. D. Daniels, O. Farkas, J. B. Foresman, J. V. Ortiz, J. Cioslowski, and D. J. Fox. Gaussian 09, 2013.
- [68] R. Dovesi, R. Orlando, B. Civalleri, C. Roetti, V. R. Saunders, and C. M. Zicovich-Wilson. CRYSTAL09. *Zeitschrift für Kristallographie*, 220(1):571–573, 2005.
- [69] H.-J. Werner, P. J. Knowles, G. Knizia, F. R. Manby, and M. Schütz. Molpro: a general purpose quantum chemistry program package. *WIREs: Computational Molecular Science*, 2(1):242–253, 2012.
- [70] J. R. Shewchuk. *An Introduction to the Conjugate Gradient Method without the Agonizing Pain*. Carnegie Mellon University Pittsburgh, 1st edition, 1994.
- [71] M. J. D. Powell. An efficient method for finding the minimum of a function of several variables without calculating derivatives. *Computer Journal*, 7(2):155–162, 1964.
- [72] I. Maros. *Computational Techniques of the Simplex Method*. Kluwer Academic Publishers, 1st edition, 2003.
- [73] D. E. Williams. Improved intermolecular force field for crystalline hydrocarbons containing four- or three-coordinated carbon. *Journal of Molecular Structure*, 485-486:321–347, 1999.

- [74] A. J. Stone and M. Alderton. Distributed multipole analysis. *Molecular Physics*, 56(5):1047–1064, 1985.
- [75] Harmoniki.png. <https://en.wikipedia.org/wiki/File:Harmoniki.png>. Accessed: 2016-06-20.
- [76] S. L. Price, M. Leslie, G. W. A. Welch, M. Habgood, L. S. Price, P. G. Karamertzanis, and G. M. Day. Modelling organic crystal structures using distributed multipole and polarizability-based model intermolecular potentials. *Physical Chemistry Chemical Physics*, 12(30):8478–8490, 2010.
- [77] S. Brodersen, S. Wilke, F. J. J. Leusen, and G. Engel. A study of different approaches to the electrostatic interaction in force field methods for organic crystals. *Physical Chemistry Chemical Physics*, 5(21):4923–4931, 2003.
- [78] G. M. Day, W. D. S. Motherwell, and W. Jones. Beyond the isotropic atom model in crystal structure prediction of rigid molecules: Atomic multipoles versus point charges. *Crystal Growth & Design*, 5(3):1023–1033, 2004.
- [79] A. D. Buckingham and P. W. Fowler. A model for the geometries of van der waals complexes. *Canadian Journal of Chemistry*, 63(7):2018–2025, 1985.
- [80] A. D. Buckingham, P. W. Fowler, and A. J. Stone. Electrostatic predictions of shapes and properties of van der waals molecules. *International Reviews in Physical Chemistry*, 5(2-3):107–114, 1986.
- [81] G. M. Day, S. L. Price, and M. Leslie. Elastic constant calculations for molecular organic. *Crystal Growth & Design*, 1(1):13–27, 2000.
- [82] J. Nyman, O. S. Pundyke, and G. M. Day. Accurate force fields and methods for modelling organic molecular crystals at finite temperatures. *Physical Chemistry Chemical Physics*, 18(23):15828–15837, 2016.
- [83] A. J. Stone. GDMA: A program for performing distributed multipole analysis of wave functions calculated using the gaussian program system [1.0], 1999.
- [84] C. P. Brock and J. D. Dunitz. Towards a grammar of crystal packing. *Chemistry of Materials*, 6(8):1118–1127, 1994.
- [85] M. Walter. *Polymorphism in Physics and Chemistry of the Organic Solid State*. Wiley Interscience, 1st edition, 1965.
- [86] J. Bernstein. *Polymorphism in Molecular Crystals*. Oxford University Press, 1st edition, 2008.

- [87] A. I. Kitiagorodsky. *Molecular Crystals and Molecules*. Academic Press Inc. (London) Ltd., 1st edition, 1973.
- [88] S. Chen, I. A. Guzei, and L. Yu. New polymorphs of ROY and new record for coexisting polymorphs of solved structures. *Journal of the American Chemical Society*, 127(27):9881–9885, 2005.
- [89] S. L. Price. Computed crystal energy landscapes for understanding and predicting organic crystal structures and polymorphism. *Accounts of Chemical Research*, 42(1):117–126, 2009.
- [90] S. L. Price. Why don’t we find more polymorphs? *Acta Crystallographica Section B*, 69(1):313–328, 2013.
- [91] W. Kabsch. A solution for the best rotation to relate two sets of vectors. *Acta Crystallographica Section A*, 32(5):922–923, 1976.
- [92] B. Hayes. Quasirandom ramblings. *American Scientist*, 99(4):282–287, 2011.
- [93] I.M. Sobol. On the distribution of points in a cube and the approximate evaluation of integrals. *USSR Computational Mathematics and Mathematical Physics*, 7(4):86–112, 1967.
- [94] D. H. Case, J. E. Campbell, P. J. Bygrave, and G. M. Day. Convergence properties of crystal structure prediction by quasi-random sampling. *Journal of Chemical Theory and Computation*, 12(2):910–924, 2016.
- [95] N. Metropolis, A.W. Rosenbluth, M.N. Rosenbluth, A.H. Teller, and E. Teller. Equation of state calculation by fast computing machines. *Journal of Chemical Physics*, 21(6):1087–1092, 1953.
- [96] D. J. Wales and J. P. K. Doye. Global optimization by basin-hopping and the lowest energy structures of Lennard-Jones clusters containing up to 110 atoms. *Journal of Physical Chemistry A*, 101(28):5111–5116, 1997.
- [97] P. G. Karamertzanis and C. C. Pantelides. Ab initio crystal structure prediction—I. rigid molecules. *Journal of Computational Chemistry*, 26(3):304–324, 2005.
- [98] P. G. Karamertzanis and C. C. Pantelides. Ab initio crystal structure prediction. II. flexible molecules. *Molecular Physics*, 105(2-3):273–291, 2007.
- [99] C. Hermite. Sur un nouveau developpement en serie de fonctions. *Comptes Rendus de l’Academie des Sciences*, 58:93–100, 1864.

- [100] D. A. Bardwell, C. S. Adjiman, Y. A. Arnautova, E. Bartashevich, S. X. M. Boerrigter, D. E. Braun, A. J. Cruz-Cabeza, G. M. Day, R. G. Della Valle, G. R. Desiraju, B. P. van Eijck, J. C. Facelli, M. B. Ferraro, D. Grillo, M. Habgood, D. W. M. Hofmann, F. Hofmann, K. V. J. Jose, P. G. Karamertzanis, A. V. Kazantsev, J. Kendrick, L. N. Kuleshova, F. J. J. Leusen, A. V. Maleev, A. J. Misquitta, S. Mohamed, R. J. Needs, M. A. Neumann, D. Nikylov, A. M. Orendt, R. Pal, C. C. Pantelides, C. J. Pickard, L. S. Price, S. L. Price, H. A. Scheraga, J. van de Streek, T. S. Thakur, S. Tiwari, E. Venuti, and I. K. Zhitkov. Towards crystal structure prediction of complex organic compounds – a report on the fifth blind test. *Acta Crystallographica Section B*, 67(535-551), 2011.
- [101] Accelrys Inc. Cerius2, 1997.
- [102] M. Le Gendre. Recherches sur l’attraction des spheroides homogenes. *Memoires de Mathematiques et de Physique, presentes a l’Academie Royale des Sciences, par divers savans, et lus dans ses Assemblies*, pages 411–435, 1785.
- [103] D. J. Struik. *A Source Book in Mathematics 1200–1800*. Harvard University Press, 1st edition, 1969.
- [104] A. Stroud, K. *Further Engineering Mathematics*. Macmillan Press Ltd, 3rd edition, 1996.
- [105] S. L. Price. Quantifying intermolecular interactions and their use in computational crystal structure prediction. *CrystEngComm*, 6(61):344–353, 2004.
- [106] P. Ewald. Die berechnung optischer und elektrostatischer gitterpotentiale. *Annalen der Physik*, 369(3):253–287, 1921.
- [107] A. J. Stone. *The Theory of Intermolecular Forces*. Oxford University Press, 2nd edition, 2013.
- [108] S. Grimme, S. Ehrlich, and L. Goerigk. Effect of the damping function in dispersion corrected density functional theory. *Journal of Computational Chemistry*, 32(7): 1456–1465, 2011.
- [109] B. Civalleri, C. M. Zicovich-Wilson, L. Valenzano, and P. Ugliengo. B3LYP augmented with an empirical dispersion term (B3LYP-D*) as applied to molecular crystals. *CrystEngComm*, 10(4):405–410, 2008.
- [110] J. E. Jones. On the determination of molecular fields. II. from the equation of state of a gas. *Proceedings of the Royal Society of London. Series A*, 106(738): 463–477, 1924.

- [111] R. A. Buckingham. The classical equation of state of gaseous helium, neon and argon. *Proceedings of the Royal Society of London. Series A. Mathematical and Physical Sciences*, 168(933):264–283, 1938.
- [112] M. A. Neumann. Tailor-made force fields for crystal-structure prediction. *Journal of Physical Chemistry*, 112(32):9810–9829, 2008.
- [113] D. E. Williams. Improved intermolecular force field for crystalline oxohydrocarbons including O—H . . O hydrogen bonding. *Journal of Computational Chemistry*, 22(1):1–20, 2000.
- [114] D. E. Williams. Improved intermolecular force field for molecules containing H, C, N, and O atoms, with application to nucleoside and peptide crystals. *Journal of Computational Chemistry*, 22(11):1154–1166, 2001.
- [115] A. Abraha and D. E. Williams. Spherical and aspherical intermolecular force fields for sulfur allotropes. *Inorganic Chemistry*, 38(19):4224–4228, 1999.
- [116] D. E. Williams and D. J. Houpt. Fluorine nonbonded potential parameters derived from crystalline perfluorocarbons. *Acta Crystallographica Section B*, 46(3):286–295, 1986.
- [117] D. E. Williams and T. R. Stouch. Characterization of force fields for lipid molecules: Applications to crystal structures. *Journal of Computational Chemistry*, 14(9):1066–1076, 1993.
- [118] E. O. Pyzer-Knapp, H. P. G. Thompson, and G. M. Day. An optimized intermolecular force field for hydrogen bonded organic molecular crystals using atomic multipole electrostatics. *Acta Crystallographica Section B*, 2(34):183–190, 2016.
- [119] A. L. Hickey and C. N. Rowley. Benchmarking quantum chemical methods for the calculation of molecular dipole moments and polarizabilities. *The Journal of Physical Chemistry A*, 118(20):3678–3687, 2014.
- [120] H. P. G. Thompson and G. M. Day. Which conformations make stable crystal structures? Mapping crystalline molecular geometries to the conformational energy landscape. *Chemical Science*, 5(8):3173–3182, 2014.
- [121] M. L. Connolly. Analytical molecular surface calculation. *Journal of Applied Crystallography*, 16(5):548–558, 1983.
- [122] B. P. van Eijck and J. Kroon. Structure predictions allowing more than one molecule in the asymmetric unit. *Acta Crystallographica Section B*, 56(3):535–542, 2000.

- [123] W. T. M. Mooij, B. P. van Eijck, and J. Kroon. Ab initio crystal structure predictions for flexible hydrogen-bonded molecules. *Journal of the American Chemical Society*, 122(14):3500–3505, 2000.
- [124] P. G. Karamertzanis and S. L. Price. Energy minimization of crystal structures containing flexible molecules. *Journal of Chemical Theory and Computation*, 2(4):1184–1199, 2006.
- [125] M. J. Frisch, G. W. Trucks, H. B. Schlegel, G. E. Scuseria, M. A. Robb, J. R. Cheeseman, V. G. Zakrzewski, J. A. Montgomery Jr., R. E. Stratmann, J. C. Burant, S. Dapprich, J. M. Millam, A. D. Daniels, M. C. Kudin, K. N. andStrain, O. Farkas, J. Tomasi, V. Barone, M. Cossi, R. Cammi, B. Mennucci, C. Pomelli, S. Adamo, C. andClifford, J. Ochterski, G. A. Petersson, P. Y. Ayala, Q. Cui, K. Morokuma, D. K. Malick, K. Rabuck, A. D. andRaghavachari, J. B. Foresman, J. Cioslowski, J. V. Ortiz, A. G. Baboul, B. B. Stefanov, G. Liu, A. Liashenko, P. Piskorz, I. Komaromi, R. Gomperts, R. L. Martin, D. J. Fox, T. Keith, M. A. Al-Laham, C. Y. Peng, A. Nanayakkara, C. Gonzalez, M. Challacombe, P. M. W. Gill, B. G. Johnson, W. Chen, M. W. Wong, J. L. Andres, M. Head-Gordon, E. S. Replogle, and J. A. Pople. Gaussian 98, 1998.
- [126] A. V. Kazantsev, P. G. Karamertzanis, C. S. Adjiman, and C. C. Pantelides. Efficient handling of molecular flexibility in lattice energy minimization of organic crystals. *Journal of Chemical Theory and Computation*, 7(6):1998–2016, 2011.
- [127] J. E. Dennis and B. Schanabel, R. *Numerical Methods for Unconstrained Optimization and Nonlinear Equations*. Prentice-Hall, 1st edition, 1983.
- [128] J. A. Chisholm and W. D. S. Motherwell. COMPACK: a program for identifying crystal structure similarity using distances. *Journal of Applied Crystallography*, 38(1):228–231, 2005.
- [129] I. Krivý and B. Gruber. A unified algorithm for determining the reduced (Niggli) cell. *Acta Crystallographica Section A*, 32(2):297–298, 1976.
- [130] T. G. Cooper, K. E. Hejczyk, W. Jones, and G. M. Day. Molecular polarization effects on the relative energies of the real and putative crystal structures of valine. *Journal of Chemical Theory and Computation*, 4(10):1795–1805, 2008.
- [131] J. Nyman and G. M. Day. Static and lattice vibrational energy differences between polymorphs. *CrystEngComm*, 17(28):5154–5165, 2015.
- [132] W. Kabsch. A discussion of the solution for the best rotation to relate two sets of vectors. *Acta Crystallographica Section A*, 34(5):827–828, 1978.

- [133] J. P. M. Lommerse, W. D. S. Motherwell, H. L. Ammon, J. D. Dunitz, A. Gavezzotti, D. W. M. Hofmann, F. J. J. Leusen, W. T. M. Mooij, S. L. Price, B. Schweizer, M. U. Schmidt, B. P. van Eijck, P. Verwer, and D. E. Williams. A test of crystal structure prediction of small organic molecules. *Acta Crystallographica Section B*, 56(4):697–714, 2000.
- [134] W. D. S. Motherwell, H. L. Ammon, J. D. Dunitz, A. Dzyabchenko, P. Erk, A. Gavezzotti, D. W. M. Hofmann, F. J. J. Leusen, J. P. M. Lommerse, W. T. M. Mooij, S. L. Price, H. Scheraga, B. Schweizer, M. U. Schmidt, B. P. van Eijck, P. Verwer, and D. E. Williams. Crystal structure prediction of small organic molecules: a second blind test. *Acta Crystallographica Section B*, 58(4):647–661, 2002.
- [135] G. M. Day, T. G. Cooper, A. J. Cruz-Cabeza, K. E. Hejczyk, H. L. Ammon, S. X. M. Boerrigter, J. S. Tan, R. G. Della Valle, E. Venuti, J. Jose, S. R. Gadre, G. R. Desiraju, T. S. Thakur, B. P. van Eijck, J. C. Facelli, V. E. Bazterra, M. B. Ferraro, D. W. M. Hofmann, M. A. Neumann, F. J. J. Leusen, J. Kendrick, S. L. Price, A. J. Misquitta, P. G. Karamertzanis, G. W. A. Welch, H. A. Scheraga, Y. A. Arnautova, M. U. Schmidt, J. van de Streek, A. K. Wolfq, and B. Schweizerr. Significant progress in predicting the crystal structures of small organic molecules – a report on the fourth blind test. *Acta Crystallographica Section B*, 65(107-125), 2009.
- [136] M. A. Neumann and M. A. Perrin. Energy ranking of molecular crystals using density functional theory calculations and an empirical van der waals correction. *Journal of Physical Chemistry B*, 109(32):15531–15541, 2005.
- [137] M. A. Neumann, F. J. J. Leusen, and J. Kendrick. A major advance in crystal structure prediction. *Angewandte Chemie International Edition*, 47(13):2472–2430, 2008.
- [138] G. M. Day, J. Chisholm, N. Shan, W. D. S. Motherwell, and W. Jones. An assessment of lattice energy minimization for the prediction of molecular organic crystal structures. *Crystal Growth & Design*, 4(6):1327–1340, 2004.
- [139] W. Doering. *Abstracts of the American Chemical Society Meeting*, page 24, 1951.
- [140] Accelrys Inc. MS modelling, 2004.
- [141] J. P. Perdew and Y. Wang. Accurate and simple analytic representation of the electron-gas correlation energy. *Physical Review B*, 45(23):13244–13249, 1992.
- [142] B. Delley. An all electron numerical method for solving the local density functional for polyatomic molecules. *The Journal of Chemical Physics*, 92(1):508–517, 1990.

- [143] C. Glidewell, J. N. Low, and J. L. Wardell. Conformational preferences and supramolecular aggregation in 2-nitrophenylthiolates: disulfides and thiosulfonates. *Acta Crystallographica*, 56(6):893–905, 2000.
- [144] A. M. Belenguer, T. Friščić, G. M. Day, and J. K. M. Sanders. Solid-state dynamic combinatorial chemistry: reversibility and thermodynamic product selection in covalent mechanosynthesis. *Chemical Science*, 2(4):696–700, 2011.
- [145] P. J. Bygrave, D. H. Case, and G. M. Day. Is the equilibrium composition of mechanochemical reactions predictable using computational chemistry? *Faraday Discussions*, 170(1):41–57, 2014.
- [146] P. J. Winn, G. G. Ferenczy, and C. A. Reynolds. Toward improved force fields. 1. multipole derived atomic charges. *The Journal of Physical Chemistry A*, 101(30):5437–5445, 1997.
- [147] G. G. Ferenczy, P. J. Winn, and C. A. Reynolds. Toward improved force fields. 2. effective distributed multipoles. *The Journal of Physical Chemistry A*, 101(30):5446–5455, 1997.
- [148] C. M. Breneman and K. B. Wiberg. Determining atom-centered monopoles from molecular electrostatic potentials. the need for high sampling density in formamide conformational analysis. *Journal of Computational Chemistry*, 11(3):361–373, 1990.
- [149] R. Stafford. Random points in an n-dimensional hypersphere - file exchange. *MATLAB Central*, 2005.
- [150] G. Marsaglia. Choosing a point from the surface of a sphere. *Journal of Mathematics and Statistics*, 43(2):645–646, 1972.
- [151] L. Pauling. *The Nature of the Chemical Bond*. Cornell University Press, 3rd edition, 1960.
- [152] E. H. Moore. On the reciprocal of the general algebric matrix. *Bulletin of the American Mathematical Society*, 26(9):394–395, 1920.
- [153] A. Bjerhammer. Application of calculus of matrices to method of least squares; with special references to geodetic calculations. *Royal Institute of Technology*, 49, 1951.
- [154] R. Penrose. A generalized inverse for matrices. *Proceedings of the Cambridge Philosophical Society*, 51(1):406–413, 1955.

- [155] S. M. Stigler. Gauss and the invention of least squares. *The Annals of Statistics*, 9(3):465–474, 1981.
- [156] J. S. Chickos. Enthalpies of sublimation after a century of measurement: A view as seen through the eyes of a collector. *Netsu Sokutei*, 30(3):116–124, 2003.
- [157] G. A. Kaminski, R. A. Friesner, J. Tirado-Rives, and W. L. Jorgensen. Evaluation and reparametrization of the OPLS-AA force field for proteins via comparison with accurate quantum chemical calculations on peptides. *The Journal of Physical Chemistry B*, 105(28):6474–6487, 2001.
- [158] A. Perry. Technical note - a modified conjugate gradient algorithm. *Operations Research*, 26(6):1073–1078, 1978.
- [159] B. Mennucci. Polarizable continuum model. *Computational Molecular Science*, 2(3):386–404, 2012.

An idle conversation between a chemical physicist and a chemist in the office on a rainy Thursday morning,

“That’s all chemistry is, finding low energy pathways.”,

“Tell that to an organic chemist...”