**University of Southampton**

Faculty of Engineering and Physical Sciences
School of Chemistry

# Enhancing Crystal Structure Prediction Methods for Flexible Small Molecule Pharmaceuticals

Volume 1 of 1

*by*

**James Owen David Bramley**
MChem
ORCiD: 0000-0003-0215-0355

*A thesis for the degree of*
*Doctor of Philosophy*

November 2025

University of Southampton

Abstract

Faculty of Engineering and Physical Sciences
School of Chemistry

Doctor of Philosophy

**Enhancing Crystal Structure Prediction Methods for Flexible Small Molecule Pharmaceuticals**

by James Owen David Bramley

This thesis presents methods for crystal structure prediction, introducing techniques that improve the identification of conformations used as seeds in the search for crystal structures. The proposed approach enhances sampling of the conformational hypersurface, leading to superior initial structures that facilitate a more thorough exploration of conformational space. Additionally, a Monte Carlo simulated annealing method has been developed, integrating experimental and computational techniques to effectively determine crystal structures. This method has demonstrated success for rigid molecules and polymorphs under various conditions. A Monte Carlo refinement procedure has also been utilised to enable precise matching of crystal structure prediction datasets to experimental data for both flexible and rigid molecules. These methodologies hold promising potential for diverse applications in crystal structure prediction but require further research to ensure their robustness for practical workloads.

# Contents

# List of Figures

# List of Tables

# Declaration of Authorship

I declare that this thesis and the work presented in it is my own and has been generated by me as the result of my own original research.

I confirm that:

1. This work was done wholly or mainly while in candidature for a research degree at this University;
2. Where any part of this thesis has previously been submitted for a degree or any other qualification at this University or any other institution, this has been clearly stated;
3. Where I have consulted the published work of others, this is always clearly attributed;
4. 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;
5. I have acknowledged all main sources of help;
6. Where the thesis is based on work done by myself jointly with others, I have made clear exactly what was done by others and what I have contributed myself;
7. None of this work has been published before submission

Signed:..... Date:.....

# Acknowledgements

I would like to thank Prof. Graeme Day for providing excellent support over the last 4 years and always being a friendly and approachable supervisor which is greatly appreciated. I wish to thank the whole of the Day Group, both former and current members, for their ideas, suggestions and assistance with many small problems. I would like to express my gratitude to Johnson Matthey for funding my studies and the supervisory support of Dr. Christopher Perry.

We thank the Engineering and Physical Sciences Research Council for funding the UK National Crystallography Service (grant No. EP/W02098X/1) which provided the experimental facilities for these studies.

We are grateful to the UK Materials and Molecular Modelling Hub for computational resources, which is partially funded by EPSRC (EP/T022213/1, EP/W032260/1 and EP/P020194/1). I am also grateful for iridis 5 and their HPC team, for providing excellent support with the use of their systems. I would like to acknowledge that this work used the ARCHER2 UK National Supercomputing Service (https://www.archer2.ac.uk).

I would like to thank my parents and family for their financial and loving support over the years, who have always been there for me. Grateful thanks go to Amy Witzmann and Catherine Crockett for their friendship and support through my studies in Chemistry. I would like to give a special thank you to Dr. Jack Hodgson for always being a fantastic friend and drinking companion over the last 8 years and without which I would not be in the position I am today.

*Dedicated to my Mum and Dad and their forever loving support.*

# Definitions and Abbreviations

| | |
|---|---|
| **API** | Active Pharmaceutical Ingredient |
| **BFDH** | Bravais-Friedel-Donnay-Harker |
| **BFGS** | Broyden–Fletcher–Goldfarb–Shanno |
| **BH** | Basin Hopping |
| **CREST** | Conformer Rotamer Ensemble Sampling Tool |
| **CCDC** | Cambridge Crystallographic Data Centre |
| **cDTW** | constrained Dynamic Time Warping |
| **CIF** | Crystallographic Information File |
| **CLG** | Crystal Landscape Generator |
| **CPU** | Central Processing Unit |
| **CSD** | Cambridge Structural Database |
| **CSP** | Crystal Structure Prediction |
| **CV** | Collective Variable |
| **DCM** | dichloromethane |
| **DFT** | Density Functional Theory |
| **DFTB** | Density Functional Tight Binding |
| **DMA** | Distributed Multipole Analysis |
| **DSC** | Differential Scanning Calorimetry |
| **DTW** | Dynamic Time Warping |
| **GA** | Genetic Algorithm |
| **GTO** | Gaussian Type Orbital |

**HF** Hartree Fock

**HPC** High Performance Computing

**LMCS** Low Mode Conformer Search

**MBD** Many Body Dispersion

**MC** Monte Carlo

**MCSA** Monte Carlo Simulated Annealing

**MD** Metadynamics

**MDS** Multi Dimensional Scaling

**ML** Machine Learning

**MMFF** Merck Molecular Force Field

**NMR** Nuclear Magnetic Resonance

**PC** Principal Component

**PCA** Principal Component Analysis

**PCM** Polarisable Continuum Model

**PES** Potential Energy Surface

**PXRD** Powder X-ray Diffraction Pattern

**QM** Quantum Mechanical

**QR** Quasi Random

**RMS** Root Mean Square

**RMSD** Root Mean Square Deviation

**RNA** Ribonucleic Acid

**SASA** Solvent Accessible Surface Area

**SCC** Self Consistent Charge

**SCF** Self Consistent Field

**SCXRD** Single Crystal X-Ray Diffraction

**ss-NMR** solid state Nuclear Magnetic Resonance

**STO** Slater Type Orbital

**THF** tetrahydrofuran

**VASP** Vienna Ab initio Simulation

**xTB** Extended Tight Binding

# Chapter 1

# Introduction

This thesis explores the prediction of organic molecular crystal lattices with a specific focus on flexible small-molecule pharmaceuticals. The primary goal is to enhance the accuracy of crystal structure predictions and introduce novel approaches for predicting the structures of flexible molecules. In refining current CSP methods, we developed new techniques where existing ones fell short. This work presents broad method development with significant potential for future advancements. Each chapter delves into different aspects of the methodology, from conformer prediction to other CSP strategies.

## 1.1 Chapter Overview

Chapter 2 provides the necessary background on CSP. We discuss the history of organic molecular CSP, evaluating its accuracy and applications to various molecular structures. We also review the six blind tests conducted over the past 25 years by scientists worldwide, which highlights the progress of CSP and its future possibilities. We explain the concepts of conformers and conformations and their importance in CSP. Additionally, we cover foundational concepts such as how crystals are described, the role of solvates, and the phenomenon of polymorphism, all of which are essential for understanding the rest of the thesis.

In Chapter 3, we outline the relevant theories, methods, and programs used within this thesis. We provide an overview of Density Functional Theory (DFT) and its applications in electronic structure theory, as well as the statistical methods used throughout this work. We also explain how CSP calculations are performed from the ground up, giving a solid theoretical grounding.

In Chapter 4, we present and demonstrate the CSP workflow applied to various crystal systems associated with the $7^{th}$ CSP Blind Test. We outline our methodology in detail, focusing on the strategies used to tackle the prediction of highly flexible molecules, which required advanced conformational search techniques. The molecules investigated in this test posed significant challenges due to their flexibility, and we highlight how our methods addressed these challenges effectively.

Chapter 5 focuses on methods for identifying molecular conformations to seed CSP methods. We explore two different conformer clustering methods, using a series of flexible, drug-like molecules. We demonstrate how our methodology improves upon previous approaches for generating conformers.

In Chapter 6, we introduce a novel Monte Carlo Simulated Annealing (MCSA) approach for predicting crystal structures using experimental data, such as Powder X-ray Diffraction Patterns (PXRDs) and Nuclear Magnetic Resonance (NMR). We demonstrate the effectiveness and versatility of this method compared to other approaches, using a series of flexible and non-flexible molecular systems. We also explore how this method can be parametrised and combined with different experimental data to yield more accurate results than typical CSP methods. Additionally, we discuss the use of a Basin Hopping (BH) approach and how it is integrated into our CSP software, cspy [1].

In Chapter 7, we present a comprehensive CSP workflow for the flexible molecule idelalisib and its solvates. We explore the challenges of dealing with flexible molecules, such as their complex conformational space and multiple degrees of freedom. We apply the methodologies described in Chapter 6 to address these challenges and optimise the prediction of crystal structures. Additionally, we conduct a rigorous post-CSP analysis using a range of Quantum Mechanical (QM), tight binding and forcefield methods.

These methods allow for a thorough refinement and evaluation of the predicted structures, ensuring the reliability of the results.

In Chapter 8, we explore an alternative sampling method developed in collaboration with colleagues from the Day group, aimed at identifying conformations that are likely to be present in the crystal. This method aims to improve the generation of conformers by enabling a more thorough sampling of the potential energy landscape. We demonstrate the entire workflow, using two distinct approaches to address the challenges posed by highly flexible molecules with many degrees of freedom. By applying this advanced sampling technique, we show how it enhances the exploration of the conformational space, leading to more accurate predictions of crystal structures in complex molecular systems.

# Chapter 2

# Background

Organic crystal structures are defined by the periodic packing of one or more organic molecules in 3 dimensions. The specific packing of these organic molecules is determined by intermolecular and intramolecular interactions, which in effect dictates physical solid state properties such as melting point, density, optical activity and solubility.

The number of packings possible for even a single small rigid organic molecule can be many and slight changes to the crystallisation process can result in different arrangement of molecules within a crystal structure. If the molecular unit possesses observable metastable crystal forms; the structures are known as polymorphs. Due to their differing arrangement of molecules, different polymorphs might possess many different physiochemical properties. With sufficient understanding of forces, orientation and conformations of molecules, it is possible to predict the crystal structure of a molecule or set of molecules computationally.

## 2.1 Descriptions of Crystal Structures

Crystal structures are not only relevant to organic molecular systems but also have relevance to inorganic chemistry and other disciplines. However, we will only describe crystal systems for organic molecules to remain relevant to the subject of this thesis.

### 2.1.1 Bravais Lattices

Organic crystal structures can be described by the smallest possible periodic molecular unit repeating in all 3 Cartesian dimensions in which we call the unit cell. This molecular unit could contain a single molecule, fractions of molecules or whole molecules.

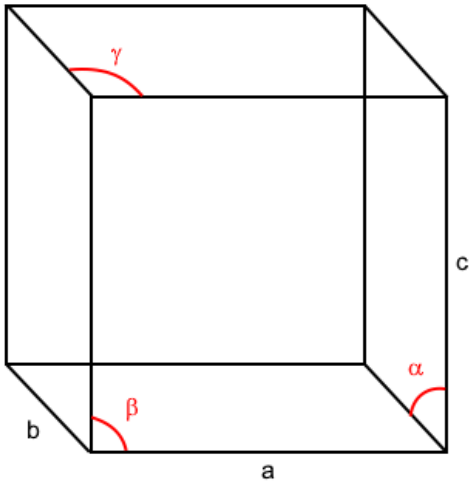



FIGURE 2.1: An empty unit cell with sides of length a, b, c and angles $\alpha$, $\beta$ and $\gamma$. For different unit cells, each of these parameters could vary. The combination of unit cell parameters dictates the crystal system.

The lengths a, b and c correspond to the lengths on each side of the unit cell and angles $\alpha$, $\beta$ and $\gamma$ correspond to the angle between sides b and c, a and c, and a and b respectively. Unit cells are classified by their shape depending on the angles and unit cell lengths present in the 3-dimensional body. The system of a crystal can be described as either triclinic, monoclinic, orthorhombic, tetragonal, hexagonal and cubic as summarised in Table 2.1.

| Crystal System | Symbol | Unit Cell Lengths | Unit Cell Angles |
|---|---|---|---|
| Triclinic | P | $a \neq b \neq c$ | $\alpha \neq \beta \neq \gamma \neq 90°$ |
| Monoclinic | C | $a \neq b \neq c$ | $\alpha = \gamma = 90°$, $\beta \neq 90°$ |
| Orthorhombic | O | $a \neq b \neq c$ | $\alpha = \beta = \gamma = 90°$ |
| Tetragonal | T | $a = b \neq c$ | $\alpha = \beta = \gamma = 90°$ |
| Trigonal | R | $a = b = c$ | $\alpha = \beta = \gamma \neq 90°$ |
| Hexagonal | H | $a = b \neq c$ | $\alpha = \beta = 90°$, $\gamma = 120°$ |
| Cubic | F | $a = b = c$ | $\alpha = \beta = \gamma = 90°$ |

TABLE 2.1: Classification of unit cells based on unit cell parameters a, b, c, $\alpha$, $\beta$, $\gamma$. These parameters indicate which crystal system a structure belongs to.

The positions in space that define the periodic arrangement of a crystal are known as lattice points. A unit cell can contain one or multiple lattice points, depending on the symmetry and centring of the lattice. Each lattice point can be occupied by one molecule, multiple molecules, or parts of molecules. The way lattice points are arranged within the unit cell determines the cell's centring type, as illustrated in Figure 2.2.

| Symbol | Lattice Type | Description |
|---|---|---|
| P | Primitive | Lattice point located only at unit cell corners |
| I | Body-centred | Primitive lattice type with an additional lattice point located at the centre of the unit cell |
| F | Face-centred | Primitive lattice type with an additional lattice point located at the centre of each face of the unit cell |
| S | Base-centred | Primitive lattice type with additional lattice points located in the centre of one pair of opposite faces |
| R | Rhombohedral lattice | Primitive lattice which is typically used to describe hexagonal crystal systems |

TABLE 2.2: Centring Types of crystal systems. These describe where lattice points are located within the unit cell which determine the lattice type.

Whilst any lattice type can possess any centring, many combinations are redundant as certain arrangements are equivalent. Therefore, only a subset of these combinations is necessary to describe all distinct lattice types. In 3-dimensions, there are 14 possible Bravais lattice types that describe the shape and symmetry of the unit cells which form the foundation of a the crystal's overall symmetry.

### 2.1.2 Space Groups

In addition to the lattice types, there are 32 crystallographic point groups, which describe the sets of symmetry operations that leave at least one point fixed, typically the origin. These point groups encompass all symmetry operations such as rotations, reflections, and inversions, and are used to classify the symmetry of molecules. A molecule belongs to a particular point group if it remains unchanged under the operations in that group.

While point groups describe local symmetries, space groups extend this concept by combining point group symmetries with translational symmetry, thereby describing the full symmetry of a crystal lattice. Altogether, there exist 230 distinct space groups in three dimensions, representing all possible combinations of symmetry operations in crystals. Space groups are characterised by various symmetry operations, including rotation, reflection, screw axes, and glide planes which are described in Table 2.3.

Additionally, crystals can be described as centrosymmetric or non-centrosymmetric, depending on whether they possess a centre of symmetry. In a centrosymmetric crystal, every part of the structure has an equivalent counterpart on the opposite side of the centre, related by inversion. In contrast, non-centrosymmetric crystals lack this inversion symmetry.

### 2.1.3 Space Group Notation

To describe space groups, Hermann-Mauguin notation is used which consists of lattice type, and additional characters corresponding to the symmetries present [2].

The first symbol in this notation represents the centring type of the cell shown in Table 2.2. The following characters are the symmetry operations. Some space groups may have different operations present along each axis, such that the first position corresponds to symmetry elements related to the a-axis, the second position to the b-axis, and the third position to the c-axis. Whether the operation acts parallel or perpendicular to the axis depends on the type of symmetry element.

| Symmetry Element | Symbol | Description |
|---|---|---|
| Rotation | r | Rotation by $\frac{360^\circ}{r}$ |
| Mirror | m | Reflection symmetry across a plane perpendicular to the axis |
| Roto-mirror | r/m | Rotation by $\frac{360^\circ}{r}$ followed by reflection across the plane perpendicular to the rotation axis |
| Roto-inversion | $\bar{r}$ | Rotations by $\frac{360^\circ}{r}$ followed by an inversion |
| Glide Plane | $a, b, c, n, d$ | Reflection symmetry across a plane perpendicular to the axis. This is followed by a translation depending on the type of glide plane. For $a$, $b$ and $c$ the glide plane this is $\frac{1}{2}$ the unit cell length along axis $a$, $b$ and $c$ respectively. For $n$, $\frac{1}{2}$ translation along two in-plane axes and $d$, $\frac{1}{4}$ or $\frac{3}{4}$ translation along two in-plane axes |
| Screw Axis | $r_m$ | Rotation by $\frac{360^\circ}{r}$ followed by fractional translation equal to $\frac{m}{r}$ along the axis |

TABLE 2.3: Symmetry operations used to define space groups applied to x, y, or z axes. r = 1, 2, 3, 4, 6.

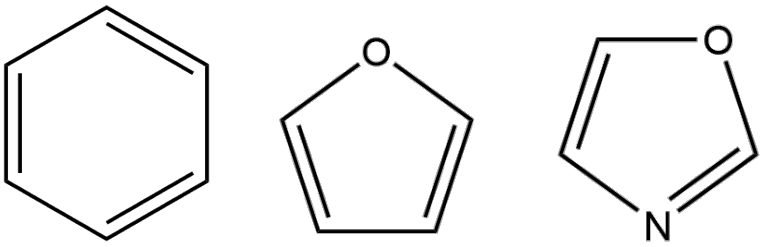
For all space groups, inversion centres and rotation axes are restricted to specific positions defined by the symmetry of the crystal lattice. These positions are chosen such that applying the symmetry operations maps the crystal onto itself without disrupting its periodicity. Typically, symmetry elements are located at the origin, cell centres, face centres, edge centres, or other fractional coordinates that correspond to lattice translations or special symmetry sites.

### 2.1.4 Z and Z'

The number of molecules in the unit cell is denoted by Z, while Z' represents the number of distinct molecules in the asymmetric unit.

## 2.2 Conformers and Conformations

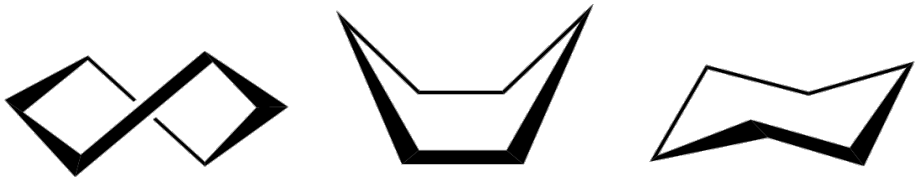
Within a molecular crystal, molecules can adopt different geometries known as conformers. Some molecules only possess a single stable molecular geometry which can be described as "rigid" and often possess no freely rotatable bonds such as those shown in Figure 2.2.




(A) Benzene (B) Furan (C) Oxazole

FIGURE 2.2: Example diagrams of rigid molecules. Each molecule contains freely rotatable bonds meaning that the molecular conformation of the molecule is fixed.

Molecules, such as those in non-conjugated ring systems such as boat and chair conformations in 6-membered carbon rings.



(A) Twist (B) Boat (C) Chair

FIGURE 2.3: Conformational isomers of cyclohexane. Cyclohexane posses no freely rotatable bonds however still has multiple conformers by moving its flexible ring.

However, molecules which do possess freely rotatable bonds may be able to adopt different conformations where the stability of each conformation is based upon a molecule's intramolecular interactions.

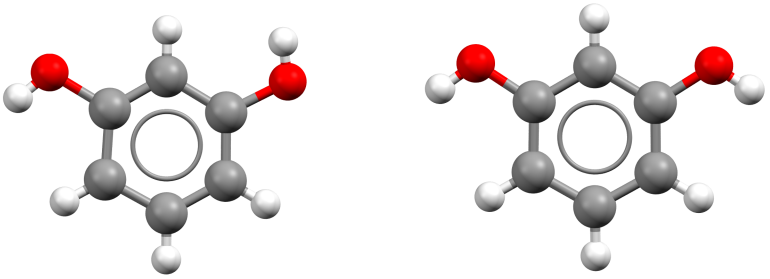


FIGURE 2.4: Ball and stick models for two different conformers of resorcinol. Resorcinol featuring a rigid ring system but contains two OH groups, allowing it to adopt different conformations.

If we consider resorcinol which shown in Figure 2.4, as an isolated molecule, as we rotate around a bond which possesses torsion, we find that certain geometries are metastable. Any change to the torsional angle, which is the angle between two planes formed by four sequentially bonded atoms, would reside in an increase in conformational energy. These configurational minima are known as conformers.

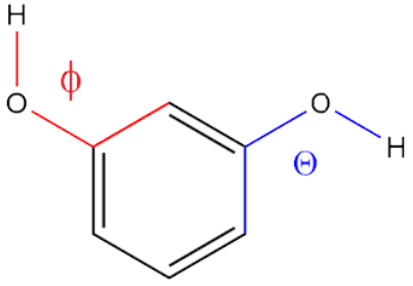



FIGURE 2.5: Resorcinol with rotatable bonds shown. The atoms that make up the torsion of $\phi$ and $\Theta$ are shown where each OH group can rotate 360°.

If we fix the torsion angle $\phi$ at 0°, whilst rotating the torsion $\Theta$, we find that two metastable configurations exist at 0° and 180° around this torsion. Although other conformers may exist, we would need to sample around torsion $\phi$ and $\Theta$ simultaneously to locate them. By doing this we are able to produce what is known as the Potential Energy Surface (PES) of the molecule. This surface describes the energy of all possible conformations of the molecule.

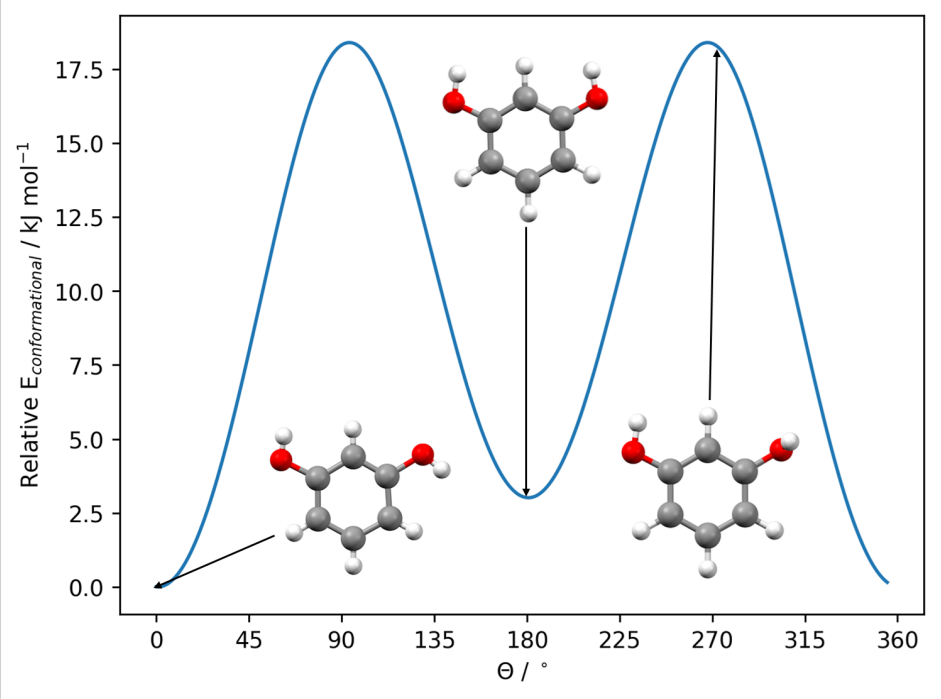



FIGURE 2.6: Effect of torsional rotation on the energy of the resorcinol molecule. The torsion angle Θ determines the energetic stability of each conformation. Two conformational minima are present at 0° and 180°, corresponding to where the OH bond is parallel to the plane of the aromatic ring. At angles of 90° and 270°, the OH bond is positioned orthogonal to the plane, resulting in a significant energy penalty.

Benzene, resorcinol and cyclohexane depicted in Figures 2.2A, 2.3 and 2.4 appear similar in structure and shape. However, conformational analysis of these structures reveal significant differences. In benzene, the stacking of the aromatic $\pi$ ring allows for significant low energy arrangements when packing into crystal structures. Whilst resorcinol also benefits from $\pi$ stacking, it also contains directional hydrogen bonding groups in which the orientation can effect the arrangements of molecules when closely packed. Cyclohexane does not contain any $\pi$ electrons and therefore relies on van der Waals forces between molecules for stability.

## 2.3 Polymorphism

Polymorphism is the phenomenon where a single chemical compound can crystallise into more than one distinct structure named a polymorph. These polymorphs can have vastly different physical properties, such as solubility, melting points and stability. Although the molecule's composition remains unchanged, the way its lattice is arranged differs, resulting in distinct forms of the same substance. In the pharmaceutical industry, for example, a drug may have several polymorphs, each with unique properties. One polymorph might dissolve quickly in the body, making it an ideal drug candidate, while another polymorph may be less soluble, rendering it ineffective [3]. Therefore polymorphism is of critical importance in pharmaceuticals, agrochemicals and materials science, as different polymorphs can lead to variations in the efficacy, safety and stability of a product [4].

Different polymorphs can form due to variations in crystallisation conditions such as temperature, pressure, and the choice of solvent. Even slight environmental changes can result in different polymorphs, presenting both opportunities and challenges for industries that require consistent material properties [5]. The formation of polymorphs is governed by thermodynamics and kinetics. From a thermodynamic perspective, the most stable polymorph has the lowest free energy, making it the preferred form under equilibrium conditions. However, if crystallisation conditions favour rapid nucleation and growth, metastable polymorphs may form, bypassing the thermodynamically stable form [6]. These metastable polymorphs can eventually convert to the stable form over time but they may also persist for extended periods influenced by factors such as temperature.

Polymorph screening, which is the process of identifying and characterising all possible polymorphs of a compound, is a crucial part of drug development and materials science [7]. Experimental methods such as Single Crystal X-Ray Diffraction (SCXRD) and Differential Scanning Calorimetry (DSC) are typically used to identify and characterise these forms, but computational approaches are increasingly being employed to predict them as discussed later on [8].

## 2.4 Crystal Structure Prediction (CSP)

CSP refers to the method of predicting the structures of solid-state materials. This thesis focuses on non-empirical methods for predicting flexible organic molecular crystals using only chemical diagrams of molecules as input.

CSP attempts to determine the configuration of molecules packed into a crystal that is lowest in energy which is determined by sampling the entirety of configurational space. If we consider both intermolecular and intramolecular forces, we wish to find the configuration that minimises the total energy, $E_{\text{total}}$:

$$E_{\text{total}} = E_{\text{intermolecular}} + E_{\text{intramolecular}}. \tag{2.1}$$

If we consider all of configurational space we can determine what is known as the crystal energy landscape.

Various approaches have been developed to address CSP, including MC simulations, evolutionary algorithms, and molecular dynamics simulations, among others [9–11] which enable an exploration of the crystal landscape.

### 2.4.1 Energy Evaluation

In CSP, accurately ranking the energy of different crystal structures is essential for identifying the most thermodynamically stable crystal form. Several computational methods, each varying in accuracy and computational cost, are used for this purpose. These include force-field based approaches, DFT, and various energy correction techniques.

Force-field methods rely on classical potentials to approximate the total energy of a system usually by summing over pairwise interactions. Due to their simplicity, force-fields are computationally inexpensive and allow for large-scale searches of configuration space. Force-fields, such as Lennard-Jones potentials for non-covalent interactions and Buckingham potentials for ionic systems, can struggle with transferability across diverse materials, which lead to inaccurate energy rankings in complex systems [12].

DFT generally offers greater accuracy than force-field methods, as it explicitly accounts for the electronic structure of the crystal. It is particularly effective for systems with complex bonding environments and often refines the energy rankings of structures initially identified through force-field calculations. However, DFT is considerably more computationally demanding. A significant challenge in DFT is the accurate treatment of dispersion forces, which stem from fluctuations in electron density that induce temporary dipoles in neighbouring molecules. These forces are crucial in molecular crystals,

frequently representing the dominant interactions in non-polar systems and still contributing substantially in polar systems. This challenge has spurred the development of dispersion-corrected DFT approaches, such as DFT-D3 and van der Waals functionals, to enhance accuracy for these materials [13].

Hybrid approaches combine the strengths of force-fields and DFT. In many CSP studies, force-fields are used for the initial screening of candidate structures, followed by DFT refinement to ensure accuracy without overwhelming computational resources.

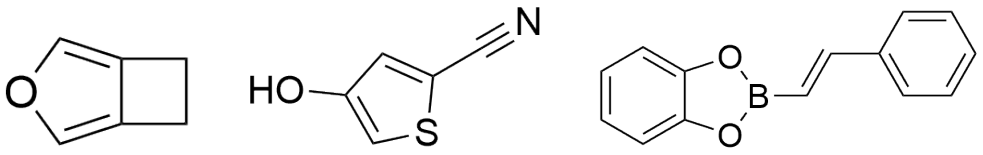
Post-DFT corrections are often applied to improve the accuracy of energy predictions. These include Many Body Dispersion (MBD) corrections for long-range interactions and vibrational corrections, which account for temperature-dependent effects. For systems where thermal expansion affects stability, the quasi-harmonic approximation is used to predict free energy at finite temperatures, further improving the reliability of the rankings [14].

Machine Learning (ML) approaches are emerging as a promising addition to CSP workflows. Trained on data from DFT or experiments, ML models can predict energy rankings with significantly reduced computational cost. However, the accuracy of these models is highly dependent on the quality of the training data and the choice of features [15, 16].

### 2.4.2 Overview of Crystal Structure Prediction Methods

CSP has evolved significantly, moving from early heuristic searches to sophisticated computational workflows that exploit increases in computer power and algorithmic efficiency. This progress has been monitored through periodic blind tests organised by the Cambridge Crystallographic Data Centre (CCDC), in which researchers predict the crystal structures expected from a given chemical diagram and occasionally supplement them with experimental data such as PXRD patterns. Here, the steady improvement from the first blind test in 1999 to the seventh in 2021 is summarised.

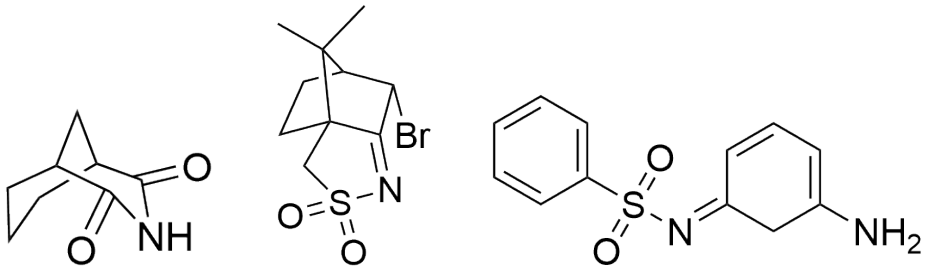
In the first blind test (1999) early CSP methods were still constrained by the cost of fully exploring the crystal-energy landscape, so most groups used empirical or semi-empirical force fields. Electrostatics were already treated with some sophisticated atomic point charges or distributed multipoles derived from *ab initio* Hartree Fock (HF) or second-order Møller–Plesset perturbation theory charge densities, but dispersion interactions and the delicate balance between many low-energy minima remained hard to capture. Several participants employed purely statistical fitness functions built from probability distributions in the Cambridge Structural Database (CSD), showing that a ranking function need not be a direct estimate of lattice energy. Rigid molecule assumptions and MC sampling had limited success for anything except the simplest, most rigid molecules [17]. Some targets are shown in Figure 2.7.




(A) I (B) II (C) III

FIGURE 2.7: Targets I-III used in the first blind test

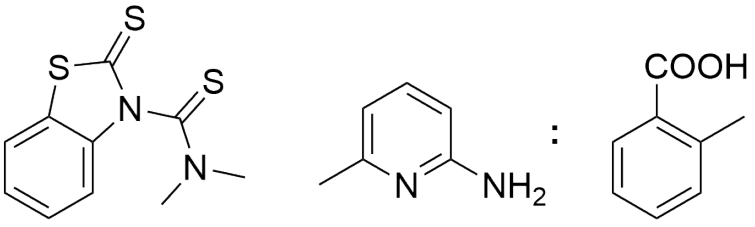
The second (2001) and third (2004) blind tests broadened the range of techniques. Most entries still relied on force-field lattice energies, but new search algorithms such as Genetic Algorithm (GA) and more systematic grid searches improved sampling efficiency [7, 18]. The third test also saw more elaborate potentials: anisotropic repulsion for halogen atoms, distributed multipoles, Angelo Gavezzotti's PIXEL electron-density integration scheme, and Detlef Hofmann's CSD-trained statistical potential [19, 20].Some targets are shown in Figure 2.8.




(A) IV (B) V (C) VI

FIGURE 2.8: Targets IV-VI for the second blind test

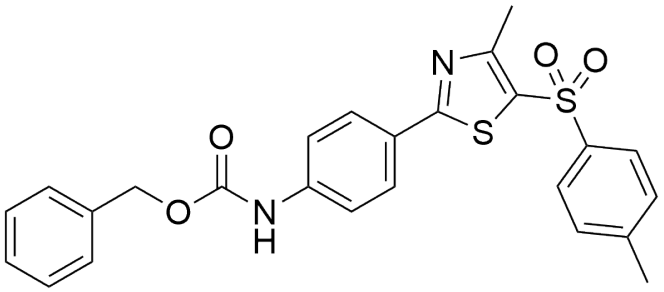
The fourth blind test (2007) marked the first use of periodic dispersion-corrected DFT. Neumann, Leusen and Kendrick applied periodic Perdew–Burke–Ernzerhof calculations supplemented by an atom–atom term and successfully ranked all four experimental structures as global minima [21]. Retrospective application to earlier blind-test molecules confirmed the power of this approach although the computational cost restricted its use to a limited number of candidate structures. Some targets are shown in Figure 2.9.




(A) XIV (B) XV

FIGURE 2.9: Targets XIV-XV for the fourth blind test

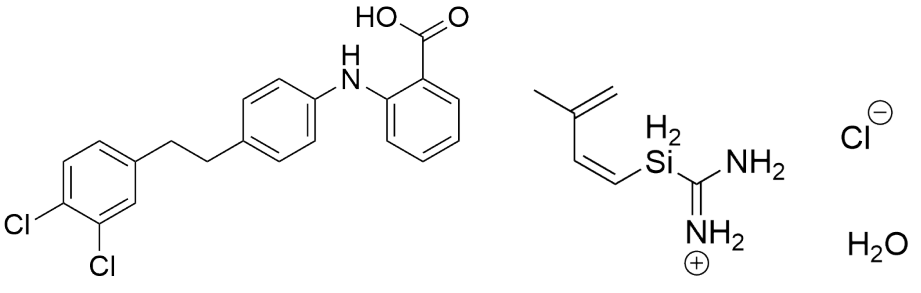
In the fifth blind test (2010) periodic DFT-D became much more widely adopted. For the first time the targets included a large, flexible, drug-like molecule, catalysing pharmaceutical interest in CSP. Molecules with rigid geometries were now routinely predicted, whereas flexible molecules possessing multiple low-energy conformations still posed a challenge [22]. Some targets are shown in Figure 2.10.



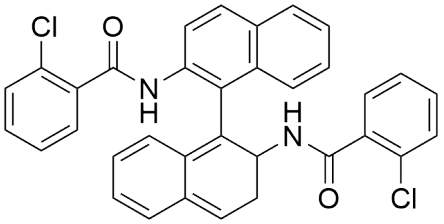

(A) XX

FIGURE 2.10: Target XX for the fifth blind test

The sixth blind test (2016) introduced five demanding highly flexible molecules, extensive hydrogen-bonding networks and a multi-component crystal. Workflows combined MC parallel tempering for structure generation with periodic DFT-D ranking, fragment-based or symmetry-adapted perturbation theory fitted potentials, vibrational free-energy corrections and even kinetic MC simulations of nucleation. One such MC + DFT-D pipeline predicted every experimental structure correctly [23]. Some targets are shown in Figure 2.11.




(A) XXIII (B) XXIV




(C) XXVI

FIGURE 2.11: Targets XXIII, XXIV and XXVI for the sixth blind test

There has been a clear trend that improved sampling algorithms, using more accurate energy models and greater computing resources have progressively increased the reliability of CSP, even for large flexible molecules and multi-component systems. Between 2020 and 2022, the seventh blind test was conducted. Details of the methods employed are discussed in Chapter 4.

### 2.4.3 Quasi Random Crystal Structure Prediction (QR-CSP)

Historically, CSP efforts within the Day group have focused predominantly on small, rigid molecules, as these systems present a more manageable search space and lower computational requirements. However, advancements in computational power and efficiency have facilitated the study of larger, more flexible molecules, which are of growing significance in the pharmaceutical field due to the impact of polymorphism on drug development. These technological improvements now allow for the prediction of crystal structures in considerably more complex systems.

In this thesis, crystal landscapes are generated using a Quasi Random (QR) routine. Molecular conformations are sampled using conformer search methods described in section 3.3, and molecular multipoles are calculated from these conformations as detailed in sections 3.1.10.1, 3.1.11, and 3.1.12. Crystals are produced by packing asymmetric units into various space groups, as discussed in section 2.4.3.2. Molecular multipoles for the structures are calculated through Distributed Multipole Analysis (DMA) and subsequently minimised using force-field approaches to yield chemically viable structures using DMACRYS which is described in section 3.1.10. Duplicate structures are removed using PLATON described in section 3.4.2.

An overview of the workflow is presented in Figure 2.12.

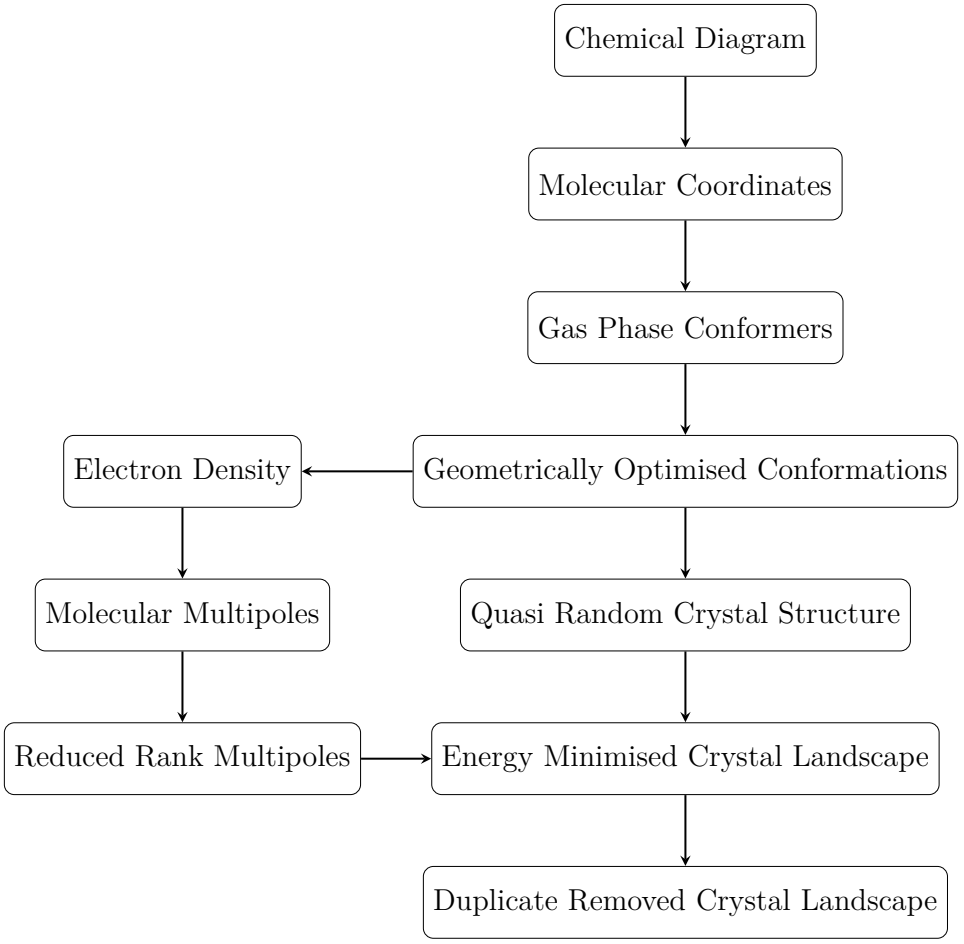



FIGURE 2.12: Workflow for Quasi Random Crystal Structure Prediction

In this section, we will delve deeper into each of these components.

#### 2.4.3.1 Conformer Searching

Our goal in a conformational search is to identify all conformers such as those in Figure 2.6 or at least those possible to be observed within a crystal structure for the purposes of CSP. We can achieve this by exploring the molecular PES.

The crystal landscape generator described in section 2.4.3.2 packs molecules into a crystal lattice. For rigid molecules, only a single conformation needs to be packed. However, in the case of flexible molecules, all relevant molecular configurations must be considered, as they influence the energetics and stability of the crystal. From Equation 2.1, it is known that both $E_{\text{intermolecular}}$ and $E_{\text{intramolecular}}$ must be minimised to produce a low energy crystal structure that may be experimentally observable. Consequently, it is necessary to identify conformations likely to form stable crystal structures. Since conformers represent energy minima, they can serve as starting points.

It is established that, for flexible molecules, the lowest energy conformer, also known as the conformational global minimum, does not always appear in the experimental crystal structure [24]. This is because crystals can achieve lower overall energy by sacrificing intramolecular energy and introducing strain, allowing molecules to maximise intermolecular interactions and thus lower the intermolecular energy of the crystal. Analyses of crystal structures have shown that the conformations adopted by molecules within crystals are often similar to those of conformers. Therefore, it is crucial to consider conformers when exploring the crystal energy landscape.

There are a variety of algorithms that have been developed to reduce computational cost and effort in a conformational search to locate conformational minima which are discussed in section 3.3.

Once conformers have been determined, they are then optimised at high level of theory such as DFT to ensure geometries are as accurate as possible.

#### 2.4.3.2 Crystal Landscape Generator (CLG)

To construct a crystal energy landscape, configurational space must be explored to identify crystals with the lowest energy. For small systems, this can be accomplished using grid-based methods, which sample degrees of freedom at fixed intervals to enable a comprehensive search of the landscape. However, as configurational space grows larger, sampling becomes increasingly challenging, as it is unclear how much sampling is necessary to adequately explore the landscape.

Random sampling offers an alternative approach, searching the landscape stochastically. This method is advantageous because extensive sampling can provide a good representation of the landscape. However, the inherent randomness means there is no certainty that the landscape has been thoroughly explored. Even with a large number of sampled points, certain regions of configurational space may remain unsampled as shown in Figure 2.13.

Instead of relying solely on stochastic methods, it is possible to balance the systematic nature of grid searches with the randomness of random searches. This approach, known as QR sampling, can more effectively sample a landscape by retaining information from previously sampled points. Such QR sampling enables more even spacing across the search space while preserving the beneficial qualities of randomness. Moreover, QR sampling offers the advantage of reproducibility.

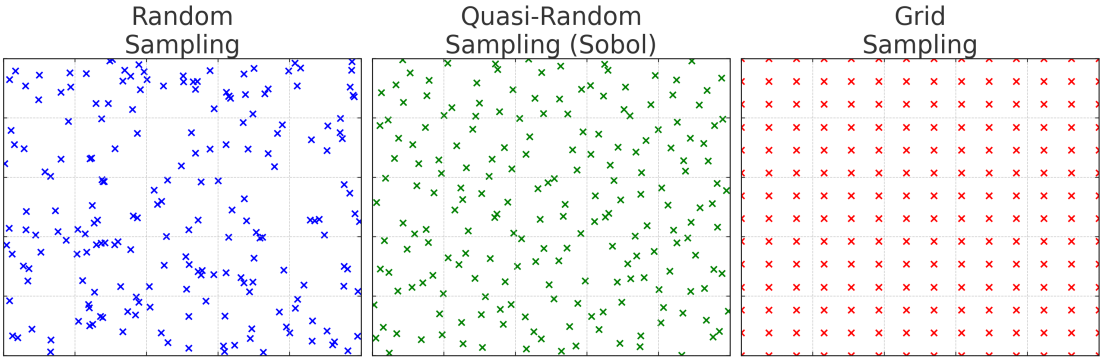



FIGURE 2.13: Comparison of different methods to sample a 2D plane using 200 data points. Random sampling may lead to uneven distribution of sampling points. Grid sampling requires a fixed interval of between data point. Quasi-Random sampling allows for an even sampling of space without relying setting fixed intervals between points.

### Structure Generation

The Crystal Landscape Generator (CLG) takes geometry optimised molecules and generates feasible crystal structures [25]. To generate the crystal structure of a molecule, numbers generated through low-discrepancy means using the Sobol method [26].

$$(x_1, x_2, x_3, \ldots, x_N), \tag{2.2}$$

where each $x_i \in [0, 1)$. These QR numbers are then mapped to structural parameters that define a trial crystal structure.

Molecular orientations are determined using the Shoemake method [27]. Here, uniform random rotations are generated directly as quaternions rather than relying on Euler angles, thereby avoiding biases and ensuring uniform sampling of orientations. The method transforms values of $x_i$ into angles and square roots that evenly distribute points over the surface of a four-dimensional hypersphere. Specifically, two numbers define angles around circles, analogous to longitude and latitude, while the third governs the relative weighting between two hemispherical components of the hypersphere. These values are combined using sine and cosine functions to produce four components $w, x, y, z$ that form a unit quaternion. This quaternion encodes both the axis of rotation and the angle of rotation, providing a smooth and uniform means of representing molecular orientations in three-dimensional space.

For unit cell angles $\theta$ ($\alpha$, $\beta$, or $\gamma$), the mapping avoids extreme angles that could produce very flat or highly skewed unit cells, which are problematic for energy minimisation. The angles are sampled using:

$$\theta = \arccos\Big(1 - 2x_i\Big), \tag{2.3}$$

to ensure a uniform distribution of $\cos(\theta)$, which naturally avoids extreme values.

For the unit cell lengths, a molecule's "shadow" is calculated along each cell axis to estimate physically reasonable bounds for the cell lengths. The length of lattice vector $l$ (a, b or c) is sampled as:

$$l_j = c \cdot \left[s_j^{\min} + x_i \cdot \left(N_{\text{mol}} \cdot s_j^{\max} - s_j^{\min}\right)\right], \tag{2.4}$$

where $s_j^{\min}$ and $s_j^{\max}$ are the minimum and maximum projections (shadows) of the molecule along lattice vector $j$, $N_{\text{mol}}$ is the number of molecules in the unit cell and $c$ is a scaling factor equal to 0.75. This ensures that cell lengths are sufficient to accommodate the molecules without being excessively large.

The positions of molecules are directly mapped to the fractional coordinates of the unit cell, ensuring even sampling of all possible positions.

When a crystal is generated, molecular clashes may occur. In such cases, the unit cell undergoes expansion, increasing the intermolecular distances between molecules which helps maintain thorough sampling of configurational space.

Clashes are resolved by enclosing each molecule within its convex hull which is a three-dimensional polyhedron encompassing all its atoms. If there exists an axis along which the projections of two convex hulls do not overlap, the molecules do not intersect. If an overlap is detected, the cell is expanded. In the case of expansion, the following equation is used:

$$\Delta l_j = \eta + \left| \frac{v_j^{\text{overlap}}}{v_j^{\text{centroid}}} \right|, \tag{2.5}$$

where $v^{\text{overlap}}$ represents the minimum translation vector required to separate overlapping convex hulls, and $v^{\text{centroid}}$ denotes the vector between the molecular centroids. The parameter $\eta$ is a small tolerance value equal to 0.001 Å.

If the clashes cannot be resolved such as if the expansion leads the unit cell volume becoming very large or the unit cell becomes very flat, the cell is discarded and treated as invalid.

Therefore, the CLG positions molecules at sensible distances from one another whilst avoiding molecules clashing. The generation provides starting points for minimisation

with a set of diverse molecular characteristics which can then undergo geometric optimisation.

To further enhance computational efficiency, the internal symmetry of a crystal's unit cell, as described in section 2.1.2, can be exploited. Analysis of crystals in the CSD has shown that most molecules crystallise in a limited set of common space groups [28]. Consequently, focusing on these prevalent space groups allows for a more rigorous search within a reduced portion of configurational space and therefore decreases the number of degrees of freedom that must be sampled, reducing computational cost.

#### Asymmetric Units

The smallest portion of the crystal structure that can generate the entire crystal using symmetry operations is called the asymmetric unit. In this way, the unit cell can be described by its internal symmetry.

For many systems, only one molecule is required to build a crystal and, in such cases, there is one molecule in the asymmetric unit. However, for some systems, such as solvates and co-crystals, more than one molecule is needed to describe the crystal structure. Therefore, when building the crystal, multiple molecules must be placed at each lattice point, giving rise to crystals with Z' > 1. These molecules may be identical but adopt different conformations, or they may be entirely different molecules.

If a molecule is centrosymmetric, it may allow for Z' < 1, meaning that less than a full molecule is needed to describe the asymmetric unit. For example, if a molecule has a plane of symmetry, only half of it is required to reproduce the rest of the molecule, and by extension, the unit cell. Additionally, centrosymmetric systems can occur where Z' = 1, and the asymmetric unit contains multiple molecular fragments or fractions of molecules.

### 2.4.4 Predicting Polymorphs

The prediction of polymorphs using computational techniques can save significant time and resources. However, despite substantial advancements, predicting polymorphs accuracy remains challenging.

Polymorphs can be identified by determining the configurations corresponding to local minima on the crystal energy landscape. This landscape is theoretical, and many local minima may not be experimentally observable because CSP methods often omit temperature from their calculations. If the energy barrier between crystal structures is small, crystals may have sufficient thermal energy to transition into a different energy basin.

In contrast, crystals separated by large energy barriers are more stable and are therefore less prone to converting into alternative configurations.

The stability of a polymorph is influenced by the energy barriers between crystal structures, which help predict whether it will be experimentally observed. These barriers can give rise to metastable forms that may only appear under certain conditions. While computational approaches are improving in their ability to simulate real-world conditions, this remains an area where experimental methods may be necessary to validate predictions [23].

Molecules with flexible structures, such as large organic compounds, present a further challenge [29]. Each conformer may lead to a different crystal packing, significantly increasing the number of potential polymorphs. Historical CSP methods are well-suited to small, rigid molecules but struggle to accurately model polymorphs of these molecules due to the significant number of degrees of freedom need to be explored complicating the search for stable polymorphs [28]. Accurate prediction of polymorphs requires the use of high-level computational techniques, such as DFT simulations, which are computationally expensive. As the size and flexibility of the molecule increase, so too does the computational cost of accurately predicting the crystal structures. CSP methods must balance computational efficiency with accuracy, often relying on approximations or ML models to reduce the search space [15].

Despite the challenges, there have been significant advancements in CSP methods. ML and artificial intelligence techniques are now being employed to guide the search for polymorphs by learning from existing databases of crystal structures and predicting the most likely polymorphs for new compounds. Genetic algorithms, which mimic the process of natural selection, have also proven useful in identifying stable polymorphs by iteratively refining possible crystal structures and selecting the most promising candidates [30]. Threshold algorithms have been employed to determine energy barriers between hypothetical crystal structures determining polymorph stability [31].

Additionally, utilising a mixture of faster throughput classical forcefield techniques have been used to ensure thorough searching of a crystal landscape and post QM methods used on low energy structures for greater accuracy of predictions for complex molecules. This approach allows for a more accurate calculation of lattice energies, which is critical in determining the most stable polymorph [32].

CSP therefore allows us to analyse and investigate the crystal structures of molecules before synthesis. In materials discovery, this capability can be used to design substances *in silico* to achieve desired properties and to determine whether synthesis is necessary or appropriate. Our workflow has shown notable success particularly in predicting the structures of rigid molecules [33]. However, predicting the structures of flexible molecules remains a significant challenge due to their high number of degrees of freedom.

# Chapter 3

# Theory, Methods and Programs

This chapter provide an overview of the theoretical frameworks and methodologies employed throughout this thesis.

Section 3.1 discusses DFT, its origins, and its application in performing energy evaluations and geometry optimisations for molecular and crystalline systems. Electronic structure calculations, semi-empirical approaches, and force-field calculations are examined, highlighting their roles in modelling molecular behaviour. Additionally, DMA is described as a technique for assessing molecular conformations and charge distributions.

Section 3.2 focuses on statistical methods, including Principal Component Analysis (PCA) and k-means clustering. These techniques are utilised to reduce data complexity and analyse structural patterns within datasets.

Section 3.3 explores conformer search methods used to identify molecular conformations suitable for crystal packing. Molecular simulation techniques are also discussed, featuring three distinct conformational search tools, RDKit, Conformer Rotamer Ensemble Sampling Tool (CREST), and an Low Mode Conformer Search (LMCS), and their applications in exploring PESs.

Section 3.4 details methods for comparing molecules and molecular crystals. This includes calculations of surface areas and PXRD analysis. The section explains how these methods facilitate structural comparisons and describes approaches for predicting NMR chemical shifts.

## 3.1 Energy Models

### 3.1.1 Hartree-Fock Theory

HF theory is a fundamental method in quantum chemistry for approximating the wavefunctions and energies of many-electron systems. It is an extension of the Hartree method, which itself approximates the many-electron wavefunction as a product of single-electron wavefunctions [34]. In HF theory however, the Pauli exclusion principle is explicitly incorporated using an anti-symmetrised wavefunction, known as a Slater determinant, to describe the electron configuration [35].

#### HF Equations

The starting point of HF theory is the many-electron Schrödinger equation, which describes the total electronic energy of a system:

$$\hat{H}\Psi(\mathbf{r}_1, \mathbf{r}_2, \ldots, \mathbf{r}_N) = E\Psi(\mathbf{r}_1, \mathbf{r}_2, \ldots, \mathbf{r}_N)\,, \tag{3.1}$$

where $\hat{H}$ is the Hamiltonian of the system, $\Psi$ is the many-electron wavefunction, $E$ is the total energy, and $\mathbf{r}_i$ are the positions of the electrons.

As mentioned previously, the wavefunction $\Psi$ is approximated as a Slater determinant of single-electron wavefunctions $\psi_i(\mathbf{r}_i)$, ensuring the antisymmetry property required by the Pauli exclusion principle:

$$\Psi(\mathbf{r}_1, \mathbf{r}_2, \ldots, \mathbf{r}_N) = \frac{1}{\sqrt{N!}} \begin{vmatrix} \psi_1(\mathbf{r}_1) & \psi_2(\mathbf{r}_1) & \cdots & \psi_N(\mathbf{r}_1) \\ \psi_1(\mathbf{r}_2) & \psi_2(\mathbf{r}_2) & \cdots & \psi_N(\mathbf{r}_2) \\ \vdots & \vdots & \ddots & \vdots \\ \psi_1(\mathbf{r}_N) & \psi_2(\mathbf{r}_N) & \cdots & \psi_N(\mathbf{r}_N) \end{vmatrix}. \tag{3.2}$$

The HF method seeks to minimise the energy by varying the orbitals $\psi_i(\mathbf{r}_i)$, subject to the orthonormality constraint, leading to a set of coupled integro-differential equations known as the HF equations [36]:

$$\hat{F}[\psi_i]\psi_i = \epsilon_i\psi_i\,, \tag{3.3}$$

where $\hat{F}$ is the Fock operator, and $\epsilon_i$ are the orbital energies. The Fock operator consists of three terms:

$$\hat{F}[\psi_i] = \hat{h} + \sum_{j=1}^{N} \left( \hat{J}_j - \hat{K}_j \right) , \tag{3.4}$$

where $\hat{h}$ represents the one-electron part of the Hamiltonian, which includes the kinetic energy of the electron and its attraction to the nucleus. The term $\hat{J}_j$, known as the Coulomb operator, accounts for the repulsive interaction between the electron in orbital $\psi_i$ and the electron in orbital $\psi_j$. Finally, $\hat{K}_j$ is the exchange operator, which arises from the antisymmetry of the wavefunction and represents the exchange interaction between electrons with the same spin.

The Fock operator depends on the orbitals $\psi_i$, making the HF equations non-linear. Therefore, the solution requires an iterative procedure known as the Self Consistent Field (SCF) method where the orbitals are updated until the solution converges to a consistent set of orbitals and energies [35].

HF theory explicitly accounts for the exchange interaction which arises due to the antisymmetrisation of the wavefunction and results in the exchange operator $\hat{K}_j$. This is a uniquely QM effect that has no classical analogue [34]. Additionally, HF employs the mean-field approximation, where each electron is considered to move in an averaged potential generated by the other electrons. While this simplifies the interactions, it neglects electron correlation, which refers to the instantaneous, real-time interactions between electrons that are not fully captured in this approximation [35]. Although exchange interactions are accounted for, the dynamic correlation of electron motions is not, which can significantly impact the accurate description of many-electron systems. In turn, this limitation necessitates the development of more advanced methods, such as post-HF techniques and DFT, which provide better treatment of electron correlation [37].

### 3.1.2 Density Functional Theory

One such method is DFT, which approaches the problem of electron correlation differently. Instead of focusing on the wavefunction, DFT expresses the energy of a system as a function of the electron density, $\rho(\mathbf{r})$, greatly simplifying the treatment of many-electron systems [37]. DFT therefore offers a more computationally efficient way to incorporate electron correlation effects through exchange-correlation functionals, which approximate both the exchange and correlation energies. In the following section, we will discuss the principles of DFT and how it addresses some of the limitations inherent in HF theory, particularly with respect to electron correlation.

In quantum mechanics, we can obtain the electron density of a system by taking the square modulus of the wavefunction and integrating over the coordinates of $N-1$ electrons:

$$\rho(\mathbf{r}) = \int |\Psi(\mathbf{r}_1, \mathbf{r}_2, \ldots, \mathbf{r}_N)|^2 \, d\mathbf{r}_2 \ldots d\mathbf{r}_N. \tag{3.5}$$

According to the Schrödinger equation:

$$E \left|\Psi\right\rangle = \hat{H} \left|\Psi\right\rangle . \tag{3.6}$$

Evaluation of the electronic Hamiltonian operator given by the Schrödinger equation yields the total energy of the system, which is composed of the sum of its kinetic and potential parts.

$$E_{total} = E_{kinetic} + E_{potential} \tag{3.7}$$

The potential terms are derived from electron–nuclei, electron–electron, and nuclei–nuclei interactions:

$$H = -\frac{1}{2} \sum_{i=1}^{N_{\text{elec}}} \nabla_i^2 - \sum_{a=1}^{N_{\text{nuclei}}} \sum_{i=1}^{N_{\text{elec}}} \frac{Z_a}{|\mathbf{R}_a - \mathbf{r}_i|} + \sum_{i=1}^{N_{\text{elec}}} \sum_{j>i}^{N_{\text{elec}}} \frac{1}{|\mathbf{r}_i - \mathbf{r}_j|} + \sum_{a=1}^{N_{\text{nuclei}}} \sum_{b>a}^{N_{\text{nuclei}}} \frac{Z_a Z_b}{|\mathbf{R}_a - \mathbf{R}_b|} , \tag{3.8}$$

where $\mathbf{r}$ and $\mathbf{R}$ are the positions of an electron and a nucleus respectively, $Z$ is the charge of the nucleus, and $\nabla$ is the Laplacian.

Due to the computationally intensive solution to the equation, it is not feasible to solve for a system of more than a few electrons; therefore we make assumptions to reduce the computational cost without significant loss in accuracy. If we apply the Born-Oppenheimer approximation, we treat electrons as QM objects and nuclei as point charges. Therefore, we treat nuclei-nuclei interactions as a function of atomic positions which can enable the solving of the electronic energy with the nuclear energy as constant. This enables the determination of the total energy for that set of atomic positions. Consequently, this allows for energy minimisation by varying the atomic positions until the energy is minimised.

If we assume that we can take two different external potentials $V_{\text{ext}}$ and $V'_{\text{ext}}$ which describe the electron–nuclei interactions that result in the same electron density $\rho$, we determine that there must also be two wavefunctions associated with these potentials, $\Psi$ and $\Psi'$. If we take $\Psi'$ as an approximate wavefunction for $H$, we obtain from the variational principle that:

$$\langle \Psi' | H | \Psi' \rangle > E_0 \tag{3.9}$$

$$\langle \Psi' | H' | \Psi' \rangle + \langle \Psi' | H - H' | \Psi' \rangle > E_0 \tag{3.10}$$

$$E_0' + \langle \Psi' | V_{\text{ext}} - V_{\text{ext}}' | \Psi' \rangle > E_0 \tag{3.11}$$

$$E_0' + \int \rho(\mathbf{r}) \Big( V_{\text{ext}} - V_{\text{ext}}' \Big) \, d\mathbf{r} > E_0 \tag{3.12}$$

Similarly, if we take $\Psi$ as an approximate wavefunction of $H'$, we get:

$$E_0' - \int \rho(\mathbf{r}) \Big( V_{\text{ext}} - V_{\text{ext}}' \Big) \, d\mathbf{r} > E_0. \tag{3.13}$$

Through addition of Equations 3.12 and 3.13, we find that:

$$E_0' + E_0 > E_0' + E_0. \tag{3.14}$$

This equation states that the assumption that any two potentials can create the same electron density is incorrect, and therefore the energy is a unique function of the electron density. Therefore, we can bypass the wavefunction, providing a basis for DFT.

We also know that the approximate electron density integrates to the number of electrons in the system:

$$\int \rho(\mathbf{r}) \, d\mathbf{r} = N_{\text{elec}}. \tag{3.15}$$

For any density that is not the ground-state density:

$$E_0[\rho'] \geq E_0[\rho] \, . \tag{3.16}$$

Finally, there is a variational way in which one can solve for the electron density, which we will discuss in Kohn-Sham theory.

### Kohn–Sham Theory

Kohn–Sham theory is an approach that reformulates an $N$-electron problem as multiple single-particle problems. This is achieved by introducing a set of orbitals that represent a system of non-interacting electrons. A correction term is then added to account for the difference between this model and the true interacting system. We can thus rewrite

the Schrödinger equation in terms of the electron density. The Kohn–Sham functional is therefore expressed as:

$$E[\rho] = T_s[\rho] + E_{\mathrm{ne}}[\rho] + E_{\mathrm{ee}}[\rho] + E_{\mathrm{xc}}[\rho] \,, \tag{3.17}$$

where $T_s[\rho]$ is the kinetic energy of the non-interacting electrons, $E_{\mathrm{ne}}[\rho]$ is the nucleus–electron interaction energy functional, and $E_{\mathrm{ee}}[\rho]$ is the electron–electron interaction energy functional. $E_{\mathrm{xc}}[\rho]$ is the exchange–correlation functional, which contains both the kinetic energy not accounted for by $T_s$ and the correlation contributions arising from electron dynamics. This functional can be approximated in various ways, forming the basis of many different DFT methods. Finally, $E_{\mathrm{xc}}$ can be further decomposed into the exchange energy $E_{\mathrm{x}}$ and the correlation energy $E_{\mathrm{c}}$.

### 3.1.3 Basis Sets

Basis sets are mathematical functions used to represent the electronic wavefunctions of atoms and molecules. By using basis sets, complex partial differential equations like the Schrödinger equation can be converted into algebraic ones, making the computational solving of QM problems feasible. A molecular wavefunction, which describes the distribution of electrons in a molecule, is often approximated as a linear combination of fixed sets of mathematical functions called basis functions.

The wavefunction $\Psi_i(\mathbf{r})$ of a molecular orbital can be expressed as a linear combination of basis functions $G_\alpha(\mathbf{r})$, each representing an atomic orbital. This relationship is written as:

$$\Psi_i(\mathbf{r}) = \sum_{\alpha=1}^{N_{\mathrm{BF}}} G_\alpha(\mathbf{r})\, C_{\alpha i} \,, \tag{3.18}$$

where $\Psi_i(\mathbf{r})$ is the molecular orbital, $N_{\mathrm{BF}}$ is the total number of basis functions, $G_\alpha(\mathbf{r})$ is a basis function, and $C_{\alpha i}$ are the coefficients that determine the contribution of each basis function to the molecular orbital [38].

The exact solution for the wavefunction of the hydrogen atom is represented by Slater Type Orbital (STO)s, which provide an accurate description of atomic orbitals due to their ability to capture the electron behaviour near the nucleus. A Slater function for an atomic orbital is expressed as:

$$s_\nu(r) = e^{-\zeta_\nu |r|} = e^{-\zeta_\nu \sqrt{x^2+y^2+z^2}} \,, \tag{3.19}$$

where $\zeta_\nu$ is the orbital exponent that controls the spread of the orbital. While STOs are physically accurate, they are computationally inefficient because the integrals needed to solve molecular problems are challenging to compute [39].

To improve computational efficiency, Gaussian Type Orbitals (GTOs) are often used instead, which can be expressed as:

$$g_\nu(\mathbf{r}) = e^{-\zeta_\nu r^2} = e^{-\zeta_\nu\left(x^2+y^2+z^2\right)} \,. \tag{3.20}$$

Gaussian functions decay more rapidly than Slater functions due to the quadratic dependence on $r$, and they lack the nuclear cusp. However, GTOs are preferred in most quantum chemistry calculations because their integrals are easier to compute owing to the Gaussian product theorem [40].

### Contracted and Primitive Gaussians

To regain some of the accuracy lost by using Gaussian functions, a set of primitive Gaussians is combined to form a contracted Gaussian. A contracted Gaussian is a linear combination of several primitive Gaussian functions with different exponents. By using multiple Gaussians with varying spreads, the contracted Gaussian can approximate the behaviour of an atomic orbital more accurately [41].

$$g_\nu(\mathbf{r}) = x^k\, y^m\, z^n\, e^{-\zeta_\nu r^2} = x^k\, y^m\, z^n\, e^{-\zeta_\nu\left(x^2+y^2+z^2\right)} \,, \tag{3.21}$$

where $k$, $m$, and $n$ are the powers of the Cartesian coordinates, and $l = k + m + n$ is the angular momentum quantum number. A contracted Gaussian is formed by linearly combining multiple primitive Gaussians as follows:

$$G_a(\mathbf{r}) = \sum_{\nu=1}^{N_\alpha} g_\nu(\mathbf{r})\, c_\nu \,, \tag{3.22}$$

where $G_a(\mathbf{r})$ is the contracted Gaussian, $g_\nu(\mathbf{r})$ are the primitive Gaussians, $c_\nu$ are the contraction coefficients, and $N_\alpha$ is the number of primitive functions used in the contraction.

Basis sets are composed of these contracted Gaussian functions and are tailored to provide an appropriate balance between computational efficiency and accuracy. Minimal basis sets use a small number of Gaussian functions to describe each atomic orbital while extended basis sets, such as double-zeta or triple-zeta, use multiple functions for each orbital to better describe the flexibility of electron distributions.

### 3.1.4 Pople Basis Sets

Pople basis sets, developed by John A. Pople and collaborators, represent a widely used family of Gaussian-type basis sets in quantum chemistry. These basis sets are designed to balance computational efficiency and accuracy, particularly for HF and DFT calculations [42]. The flexibility and performance of Pople basis sets make them useful for describing electronic structures, especially in the context of molecular interactions and chemical reactions.

Pople basis sets use a split valence approach which distinguishes between core and valence electrons, allowing different levels of precision for their description. Core electrons, which are closer to the nucleus and less involved in bonding, are represented by fewer Gaussian functions. Since core orbitals undergo smaller distortions in various chemical environments, fewer functions suffice for an accurate description. Conversely, valence orbitals, which play a significant role in bonding, are represented by multiple Gaussian functions. This added flexibility captures changes in electron distribution during bonding and chemical reactions [43].

Core orbitals are typically described by contracted Gaussians which are combinations of multiple primitive Gaussian functions. This reduces the computational cost while retaining an accurate representation of core orbitals which are tightly bound and less affected by chemical reactions. On the other hand, valence orbitals are described using split functions with multiple Gaussian functions of varying exponents. This increased flexibility is crucial for valence orbitals as they adapt more readily to different chemical environments during bonding [44].

Further refinements to Pople basis sets involve the inclusion of polarisation functions, denoted by symbols such as (d), (p), or (f) following the basis set name. For example, 6-31G(d) or 6-31G(d,p) incorporates polarisation functions to account for angular distortions in the electron cloud. These functions are crucial for accurately describing complex electron distributions, including those associated with lone pairs or external fields [45]. Additionally, diffuse functions, indicated by a plus sign (+), may be appended to Pople basis sets. These functions, characterised by small exponents, are used to represent loosely bound electrons, which play a significant role in systems such as anions or excited states [46]. While these enhancements improve the accuracy of calculations, they also increase computational cost [47].

### 3.1.5 Gaussian09

Gaussian09 is a computational chemistry program that allows us to perform (a) single-point energy calculations using QM methods and (b) geometry optimisations [48].

a) A single-point energy calculation determines the energy of a molecule at a fixed geometry, without altering its atomic structure. This type of calculation is useful for evaluating the energy of a system as a reference point for different molecular conformations. Gaussian performs calculations based on parameters defined by the user, including the choice of basis set, computational method, dispersion corrections, and other settings. In single-point energy calculations, Gaussian computes the electronic wavefunction using the selected method by solving the SCF equations. This yields the ground-state electronic density and the corresponding energy of the system. Once the wavefunction is established, the program computes the total energy of the system, which includes contributions such as electronic energy, nuclear repulsion, and exchange-correlation. The final output is a single energy value, often used as a comparative reference.

b) Geometry optimisation involves adjusting the atomic positions of a molecule to find its local minimum energy geometry. Gaussian performs a single-point energy calculation to establish the initial energy and forces on the atoms. It then calculates the forces on each atom, which correspond to the negative gradient of the energy with respect to atomic positions. The gradient indicates the direction in which the energy decreases most rapidly. Utilising the Berny optimisation algorithm, Gaussian iteratively adjusts the atomic coordinates, moving the system toward a lower energy configuration [49]. The algorithm evaluates step size and direction based on the forces and the Hessian matrix, which is the second derivative of energy.

This process continues until convergence criteria are met. These criteria include the maximum force, maximum displacement of atoms, and changes in energy between iterations falling below specified thresholds. The final output provides the optimised geometry and corresponding energy.

Whilst the approach is often successful for many systems, geometry optimisation can encounter challenges such as convergence issues if the initial structure is far from a minimum or if the system has a complex PES.

### 3.1.6 Density Functional Tight Binding

Density Functional Tight Binding (DFTB) is a simplified version of DFT that approximates total energy through a second-order expansion in charge density fluctuations, thereby reducing the complexity of solving the Kohn-Sham equations, a fundamental component of DFT. Instead of explicitly calculating all electronic integrals, DFTB employs pre-computed parameters often derived from DFT calculations to describe the Hamiltonian matrix elements between atomic orbitals. This approach imparts a 'tight-binding' characteristic, allowing DFTB to retain essential electronic structure information without the computational expense of full DFT. DFTB effectively captures chemical bonding effects by considering interactions between atoms up to two and three

centres, providing reasonable accuracy in describing molecular and material properties. Its reliance on parametrised terms, rather than complex exchange-correlation energy calculations, enables significantly better scaling with system size, making it a valuable tool for studying complex chemical and material systems that are typically challenging for traditional DFT methods.

An enhanced version, known as DFTB+, incorporates several extensions to the original method, addressing some limitations, such as the need for extensive parameterisation similar to force fields [50]. The Hamiltonian matrix elements are defined in an element pair-wise manner, which requires thousands of empirical parameters, limiting the method's applicability to a broader range of elements in the periodic table. Although DFTB combines the efficiency of earlier minimal basis set methods with the improved accuracy of DFT, surpassing HF theory in performance, the use of small minimal atomic orbital basis sets restricts the accurate representation of simplified Kohn-Sham equations. This can lead to inaccuracies in predicting certain properties, such as chemical bond energies, necessitating further optimisation with tools. This issue is not unique to DFTB but is common across all semi-empirical methods.

### 3.1.7 xTB

The limitations of tight-binding methods have led to the development of Extended Tight Binding (xTB) methods, specifically designed to accurately describe molecular properties such as geometries, vibrational frequencies, and non-covalent interactions for up to thousands of atoms [51]. Whilst also employing a tight-binding approach, xTB is empirically parametrised to be more versatile, handling a broader range of elements and environments. This makes xTB more transferable across diverse chemical systems, including organic, inorganic, and transition metal complexes.

To solve for the electronic structure, xTB employs the SCF method, similar to other QM approaches like HF and DFT. The process begins with an initial guess for the electron density or molecular orbitals. Using the tight-binding Hamiltonian, xTB calculates the electronic structure of the system. The electron density is then updated based on the results of the electronic structure calculation, and this process is repeated iteratively until the electron density converges.

One of the major advantages of xTB is its speed, as it is much faster than ab initio methods. This makes xTB highly suitable for large systems and high-throughput calculations, where computational efficiency is critical. xTB also scales well, allowing it to handle molecular systems with hundreds or even thousands of atoms.

GFN1-xTB uses a similar approximation scheme to DFTB, primarily employing second-order terms with some third-order corrections, but without relying on element pair-wise parameterisation. This approach enables consistent parameterisation across a large portion of the periodic table, covering elements with proton numbers up to 86.

GFN2-xTB incorporates advanced physics, including a multipole electrostatic treatment up to quadrupole terms and the latest D4 dispersion model, eliminating the need for pair-specific parameterisation [52]. It offers improved accuracy and is computationally more efficient by avoiding the Self Consistent Charge (SCC) iterations that are typically the computational bottleneck in most semi-empirical QM methods.

### 3.1.8 Periodic DFT

In band theory, as implemented in codes like Vienna Ab initio Simulation (VASP), solids are modelled as collections of nuclei and electrons, with the electronic structure determined by solving the Schrödinger equation. This equation accounts for the interaction of electrons with both the nuclei and each other. Within this framework, chemical bonding emerges naturally as a property of the system's ground state. Both bonded and non-bonded electrons are treated equally, and the interaction between electrons and the external potential is derived directly from solving the equation.

Band theory can employ a simplified model that neglects electron–electron interactions, known as the one-electron model, where the Schrödinger equation is written as:

$$E\,\Phi = -\frac{\hbar^2}{2m_e}\,\nabla^2\Phi + V(\mathbf{r})\,\Phi\,, \tag{3.23}$$

where $E$ is the energy eigenvalue of the electron, $\hbar$ is the reduced Planck constant, $m_e$ is the electron mass, $\Phi(\mathbf{r})$ is the electron wavefunction, and $V(\mathbf{r})$ is the external potential experienced by a single electron at position $\mathbf{r}$.

This potential typically arises from the atomic nuclei and is treated as static. Since the potential $V(\mathbf{r})$ acts independently on each electron and depends only on spatial position, this formulation represents the most basic version of the Schrödinger equation used in band theory.

To account for electron–electron interactions, the one-electron formulation is extended to more complex systems using methods like HF or DFT. In DFT, the many-body Schrödinger equation is reformulated to include electron–electron interactions within the potential. DFT simplifies the problem by mapping the many-electron system onto an auxiliary system of non-interacting particles that experience an effective potential, making it computationally feasible to study large systems while accurately accounting for these interactions.

In periodic systems such as crystals, the nuclear (external) potential is a sum of Coulomb potentials from the nuclei, expressed as:

$$V(\mathbf{r}) = \sum_i V_i\Big(\mathbf{r} - \mathbf{R}_i\Big) = \sum_i \frac{Z_i}{|\mathbf{r} - \mathbf{R}_i|}\,, \tag{3.24}$$

where $Z_i$ represents the charge of nucleus $i$ at position $\mathbf{R}_i$, and $\mathbf{r}$ is the position of the electron. For a periodic crystal, the potential repeats itself at intervals determined by the crystal's translation vectors $\mathbf{T}_j$, making the potential periodic:

$$V(\mathbf{r} + \mathbf{T}_j) = V(\mathbf{r})\,. \tag{3.25}$$

The index $j$ labels all combinations of integer multiples of the reciprocal lattice basis vectors, effectively enumerating all the plane waves consistent with the crystal's periodicity. This periodicity enables the potential to be expressed as a Fourier series:

$$V(\mathbf{r}) = \sum_{\mathbf{G}_j} V(\mathbf{G}_j)\, e^{i\,\mathbf{G}_j\cdot\mathbf{r}}\,, \tag{3.26}$$

where $\mathbf{G}_j$ are the reciprocal lattice vectors, which form a reciprocal lattice corresponding to the real-space lattice defined by $\mathbf{T}_j$.

In a crystal with real-space lattice vectors $\mathbf{a}$, $\mathbf{b}$, and $\mathbf{c}$, the corresponding reciprocal lattice vectors $\mathbf{a}^*$, $\mathbf{b}^*$, and $\mathbf{c}^*$ are given by:

$$\mathbf{a}^* = \frac{2\pi}{V}\,\mathbf{b}\times\mathbf{c}\,, \tag{3.27}$$

**Plane-Wave Basis in Periodic Systems**

In plane-wave DFT, the wavefunctions are expanded as a series of plane waves. The electron wavefunction is approximated by a plane-wave trial function:

$$\mathbf{\Phi}_\mathbf{k}(\mathbf{r}) = c_\mathbf{k}\, e^{i\,\mathbf{k}\cdot\mathbf{r}}\,, \tag{3.28}$$

where $\mathbf{k}$ is the wavevector and $c_\mathbf{k}$ is a coefficient. Substituting this trial function into the Schrödinger equation for free space gives:

$$-\frac{\hbar^2}{2m_e}\,\nabla^2\mathbf{\Phi}_\mathbf{k}(\mathbf{r}) = \frac{\hbar^2\,|\mathbf{k}|^2}{2m_e}\,\mathbf{\Phi}_\mathbf{k}(\mathbf{r})\,. \tag{3.29}$$

Thus, the term $\frac{\hbar^2 |\mathbf{k}|^2}{2m_e}$ is the eigenvalue corresponding to the eigenstate $\Phi_{\mathbf{k}}(\mathbf{r})$.

#### First Brillouin Zone and Zone Folding

In periodic systems, the wavevector $\mathbf{k}$ takes values in the range $\frac{-\pi}{a} < \mathbf{k} \leq \frac{\pi}{a}$. The wavefunction takes the form of a Bloch function, combining a plane wave with a periodic function:

$$\Phi_{\mathbf{k}}(\mathbf{r}) = e^{i\,\mathbf{k}\cdot\mathbf{r}}\, u_{\mathbf{k}}(\mathbf{r})\,. \tag{3.30}$$

Here, $u_{\mathbf{k}}(\mathbf{r})$ is a periodic function with the same periodicity as the crystal lattice, and the phase factor $e^{i\,\mathbf{k}\cdot\mathbf{r}}$ reflects the translational symmetry of the lattice. The function $u_{\mathbf{k}}(\mathbf{r})$ can itself be expanded in a Fourier series:

$$u_{\mathbf{k}}(\mathbf{r}) = \sum_{\mathbf{G}_j} c_{\mathbf{k}+\mathbf{G}_j}\, e^{i\,\mathbf{G}_j\cdot\mathbf{r}}\,. \tag{3.31}$$

This Fourier expansion of $u_{\mathbf{k}}(\mathbf{r})$ allows for the treatment of periodicity and variations in the electron wavefunction within the crystal lattice.

### 3.1.9 Vienna Ab initio Simulation Package

The VASP software package performs periodic DFT calculations, enabling high-accuracy crystal structure optimisations. While DMACRYS (described in section 3.1.10) can minimise the lattice energy of crystal structures generated from CSP using rigid molecular conformations, further energy minimisation is possible by relaxing the molecular conformations within the unit cell [53–56]. VASP can be utilised to perform periodic DFT on these crystal structures, providing more accurate energy calculations. This is achieved by assuming translational invariance through periodic boundary conditions to minimise the crystal structure.

In studies presented within this thesis, VASP optimisation begins with a fixed-cell relaxation, during which the atoms within the unit cell are allowed to relax while the cell dimensions remain constant. Upon reaching a minimum-energy configuration, a variable-cell relaxation follows, permitting changes in the unit cell's shape and volume. Finally, a single-point energy calculation is performed on the fully relaxed structure, employing a higher plane-wave energy cut-off, a denser k-point mesh, and tighter convergence criteria to achieve a more accurate total energy.

Plane-wave DFT codes like VASP are generally preferred for periodic DFT due to their inherent periodic nature. This formalism allows for efficient computation of electronic

band structures, especially in systems with well-defined crystal geometries. The scalability of plane-wave DFT is advantageous for large systems, especially when combined with pseudo potentials that reduce computational complexity by focusing on valence electrons, simplifying the treatment of core electrons.

However, a notable drawback of the plane-wave approach is the substantial number of plane waves required to accurately represent localised states, such as those found in molecules or defects within crystals. This can make the method computationally intensive for systems with significant electron localisation.

### 3.1.10 DMACRYS

The crystal structures generated by the CLG are minimised using DMACRYS [57]. This program calculates the lattice energies of crystal structures by summing both van der Waals and electrostatic interactions between molecules. DMACRYS operates under a rigid-body approximation, treating molecules as fixed units during optimisation. This ensures that the overall crystal structure is energy-minimised with respect to both translational and rotational degrees of freedom.

In this thesis, repulsion and dispersion forces are modelled using the Buckingham potential, which is widely used for molecular systems and is described by Equation 3.32:

$$\phi_{ab}(r) = A\, e^{-Br} - \frac{C}{r^6}\,, \tag{3.32}$$

where $A$, $B$, and $C$ are constants, and $r$ is the interatomic distance between points $a$ and $b$. The exponential term models the repulsive interaction at short ranges, while the second term accounts for the attractive van der Waals interaction, characterised by a power-law decay proportional to $r^{-6}$ at long distances.

The electrostatic interactions are modelled using Coulomb's law, representing the force between point charges in the system. The electrostatic force between two charges $q_a$ and $q_b$ separated by distance $r$ is given by Equation 3.33:

$$F = k_e \frac{q_a q_b}{r^2}\,, \tag{3.33}$$

where $F$ is the electrostatic force, $k_e$ is Coulomb's constant $9 \times 10^9\,\mathrm{N m^2 C^{-2}}$, $q$ is the charge, and $r$ is the distance between points $a$ and $b$.

However, atoms and molecules are anisotropic in nature, meaning that they are not simple spherical point charges. The charge distribution is more complex, particularly in molecules, due to electron clouds and chemical bonds. This anisotropy necessitates a more accurate representation of the electrostatic interactions using multipole expansion.

#### 3.1.10.1 Distributed Multipole Analysis (DMA)

In DMACRYS, the electrostatic interactions are modelled using distributed multipoles, which include dipoles, quadrupoles, and higher-order terms. These are computed via the DMA method [58], which decomposes the molecular charge distribution into a set of atomic multipoles, allowing for a more accurate representation of the electrostatic potential and energy.

The use of multipole expansion extends the accuracy of the electrostatic modelling beyond simple charge-charge interactions, allowing DMACRYS to capture directional effects and polarisation that are crucial in determining the correct packing and interactions in crystal structures. This method is particularly useful for complex molecules where hydrogen bonding, polar groups, and non-spherical electron distributions play a significant role in the stabilisation of a crystal structure. In this thesis, a rigid-body approximation has been used.

Conventional multipole analysis represents atoms or molecules as point multipoles of rank $k$, where $k = 0$ corresponds to a monopole, $k = 1$ to a dipole, $k = 2$ to a quadrupole, and so on. For example, a water molecule can be approximated as a dipole, which is sufficient for simulating systems with large spatial separations. However, this model becomes inaccurate for short-range interactions, such as hydrogen bonding, where water would behave unrealistically. DMA provides a more detailed representation of a molecule's charge distribution, by modelling atoms up to hexadecapoles.

### 3.1.11 GDMA

GDMA software generates atom-centred multipoles to reproduce the electrostatics of a molecule based on data derived from Gaussian calculations [58].

An electronic wavefunction describes the distribution of electrons in a molecule. The multipole expansion expresses this distribution as a series of terms, each representing a different order of spatial distribution. The first term in this series is the monopole moment, denoted as $q$, which represents the total charge at a point. For an atom, this corresponds to the net electronic charge. The second term is the dipole moment, $\mu$, which represents the first-order distribution of charge, indicating how the charge is polarised or how the positive and negative charges are separated. Following this, the quadrupole moment, $Q$, describes the second-order distribution, providing information about the shape of the charge, such as its elongation or compression. Higher-order multipoles, such as octapoles and hexadecapoles, offer even more refined details of the charge distribution.

The total potential $V(r)$ due to a charge distribution at a distance $r$ from the atom centre can be expanded in terms of these multipole moments. This expansion is given by:

$$V(r) = \frac{1}{r}\left(q + \frac{\mathbf{r}\cdot\boldsymbol{\mu}}{r^2} + \frac{1}{2}\sum_{i,j}\frac{r_i\, r_j\, Q_{ij}}{r^4} + \dots\right), \tag{3.34}$$

where $r_i$ and $r_j$ are the $i^{\text{th}}$ and $j^{\text{th}}$ components of the vector $\mathbf{r}$, $r$ is its magnitude, $q$ and $\boldsymbol{\mu}$ represent the monopole and dipole moments respectively, and $Q_{ij}$ represents the components of the quadrupole moment tensor, describing how the charge is distributed in a non-spherical manner around the centre.

DMA distributes these multipole moments over atom centres. This process involves partitioning the total electron density into localised regions associated with each atom. For each of these regions, the local multipole moments, such as charge, dipole, and quadrupole, are calculated by integrating the electron density over the localised region using basis functions. The original electronic wavefunction or density is then represented as a sum of these local multipole moments, offering a detailed picture of the charge distribution throughout the molecule.

### 3.1.12 MULFIT

One issue with GDMA is that it can generate high-rank multipoles, which increase computational cost. The program MULFIT addresses this by refitting the multipoles to lower ranks without significantly altering the electric potential [59, 60].

As the multipole expansion of the series converges, some high-order terms may become redundant as these multipoles provide finer details of the charge distribution. Reducing the rank of a multipole minimises the use of high-order multipoles while preserving the essential features of the charge distribution. Therefore these terms can be dropped without drastically altering the charge distribution and reduce the cost of DMA.

MULFIT performs an orthogonalisation procedure to transform the multipoles into a set of orthogonal components. This transformation helps to identify linearly dependent or redundant multipoles, which can then be eliminated or merged. After this orthogonalisation, MULFIT fine-tunes the remaining multipoles to best reproduce the target properties, such as the electrostatic potential and dipole moment, using the reduced set of multipoles.

MULFIT achieves this by minimising the error $\epsilon$ between the reference potential $V_{\text{ref}}(r)$ and the potential $V_{\text{fit}}(r)$ generated by the fitted multipoles:

$$\epsilon = \sum_{r} \left( V_{\text{ref}}(r) - V_{\text{fit}}(r) \right)^2 \ . \tag{3.35}$$

where $V_{\text{fit}}(r)$ is the summation of calculated as:

$$V_{\text{fit}}(r) = \sum_{i} \frac{q_i}{|r - R_i|} + \sum_{i} \frac{\mu_i \cdot (r - R_i)}{|r - R_i|^3} + \sum_{i} \frac{Q_i \cdot (r - R_i)(r - R_i)}{|r - R_i|^5} + \dots \ , \tag{3.36}$$

where $q_i$, $\mu_i$, and $Q_i$ represent the monopole, dipole and quadrupole moments, respectively, located at centre $R_i$. The electrostatic potential is evaluated at a set of grid points $r$, and the total fitted potential $V_{\text{fit}}(r)$ is constructed as a sum of contributions from each multipole at each site $i$. MULFIT optimises these parameters to best match the reference potential while systematically reducing the rank and eliminating insignificant terms.

## 3.2 Statistical Methods

### 3.2.1 Euclidean Distance

In this work, we use the Euclidean distance to describe the real distance between any two points. The Euclidean distance is a metric used to calculate the distance between two objects in $n$-dimensional space:

$$d(A,B) = \sqrt{\sum_{j=1}^{n} (x_{jA} - x_{jB})^2}\,, \tag{3.37}$$

where $x_{jA}$ and $x_{jB}$ represent the values of the $j$-th feature for points $A$ and $B$, respectively.

### 3.2.2 Principal Component Analysis

PCA, first introduced by Pearson [61] and later refined by Hotelling [62], is an unsupervised dimensionality reduction technique that uses feature variation to represent data while preserving the global structure of the dataset [63]. PCA works by identifying new axes, known as principal components, which are linear combinations of the original features. These components capture the maximum variance in the data with the first principal component accounting for the largest variance, the second capturing the next highest variance and so on.

The data is first standardised to ensure that each feature contributes equally, especially when the variables are measured in different units. $Z_{ij}$ is the standardised value of data point $X_{ij}$ where $i$ is the data point of feature $j$. This can be calculated using:

$$Z_{ij} = \frac{X_{ij} - \mu_j}{\sigma_j}\,, \tag{3.38}$$

where $\mu_j$ is the mean and $\sigma_j$ is the standard deviation of the $j^{th}$ feature.

PCA computes the covariance matrix to assess how much each feature varies in relation to others. The covariance between two features, $j$ and $k$, is defined as:

$$C_{jk} = \frac{1}{n-1} \sum_{i=1}^{n} (Z_{ij} - \mu_i)(Z_{ik} - \mu_j)\,, \tag{3.39}$$

where $C_{jk}$ represents the covariance between feature $j$ and $k$, while $n$ is the number of data points within each feature. The covariance matrix $C$ is a square matrix, with each element describing the covariance between a pair of variables.

Once the covariance matrix is obtained, the eigenvectors and eigenvalues of $C$ are calculated. Eigenvalue decomposition is performed similarly to as described in section 3.3.2.

$$C = V\Lambda V^T \,. \tag{3.40}$$

In this equation, $V$ is a matrix whose columns are the eigenvectors, representing the directions of maximum variance in the data. $\Lambda$ is a diagonal matrix containing the corresponding eigenvalues, which indicate the amount of variance described by each eigenvector.

The dataset is then transformed onto a new coordinate system which is the projection onto the principal components.

$$X_{\text{new}} = ZV \,, \tag{3.41}$$

where $Z$ is the standardised dataset and $X_{\text{new}}$ is the dataset expressed in the new coordinate system defined by these eigenvectors. Each row in $X_{\text{new}}$ represents the original data point, but re-expressed as a combination of the new axes known as principal components.

Each principal component is defined by an eigenvector corresponding to an eigenvalue $\lambda_1, \lambda_2, \ldots$ of the data's covariance matrix. The eigenvalues quantify the amount of variance in the original dataset captured by each component. The proportion of variance explained by a principal component is determined by dividing its eigenvalue by the sum of all eigenvalues. Components associated with the highest eigenvalues are considered the most informative for representing the structure of the data.

The proportion of variance explained by each component can be visualised using a scree plot, shown in Figure 3.1.

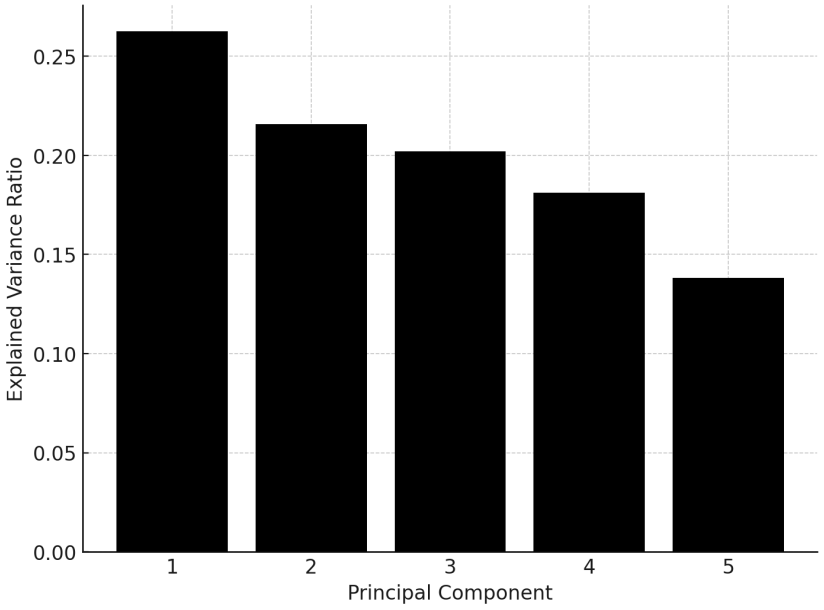



FIGURE 3.1: Example scree plot illustrating the proportion of variance explained by each principal component. A principal component accounting for a large amount of variance conveys more information about the dataset.

This plot helps determine how many components to retain after performing PCA. When a small number of components explain a large proportion of the total variance, effective data approximation is achievable. Conversely, if the variance is distributed across many components, dimensionality reduction may result in the loss of important information. The cumulative variance explained by the selected components indicates how much of the original data's variability is preserved. For example, if the first three components account for 90% of the variance, they provide a reliable approximation of the original data.

#### 3.2.2.1 Loading Scores

Loading scores represent the weights or coefficients assigned to each original feature when forming a principal component. By analysing the loading scores of each of these, it is possible to understand which features have the most influence in shaping that component. This provides insight into the structure of the data and reveals how the principal components are constructed from the original variables.

#### 3.2.2.2 Geodesic Principal Component Analysis

When clustering molecular conformations based on torsion angles, Geodesic PCA is more appropriate than PCA due to the nature of angular data. Torsion angles are inherently circular, meaning that they exist on a continuous loop. This circularity

presents significant challenges for standard linear methods like PCA, which assume that data is distributed in a Euclidean space. Geodesic PCA, however, operates on manifolds and accurately calculates geometric distances between points in circular space, ensuring correct representation of such data.

Geodesic PCA has been widely applied in the study of molecular conformations, particularly in Ribonucleic Acid (RNA) and protein research, where torsion angles dominate the conformational flexibility of the molecules [64, 65].

### 3.2.3 k-means Clustering

k-means clustering is an unsupervised machine learning technique based on Lloyd's or Elkan's algorithm [66, 67] which assigns data points into $k$ clusters where $k$ is an input parameter. These clusters can be used to assign unlabelled data into groups and categories.

The algorithm works by randomly assigning a data point to each of the $k$ clusters which act as each clusters initial "centroid". All other data points are then assigned to the cluster of the centroid closest to it based on Euclidean distance.

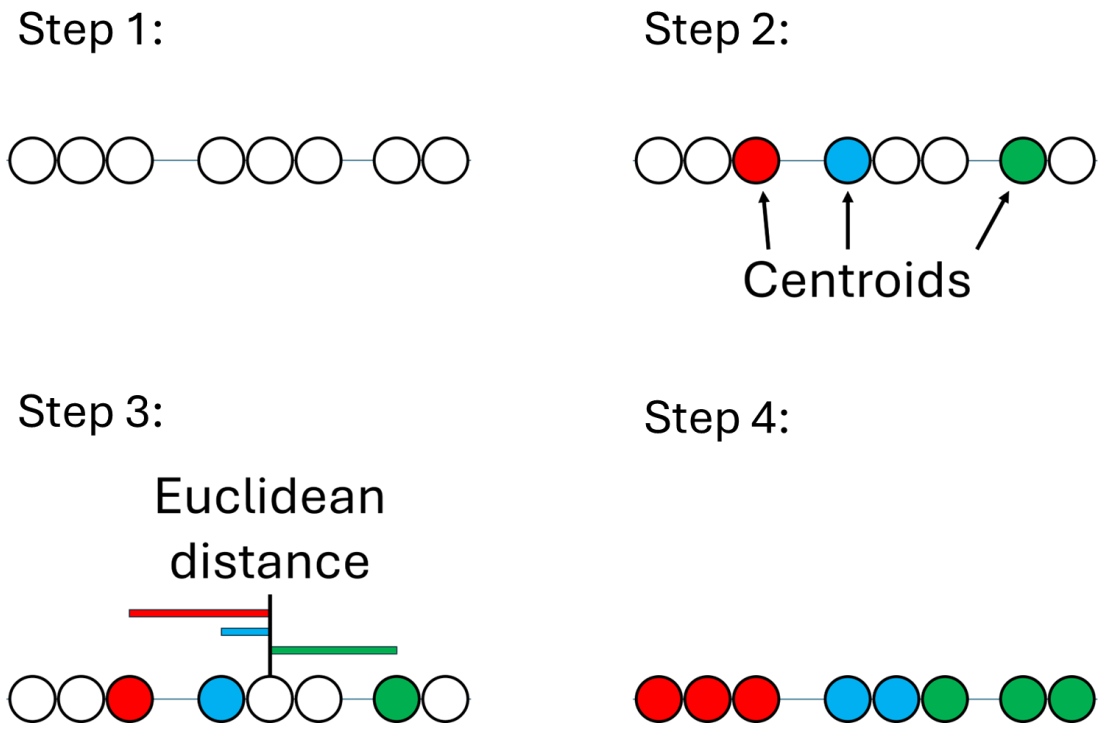



FIGURE 3.2: Stages of k-means clustering for a linear dataset. Stage 1 describes the initial unlabelled linear dataset, stage 2 describes how centroids are randomly assigned to a random data point, stage 3 describes assigning the data point to the cluster of the nearest centroid and stage 4 describes the final labelled linear dataset.

The effectiveness of the clustering process can measured by calculating the sum of square distances between all clusters. This metric describes overall, how well each data point has been assigned to a particular cluster.

$$V = \sum_{i=1}^{k} \sum_{n=1}^{N_i} d(x_{in}, c_i)^2 \,, \tag{3.42}$$

where $i$ is the $i^{th}$ cluster, $x_{in}$ is the $n^{th}$ of data point belonging to cluster $i$, $d$ is the euclidean distance, $N_i$ is the number of data points in cluster $i$ and $c_i$ the centroid of cluster $i$.

As a small value of $V$ suggests better clustering, we wish to minimise this value using an iterative approach. This will identify the optimum clustering arrangement which reduces the sum of square distances across all clusters.

Once all data points have been assigned, the centroid of each cluster is reassigned to be the mean of that cluster. Again, the $V$ is calculated for the new centroids and the process repeated until there is no change to the assignment of the clusters.

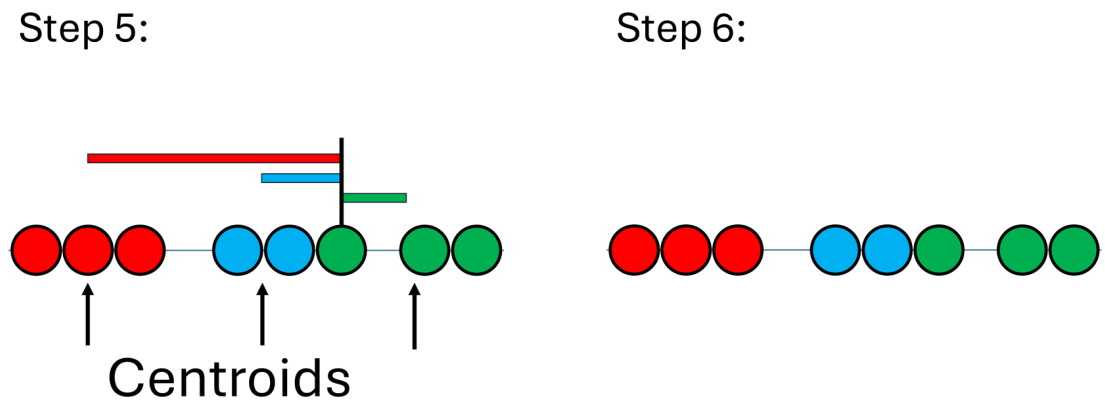



FIGURE 3.3: Stage 5 describes the position of the new centroids as the mean of all data points within its cluster and stage 6 describes the final clustering of data points.

Steps 1-6 are repeated using a different random selection of initial data points which define the first centroids. The iteration with the lowest sum of the square distances for all clusters is the ideal clustering iteration.

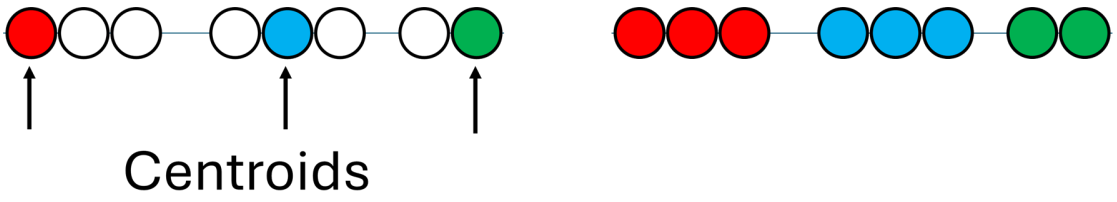



FIGURE 3.4: Initial centroids and final clustering arrangement with the smallest sum of square distances for $k$-means clustering.

The sum of square distances from this second selection is less than that of the first selection therefore is a more optimal grouping of data points.

#### 3.2.3.1 Determination of k

As the value of $k$ is an input parameter, it is useful to be able to identify its value using an automated or systematic method. In this thesis, the ideal value of clusters has been determined using an elbow plot.

Lets propose we have a dataset in which we wish to label into clusters shown in Figure 3.5

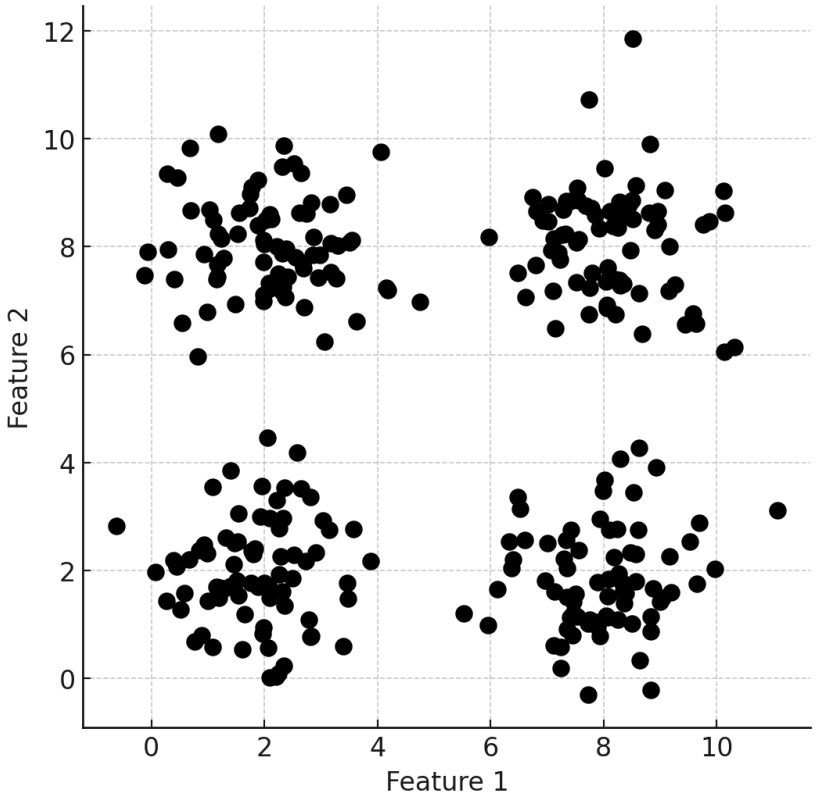
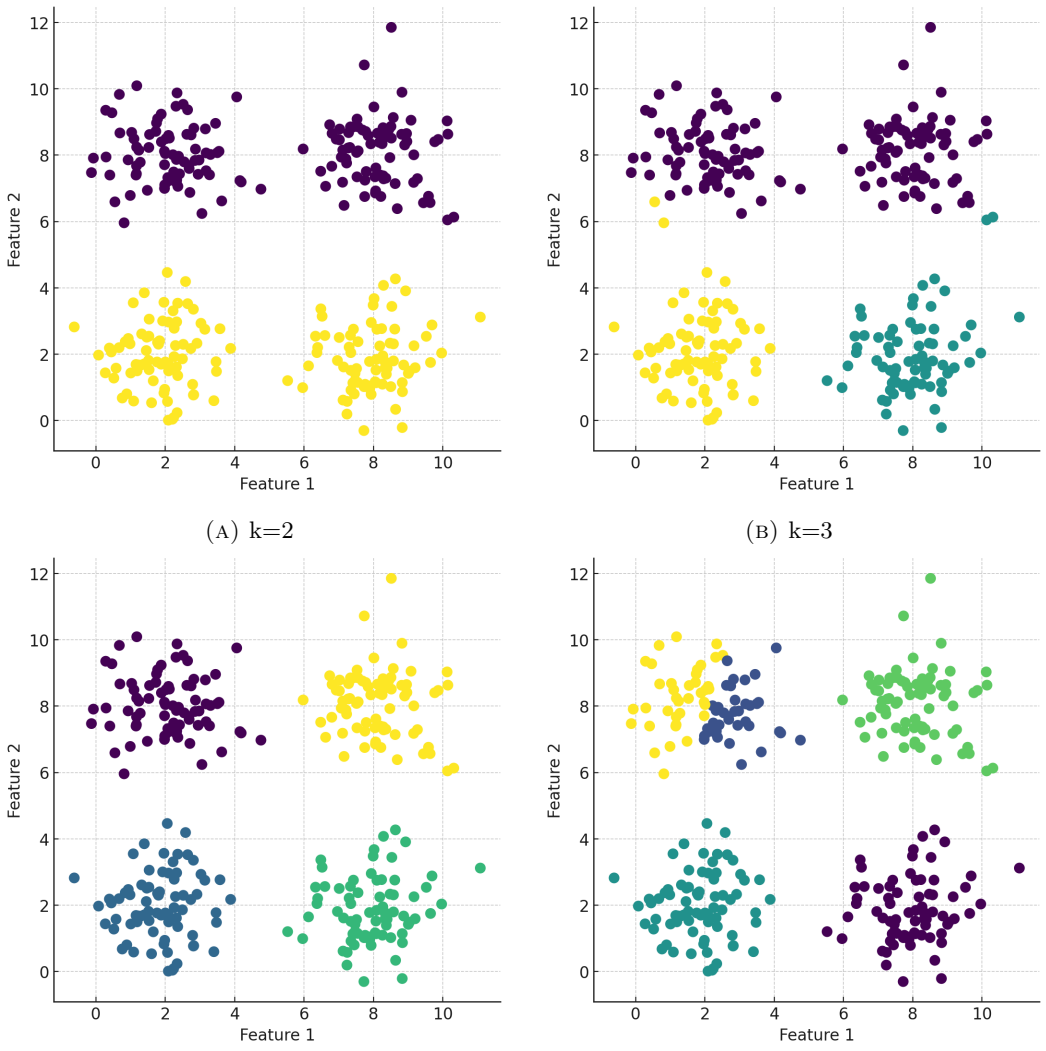



FIGURE 3.5: Example dataset for clustering using k-means clustering. Each point should be defined to a distinct cluster.

Here, we have performed k-means clustering on the dataset using a series of values of k.




(A) k=2 (B) k=3

(C) k=4 (D) k=5

FIGURE 3.6: Example assignment of data points to clusters using k-means clustering with various values of $k$. Multiple values of $k$ have been tested, and most result in poor clustering, whereas $k = 4$ produces well-defined clusters.

The sum of squared distances for each value of $k$ has been plotted and is presented in Figure 3.7.

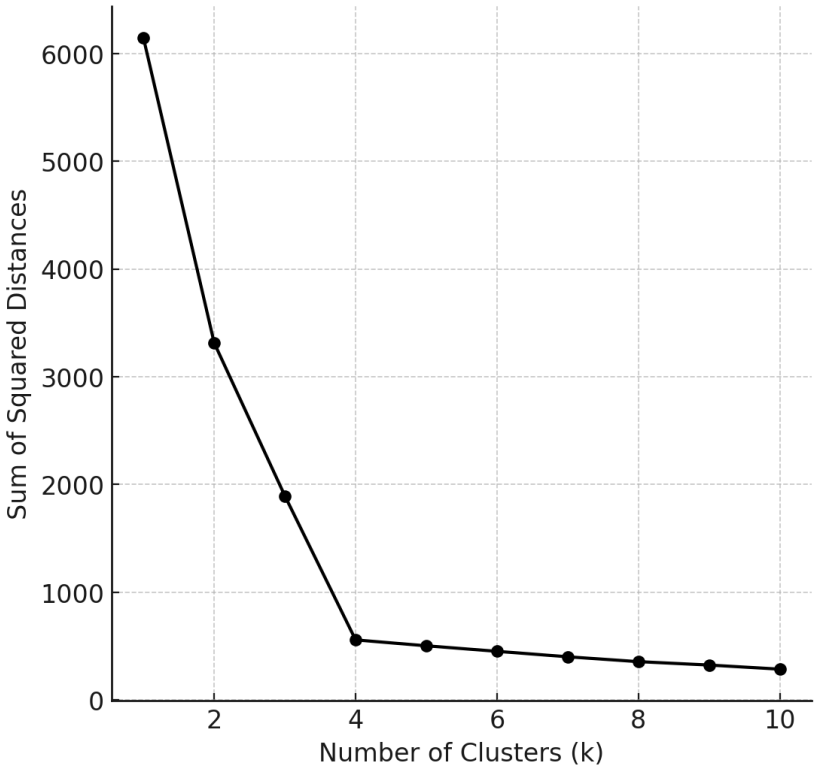



FIGURE 3.7: Example elbow plot illustrating the determination of an appropriate number of clusters for a dataset. Beyond $k = 4$, the reduction in the sum of squared distances diminishes significantly compared to earlier decreases, indicating that four clusters may adequately represent the data structure.

The sum of the square distances for all clusters will always decrease by increasing the number of clusters. Though, after the ideal value of $k$, the reduction in the sum of square distances is significantly decreases. The optimal value of $k$ is the number in which possesses the smallest value whilst also having the smallest sum of square distances. Therefore this indicates that the optimal value of $k = 4$.

### 3.2.4 Silhouette Scores

To identify how well any data point is assigned to a cluster, we can measure how close it is to its cluster rather than another cluster.

If the data point is equidistant between clusters, this suggests that it is poorly clustered. Conversely, if the data point is close to a centroid, and far away from another, then the data point has been well assigned. The metric for how well a point is assigned to a cluster is known as the silhouette coefficient which can be calculated by:

$$s_{in} = \frac{d(x_{in}, c_i) - d(x_{in}, c_j)}{max(d(x_{in}, c_i), d(x_{in}, c_j))}\,, \tag{3.43}$$

where $s_{in}$ is the silhouette coefficient, $x_{in}$ is the $n^{th}$ data point of cluster $i$, $c_i$ is the centroid of cluster $i$ and $c_j$ is the centroid of cluster $j$ that is closest to $x_{in}$ that is not $i$.

The coefficient identifies the degree of uncertainty in the assignment of any data point. Where the silhouette coefficient for a data point is close to 1, it suggests that the data point is clustered well, and far away from other clusters. Where a coefficient is 0 the data point is on the boundary between two clusters. Where the coefficient is -1 it is closer to a different cluster.

A silhouette score, $S$ is determined as the mean silhouette coefficient across all data points such that:

$$S = \frac{1}{N_{\text{total}}} \sum_{i=1}^{k} \sum_{n=1}^{N_i} s_{in} \,, \tag{3.44}$$

where $N_{\text{total}}$ is the total number of data points across all clusters.

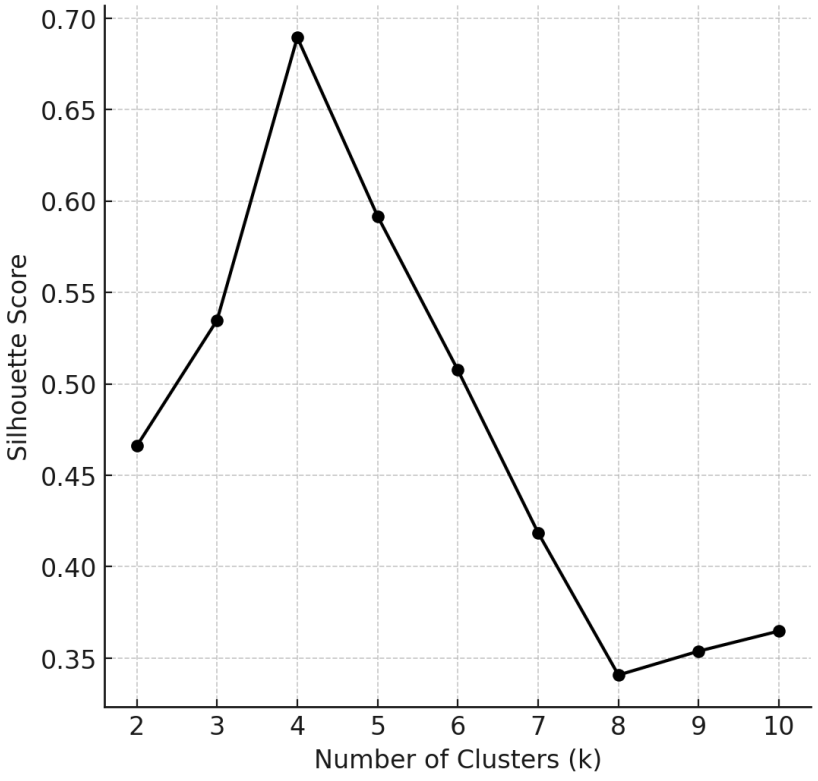



FIGURE 3.8: Example variation of silhouette scores across different values of $k$ in k-means clustering. A maximum silhouette score occurs at $k = 4$, suggesting this value optimises the separation of data points from neighbouring clusters to which they do not belong.

Plotting the silhouette scores for varying numbers of clusters enables identification of an ideal cluster count. In this instance, a maximum silhouette score is observed when $k = 4$, indicating the ideal number of clusters.

## 3.3 Conformer Search Methods

In principle, it is possible to create a full PES of a molecule by calculating the potential energy in all possible configurations by utilising a grid based search. The resolution of such a search would depend on the number of conformations generated. For systems with few degrees of freedom, this approach is feasible such as our scan of the flexible torsions $\Theta$ and $\phi$ in Figure 2.6. However this method scales poorly when working with molecules with many degrees of freedom as the complexity scales exponentially. As a result, alternative methods to explore the molecular PES must be used.

### 3.3.1 RDKit

RDKit can be used to sample conformational space by utilising distance geometry followed molecular optimisation which can produce different molecular conformations used to seed our CSP methodologies [68].

### 3.3.2 Distance Geometry

Distance geometry is used to produce a series of 3-dimensional coordinates. For this to be achieved, an initial distance bounds matrix is created consisting of upper and lower bounds for the distances between pairs of atoms within a molecule in. In this process, fixed chemical bond lengths are used, including those between directly bonded atoms, as defined by established covalent bond parameters. Bond angles are similarly constrained by the molecular geometry. These values are derived from empirical data, reflecting typical bond lengths and angles observed in structurally similar molecules. Torsion angles, which influence the spatial arrangement of atoms separated by multiple bonds, are also constrained but allow some flexibility. These are also guided by empirical data to reflect energetically favourable conformations. For atoms separated by greater distances, where direct bonding is not involved, distance bounds are estimated using empirical rules such as van der Waals radii or data from experimental techniques like NMR spectroscopy.

For example, if we wish to calculate the non-hydrogen distance bounds matrix for molecular butane, $C_4H_{10}$, we can use C-C bond distances.

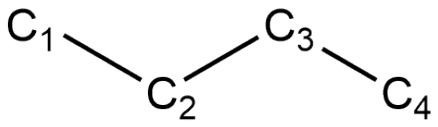



FIGURE 3.9: Molecular structure of butane with hydrogens removed for distance geometry

The distance bounds matrix $D$ where $D_{ij} = [l_{ij}, u_{ij}]$ where $l_{ij}$ and $u_{ij}$ are the lower and upper bounds between atoms $i$ and $j$ would be:

| $D_{ij}$ / Å | $C_1$ | $C_2$ | $C_3$ | $C_4$ |
|---|---|---|---|---|
| $C_1$ | 0 | [1.45, 1.60] | [2.40, 2.60] | [3.60, 4.20] |
| $C_2$ | [1.45, 1.60] | 0 | [1.45, 1.60] | [2.40, 2.60] |
| $C_3$ | [2.40, 2.60] | [1.45, 1.60] | 0 | [1.45, 1.60] |
| $C_4$ | [3.60, 4.20] | [2.40, 2.60] | [1.45, 1.60] | 0 |

TABLE 3.1: Lower and upper interatomic distances between carbon atoms in a molecule of butane for distance geometry.

Once the bounds are set, RDKit generates a distance matrix which satisfies the distance bounds matrix. For rigid molecules, the distance bounds matrix may be the distance matrix.

$$D = \begin{bmatrix} 0 & 1.54 & 2.56 & 3.93 \\ 1.54 & 0 & 1.50 & 2.60 \\ 2.56 & 1.50 & 0 & 1.51 \\ 3.93 & 2.60 & 1.51 & 0 \end{bmatrix} . \tag{3.45}$$

Multi Dimensional Scaling (MDS) is then applied to transform the distance matrix into 3D coordinates. The goal of MDS in this context is to determine a set of 3D coordinates such that the pairwise distances between atoms match the entries in the distance matrix as closely as possible.

To proceed, the matrix is double-centred, which transforms the distances into inner products.

$$D^2 = \begin{bmatrix} 0 & 2.3716 & 6.5536 & 15.4449 \\ 2.3716 & 0 & 2.2500 & 6.7600 \\ 6.5536 & 2.2500 & 0 & 2.2801 \\ 15.4449 & 6.7600 & 2.2801 & 0 \end{bmatrix} , \tag{3.46}$$

where $D^2$ is the element-wise squared distance matrix. This allows for the calculation of coordinates. The centring process involves using a centring matrix, $H$:

$$H = I - \frac{1}{n}\mathbf{1}\mathbf{1}^T = \begin{bmatrix} 0.75 & -0.25 & -0.25 & -0.25 \\ -0.25 & 0.75 & -0.25 & -0.25 \\ -0.25 & -0.25 & 0.75 & -0.25 \\ -0.25 & -0.25 & -0.25 & 0.75 \end{bmatrix}. \tag{3.47}$$

The double-centred matrix, $B$, [69] can be calculated using formula,

$$B = -\frac{1}{2}HD^2H = \begin{bmatrix} 3.8638 & 1.0544 & -1.0744 & -3.8443 \\ 1.0544 & 0.6166 & -0.5456 & -1.1254 \\ -1.0744 & -0.5456 & 0.5422 & 1.0773 \\ -3.8443 & -1.1254 & 1.0773 & 3.8925 \end{bmatrix}. \tag{3.48}$$

Elements of matrix B have been rounded to 4 decimal places. The double-centred matrix then undergoes eigenvalue decomposition via

$$B = V\Lambda V^T\,, \tag{3.49}$$

where $V$ is an orthogonal matrix of eigenvectors and $\Lambda$ is the diagonal matrix of eigenvalues, which are arranged in descending order.

This is achieved by solving:

$$\det\Big(B - \lambda I\Big) = 0\,. \tag{3.50}$$

This yields a degree-4 polynomial in $\lambda$. Numerical computation gives approximate eigenvalues:

$$\begin{aligned} \lambda_1 &\approx 8.0598 \\ \lambda_2 &\approx 0.7708 \\ \lambda_3 &\approx 0.0848 \\ \lambda_4 &\approx -0.0002 \end{aligned} \tag{3.51}$$

For each eigenvalue $\lambda_i$ is solved:

$$\Big(B - \lambda_i I\Big)\, v_i = 0\,. \tag{3.52}$$

and then each eigenvector is normalised:

$$v_i^\top v_i = 1. \tag{3.53}$$

The eigenvectors are collected into the orthogonal matrix $V$:

$$V = \begin{bmatrix} -0.5979 & -0.4722 & 0.2950 & -0.5673 \\ -0.3168 & 0.6863 & -0.6088 & -0.2039 \\ 0.1652 & -0.3974 & -0.6940 & 0.5727 \\ 0.7498 & 0.1854 & 1.0056 & 0.5558 \end{bmatrix} \tag{3.54}$$

and the eigenvalues into the diagonal matrix:

$$\Lambda = \begin{bmatrix} 8.0598 & 0 & 0 & 0 \\ 0 & 0.7708 & 0 & 0 \\ 0 & 0 & 0.0848 & 0 \\ 0 & 0 & 0 & -0.0002 \end{bmatrix} \tag{3.55}$$

Only the largest three eigenvalues are kept for 3D conformer generation which means:

$$\Lambda_{(3)}^{1/2} = \begin{bmatrix} \sqrt{8.0598} & 0 & 0 \\ 0 & \sqrt{0.7708} & 0 \\ 0 & 0 & \sqrt{0.0848} \end{bmatrix} = \begin{bmatrix} 2.8397 & 0 & 0 \\ 0 & 0.8779 & 0 \\ 0 & 0 & 0.2912 \end{bmatrix} \tag{3.56}$$

The 3D coordinates, $X$, of the atoms are then computed from the eigenvectors and eigenvalues:

$$X = V\Lambda^{1/2}\,, \tag{3.57}$$

The corresponding eigenvectors, found in the matrix $V$, provide the directions in the 3D space for placing the atoms. For butane, this results in the matrix:

$$X \approx \begin{bmatrix} -1.698 & -0.4148 & 0.0859 \\ -0.899 & 0.6025 & -0.1774 \\ 0.469 & -0.3490 & -0.2021 \\ 2.128 & 0.1628 & 0.2928 \end{bmatrix} \tag{3.58}$$

#### 3.3.2.1 Optimisation

Generated conformations initially may not correspond to local minima configurations and therefore, small adjustments to the atomic positions are made to produce a more physically realistic conformations. To ensure conformations are chemically sound, we apply an inexpensive Merck Molecular Force Field (MMFF)94 [70–74], which optimises bond lengths, bond angles, and torsion angles while keeping the molecule within its physical constraints.

### 3.3.3 Metadynamics

Metadynamics (MD) is an sampling technique designed to explore the free energy landscape of complex systems more efficiently. It is particularly valuable for studying rare events such as chemical reactions, phase transitions, and molecular conformational changes that occur on timescales inaccessible to conventional molecular dynamics methods [75, 76].

The method works by introducing a history-dependent biasing potential as the landscape is explored. This bias discourages the system from remaining trapped in metastable states, enabling it to escape and traverse less accessible regions of configuration space. As a result, the simulation can uncover new conformational states or reaction pathways that are rarely visited in standard MD.

MD also uses Collective Variables (CVs) which represent essential degrees of freedom in the system [77]. By doing so, the method simplifies the complex, high-dimensional energy landscape into a more tractable form [78]. This reduction preserves the essential physics while making the exploration of relevant thermodynamic and kinetic features computationally feasible. These CVs often correspond to slow, relevant motions such as interatomic distances or torsion angles that govern the system's transitions between metastable states [79].

During a simulation, a bias potential is continuously added to the system [80]. The bias potential is typically implemented as a series of Gaussian hills, which are added to the space of the CVs as the simulation progresses. Each time the system visits a particular

point on the free energy surface, a small Gaussian potential is deposited, raising the energy at that point and facilitating the escape from local minima [76].

The Gaussian hills described by a Gaussian function:

$$V(s,t) = \sum_i w_i \exp\left(-\frac{(s-s_i)^2}{2\sigma^2}\right), \tag{3.59}$$

where $s$ is the CV, $s_i$ is the position where the hill is added, $w_i$ represents the height (or weight) of the hill, and $\sigma$ is the width of the Gaussian. These Gaussian hills gradually fill the energy wells in the landscape [79] and enable the overcoming energy barriers to uncover new states or molecular conformations [75].

### 3.3.4 Conformer Rotamer Ensemble Sampling Tool (CREST)

CREST is a tool that provides quick and effective sampling of conformational space for a molecule which is provided by the xTB program described in section 3.1 [51, 81].

CREST offers several methodologies. However, here we will discuss the latest developments namely iMTD-GC and the iMTD-sMTD workflows.

#### iMTD-GC

The iMTD-GC workflow uses xTB with Root Mean Square Deviation (RMSD) based MD (section 3.3.3) to sample conformational space.

The CVs are given as the RMSD between previous minima on the PES during MD run with the biasing potential applied described below (Equation 3.60).

$$V_{\text{bias}} = \sum_i^n k_i e^{-\alpha {\Delta_i}^2}, \tag{3.60}$$

where $\Delta_i$ is the RMSD between minima, $n$ is the number of reference structures, $k_i$ is the pushing strength and $\alpha$ is the potential shape. The potential provides guiding forces to drive the structure away from previous minima and into unexplored conformational space. The values of $\alpha$ and $k_i$ are determined for each molecule by using a variety of biasing potentials and tested using an iterative process.

In addition, the algorithm uses genetic z-matrix crossing for more efficient sampling. This method refers to selecting two or more conformers based on their energy profiles or structural diversity and converted them into Z-matrix representations; a geometry of a molecule using internal coordinates. Portions of their internal coordinates are then

exchanged to generate a new child conformer, emulating genetic crossover such as in biological evolution.

For example, if two conformers were to undergo genetic Z-matrix crossing, a coordinate such as the H–C–C–H torsion angle might be selected as the ”gene” to be swapped. CREST could extract this torsion angle from the conformer and insert it into the scaffold of the second conformer.

This new structure is then geometry-optimised and if both energetically favourable and distinct from other members of the ensemble, it is retained and added to the conformational pool. The reference structure is then updated by utilising the genetic crossing.

The ensembles of conformers and rotamers generated from this method are collected and from here onwards will be referred to as CREST conformers generated using the iMTD-GC algorithm.

#### iMTD-sMTD

The iMTD-sMTD workflow utilises multiple MD runs and rather than updating the new structures, $V_{\text{bias}}$ is used as a global term for all runs by adding previously found minima.

The algorithm runs until convergence in the energy and number of conformations in the ensemble. For each run, new bias structures are identified using PCA and k-means clustering using torsion angles.

### 3.3.5 Low Mode Conformer Search (LMCS)

LMCS explores conformational space by investigating the direction of the eigenvector for the low-frequency vibrational modes on the PES using Eigenvector following [82]. This is achieved by the determination of eigenvectors which are described as the normal modes of vibration. We assume in this case that the path between conformations follow generally low frequency modes.

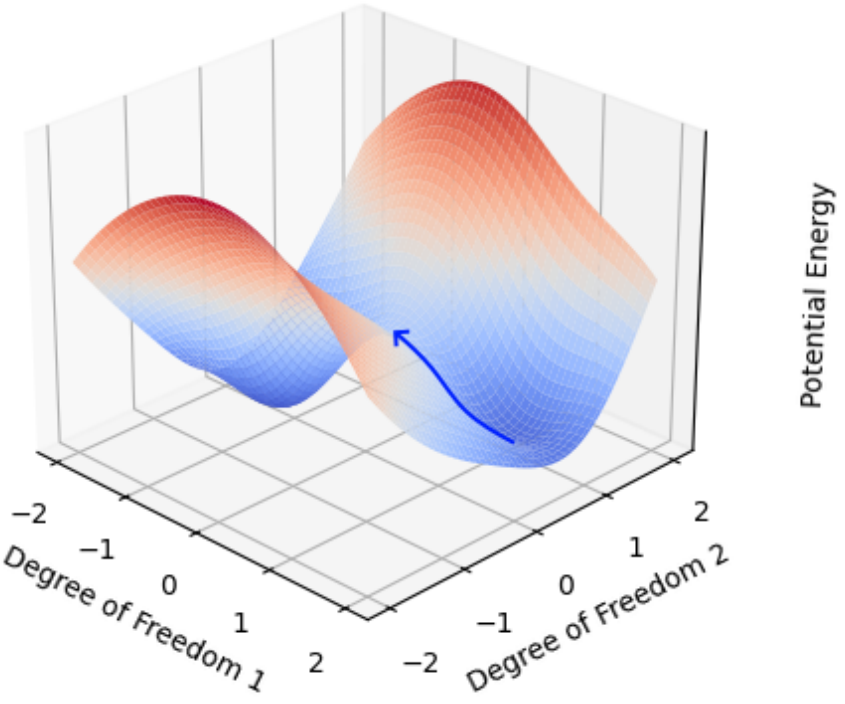



FIGURE 3.10: Movement of a trajectory on the potential energy surface using a low mode search.

To enable the method to be applicable to conformer search applications, LMCS [83, 84] determines saddle points. An initial local minimum is found by performing geometry optimisation on any molecular conformation. Saddle points are then located by making molecular perturbations of discrete step size until the change in energy becomes smaller than a defined threshold suggesting a saddle point has been reached. The resultant geometry at the saddle point is subjected to energy minimisation, identifying a new minimum [85]. If a new global minimum is found, it it is designated as the new reference point for further exploration. Subsequent perturbations and searches are then initiated from this geometry, rather than from higher-energy structures. By continuously updating the search centre to the lowest-energy conformer found so far, the method more efficiently guides the sampling process toward deeper regions of the potential energy surface, thereby improving the discovery of low-energy conformers. When the low modes are fully exhausted, MC sampling can be used to further explore the space using a random mixture of low-mode eigenvectors.

The ensembles of conformers and rotamers generated from this method are collected and from here onwards will be referred to as LMCS conformers.

## 3.4 Comparison Methods

This section describes methods used for analysis of molecules and crystal features and discusses how we can compare molecules and crystals to one another.

### 3.4.1 Shrake-Rupley Surface Area Calculations

The surface area of a molecule, particularly its Solvent Accessible Surface Area (SASA), is calculated using the Shrake-Rupley method [86]. This method estimates the SASA by placing a series of points equidistant from each atom centre and determining which of these points are accessible to a solvent probe. Points that are accessible indicate the solvent-exposed surface area, while those that are buried or occluded are ignored.

In this method, each atom in the molecule is treated as a sphere with its radius defined by the van der Waals radius of the atom. To account for solvent accessibility, an additional solvent radius is added to each atom's van der Waals radius. This results in an effective radius that reflects both the atom and the surrounding solvent. The SASA is calculated by distributing a large number of uniform points across the new surface where each point covers a patch. The areas of the exposed surfaces are calculated by the number of distributed points to provide a measure of the area.

To differentiate between exposed and buried points on the surface of each atom, for each sampled point, the algorithm checks if it is within a certain distance of another atom in the molecule. If a sampled point is close enough to a neighbouring atom (considering the van der Waals and solvent radii), it is classified as buried or inaccessible. This is done by calculating the distances between the sampled points on an atom's surface and the centres of neighbouring atoms.

Once the exposed points have been identified, the algorithm calculates the accessible surface area for each atom. Each exposed point corresponds to a small patch of the atom's surface, and the total area of these patches is summed to give the accessible surface area of the atom. The total SASA for the entire molecule is then obtained by summing the accessible surface areas of all the atoms in the molecule, accounting for overlaps between atoms.

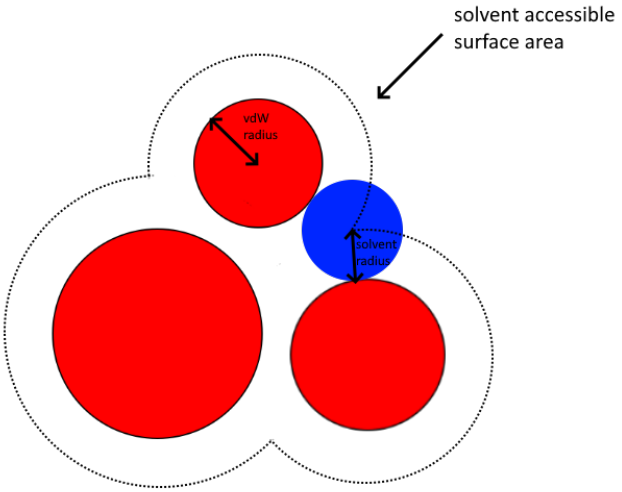



FIGURE 3.11: Shrake-Rupley Surface Area. Red circles represent atoms van der Waals surface, blue circle represents the solvent probe. The dashed line represents the solvent accessible surface area, $A_{Shrake-Rupley}$.

### 3.4.2 PLATON

PLATON is a software tool used to generate PXRD patterns for molecular crystals from the crystallographic data of a given crystal structure. It computes theoretical PXRD patterns based on the lattice parameters, atomic positions, and symmetry of the molecular crystal.

PLATON systematically generates sets of Miller indices $(hkl)$, calculates the corresponding $d$-spacings from the unit cell parameters, and determines the diffraction angles using Bragg's law.

$$n\lambda = 2d\sin(\theta), \tag{3.61}$$

where $n$ is the order of diffraction (typically $n = 1$ in PXRD), $\lambda$ is the wavelength of the incident X-ray, $d$ is the interplanar spacing, and $\theta$ is the diffraction angle.

It also calculates the structure factor $F(hkl)$ for each set of planes, which dictates the intensity of each reflection in the PXRD pattern. The structure factor is the sum of the scattering contributions from all atoms in the unit cell, considering their positions and scattering powers given by:

$$F(hkl) = \sum_j f_j e^{2\pi i(hx_j + ky_j + lz_j)} \tag{3.62}$$

where $f_j$ is the atomic scattering factor of atom $j$, and $(x_j, y_j, z_j)$ are the fractional coordinates of atom $j$. The Miller indices $h, k, l$ define the lattice plane. The intensity of each diffraction peak is proportional to the square of the structure factor:

$$I(hkl) \propto |F(hkl)|^2 \,. \tag{3.63}$$

Using the calculated structure factors, PLATON determines the intensities of the diffraction peaks. The intensity of each peak depends on the atomic positions and the scattering power of each atom in the unit cell.

This pattern produced shows the intensity of diffracted X-rays as a function of the diffraction angle, reported as $2\theta$.

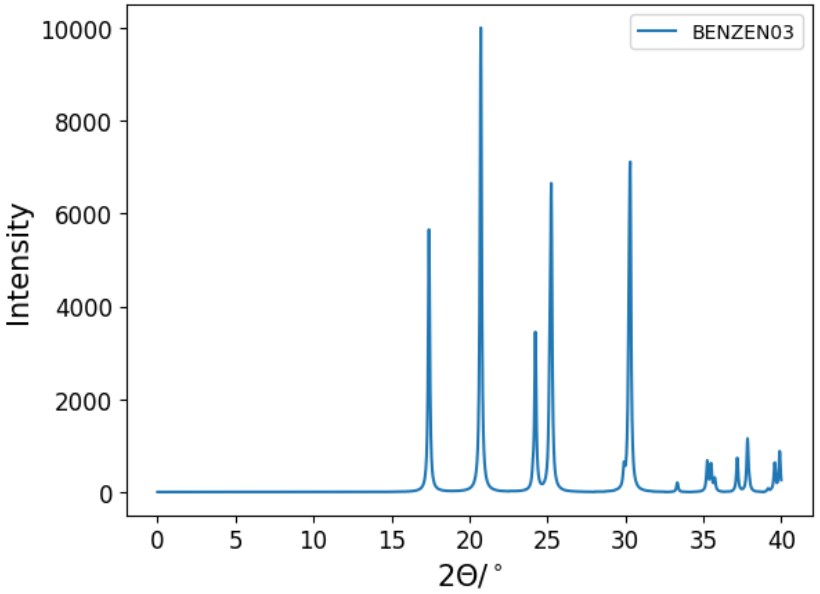



FIGURE 3.12: Simulated powder X-ray diffraction pattern of BENZEN03[87]

### 3.4.3 PXRD Comparison

Comparison of PXRD patterns can be a challenging endeavour due to the variations in the pattern shape.

#### Dynamic Time Warping

In this thesis, Dynamic Time Warping (DTW) is employed to assess the similarity between two PXRD patterns [88]. To compare two series of data; $X = (x_1, x_2, \ldots, x_n)$ and $Y = (y_1, y_2, \ldots, y_m)$, we could use the square distance between points to see how well they are aligned. For a one dimensional sequence:

$$D = \sum_{i=1}^{n} (x_i - y_i)^2 \,, \tag{3.64}$$

where $i$ is the $i^{\text{th}}$ point of series $X$, and $j$ is the $j^{\text{th}}$ point of series $Y$. However, this approach only applies if the series are of equal length, i.e. $n = m$.

A more effective method for comparing PXRD data involves allowing flexibility in the matching process through DTW. This method identifies the optimal alignment between two patterns by locally warping the positions of points and minimising their cumulative distance, even when the sequences differ in length or exhibit local shifts and distortions.

To find the distance between two points for a one dimensional sequence we use:

$$D(x_i, y_j) = (x_i - y_j)^2 \,, \tag{3.65}$$

We can calculate a cost matrix describing the pairwise distances between all points,

$$C = \begin{bmatrix} D(x_1, y_1) & D(x_1, y_2) & \cdots & D(x_1, y_m) \\ D(x_2, y_1) & D(x_2, y_2) & \cdots & D(x_2, y_m) \\ \vdots & \vdots & \ddots & \vdots \\ D(x_n, y_1) & D(x_n, y_1) & \cdots & D(x_n, y_m) \end{bmatrix} \tag{3.66}$$

We compute a cumulative cost matrix, $A$, such that:

$$A(i,j) = C(i,j) + \min \begin{cases} D(i-1, j), \\ D(i, j-1), \\ D(i-1, j-1) \end{cases} \tag{3.67}$$

We do this by first computing $A(1,1) = C(1,1)$, then calculating other elements sequentially. The value $A(n,m)$ provides us with the value of the DTW distance.

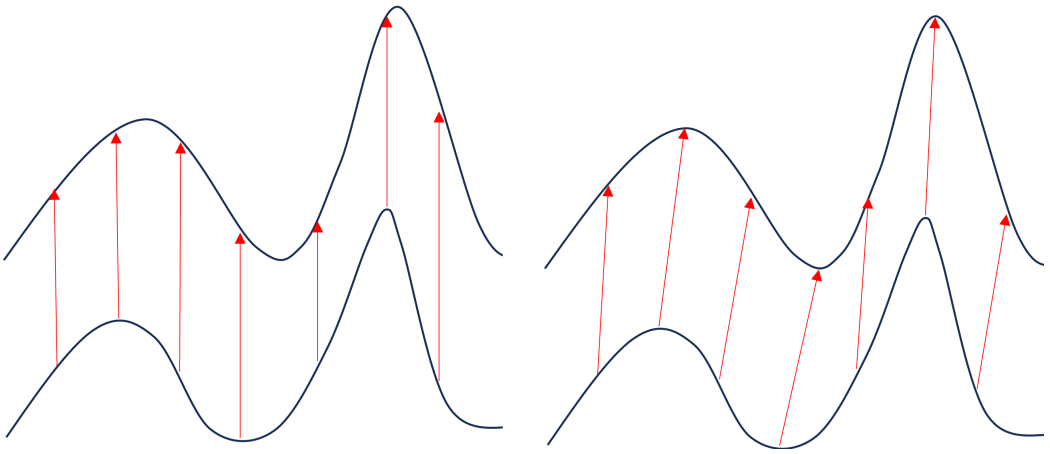
This adaptability makes the method particularly suitable for PXRD pattern analysis, where minor variations can significantly influence direct point-to-point comparisons. Although DTW is effective in identifying similarities between patterns, noise can cause peaks to be misaligned, resulting in inaccurate comparisons.

### Sakoe-Chiba Band

The Sakoe-Chiba band introduces a constraint into the DTW algorithm by restricting the warping path to a predefined band around the diagonal of the DTW cost matrix. This band is defined by a maximum allowable deviation $r$ from the diagonal, effectively constraining the search space for the optimal warping path.

In the matrix representation, the DTW algorithm uses a matrix where the cell at position $(i, j)$ represents the cost of aligning the $i^{th}$ element of $X$ with the $j^{th}$ element of $Y$. The Sakoe-Chiba band limits the path to a narrow strip around the diagonal of this matrix, meaning that if $|i - j| > r$, the corresponding cell in the DTW matrix is not considered in the path calculation. This ensures that only points within a certain range are compared, preventing excessive stretching or compression of the time series. In addition, by reducing the number of cells that need to be evaluated, the Sakoe-Chiba band can decrease the computational cost of the DTW algorithm.

This constraint ensures that the extracted distance measure remains robust and meaningful, reflecting true structural similarities rather than artifacts of experimental error. We call this distance calculated the constrained Dynamic Time Warping (cDTW) distance.



(A) Euclidean Distance (B) cDTW Distance

FIGURE 3.13: Comparison of comparing two different time series data using A) euclidean distance B) constrained dynamic time warping distance

## 3.4.4 COMPACK

COMPACK is a computational tool designed to group crystal structures after comparing their molecular packing arrangements [89]. It is primarily used to identify and remove

duplicate structures that arise during crystal structure prediction or energy minimisation, and to compare experimental and predicted crystal structures. The method emphasises the packing similarity between two crystal structures by comparing the relative positions and orientations of molecules in each structure.

The clustering process in COMPACK begins by selecting a reference molecule in each crystal structure. This reference molecule serves as the centre for comparison and a cluster of neighbouring molecules around it is considered. Typically the nearest 30 neighbour molecules are chosen to define the cluster. These neighbours are the closest molecules in the crystal lattice and their relative arrangements are compared between the two crystal structures.

#### Superposition of Molecular Clusters

COMPACK employs molecular superposition aligning clusters of neighbouring molecules in the crystal structures. A least-squares fit is performed to superimpose the reference molecule and its nearest neighbours from one structure onto the corresponding molecules in another structure [90]. During this superposition, COMPACK minimises the RMSD between the atomic positions of the two clusters.

Short intermolecular distances between atoms of adjacent molecules are focussed on, as these distances play a critical role in defining the packing motifs of the crystal. These nearest neighbour interactions contribute significantly to the stability and arrangement of the crystal [91]. By comparing these distances, COMPACK determines whether the molecular packing in two crystal structures is similar, even if there are slight differences in molecular conformations.

COMPACK calculates the RMSD between the molecular clusters after superposition. The total RMSD value represents the degree of similarity between the packing arrangements in the two crystal structures. A low RMSD indicates high similarity in molecular packing, whereas a high RMSD suggests significant differences in how the molecules are packed.

After calculating RMSD values for all pairs of structures, COMPACK clusters those with similar molecular packing. Structures with low RMSD values, indicating high packing similarity, are grouped together as duplicates or variants of the same crystal form. This approach is especially effective for identifying structures that share similar packing motifs but may differ slightly in molecular conformation or orientation. This process aids in removing duplicates from predicted crystal structures [92]. COMPACK is robust enough to detect similarities even when the predicted structure exhibits slight deviations in molecular conformation or positioning [93].

**Experimental Match**

A structure is classified as an experimental match if, within a superimposed molecular cluster, each pair of equivalent atoms in the cluster is no more than 20 % apart, and the angular differences between the two clusters are less than 20°. The RMSD for a cluster of 30 molecules is reported as $RMSD_{30}$.

### 3.4.5 ShiftMLv2

We utilise ShiftMLv2 which can be used to predict $^{1}H$ and $^{13}C$ NMR chemical shift values for organic molecular crystals using a machine learning approach [94]. The training data, derived from high-quality quantum chemical calculations, represents each atom in a molecule through features capturing its local chemical environment. ShiftMLv2 employs a neural network model trained to minimise the difference between predicted and reference chemical shifts.

The model uses supervised learning, with known chemical shifts serving as target outputs for each atomic environment. Parameters are optimised iteratively through back propagation and gradient descent. Techniques like cross-validation and regularisation are applied to prevent over fitting and ensure the model generalises well to new data.

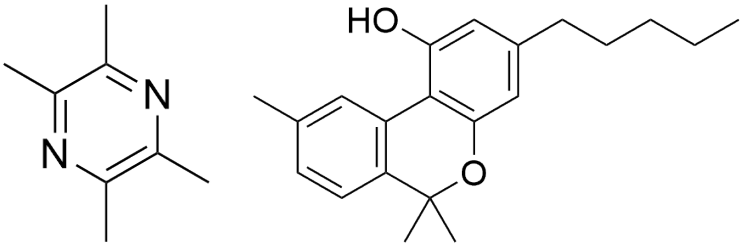
# Chapter 4

# CCDC Blind test 2021

*The following work has been in collaboration with Ramón Cuadrado, Joseph Glover, Christopher R. Taylor and Graeme M. Day. The author, Cuadrado, Glover, Taylor and Day ran CSP calculations on all sixth blind test targets and Day also provided his expert advice. In addition, the author performed conformational searches, CSPs and geometric optimisations of crystal structures which are presented here.*

The CCDC organises blind tests to assess the current state of scientific methods in the field of CSP. In these blind tests, various participants and research groups attempt to predict the crystal structures of small molecules without prior knowledge of the experimental structures.

The seventh CSP blind test, conducted from 2020 to 2022, presents unique challenges for CSP examining a range of molecules. This chapter focuses on the CSP of target XXX, depicted in Figure 4.1, and establishes the groundwork regarding the challenges and areas of development necessary for future blind CSP studies.




(A) Molecule A (B) Molecule B

FIGURE 4.1: Target XXX for the Cambridge Structural Data Centre Blind Test 2021

In this study, we have been provided that for target XXX, there are two experimentally known forms with different stoichiometries and the number of components in each

stoichiometry is $< 4$. As a result, the following stoichiometries should be investigated: 1:1, 2:1 and 1:2 of molecule A and molecule B respectively. The challenges imposed by this target is that there are two elements which both contribute considerably towards computational cost. Firstly, there are multiple stoichiometries and secondly molecule B has 5 flexible torsions. These coupled together means that there is a vast amount of configurational space to explore.

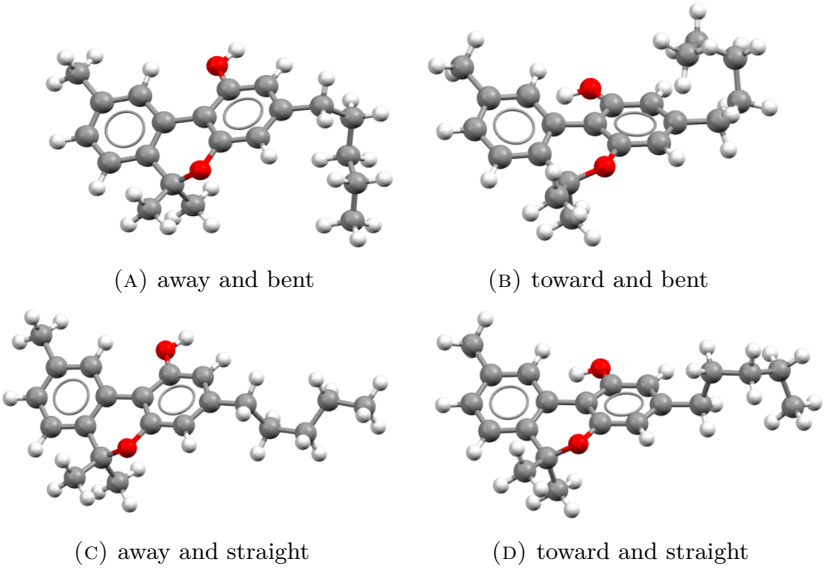
The methods employed in this study follow the workflow depicted in Figure 2.12, utilising multiple generated conformations for each of the different stoichiometries.

## 4.1 Conformer Search

The PES of molecule B was explored using conformational searches with both LMCS and CREST methods. For CREST, the iMTD-sMTD algorithm was used. The resultant conformations were optimised with DFT using PBE0/6-311G(d,p) with GD3BJ and then clustered within an RMSD threshold of 0.5 Å.

Despite using two different search methods and combining the results, conformational searches yielded few conformers suggesting the landscape was not very well explored. As a result, we attempted to improve our conformational search by utilising multiple starting conformations. We suggest that utilising different starting positions would enable better exploration of conformational space.

Conformers were manually selected which possessed distinctly different conformations as shown in Figure 4.2.



(A) away and bent (B) toward and bent

(C) away and straight (D) toward and straight

FIGURE 4.2: Starting conformations of target XXX molecule B for the CREST conformational search. Each conformation is designed to be distinct from the others to maximise exploration of the conformational space. **KEY:** Grey – carbon; white – hydrogen; red – oxygen.

Two main features were of particular interest: the position of the hydrogen atom in the hydroxyl group and the conformation of the alkyl chain.

Molecule A and the conformers of molecule B underwent geometry optimisation at the DFT level using Gaussian09 with the PBE0/6-311G(d,p) basis set and GD3BJ empirical dispersion correction. The resulting structures were then clustered using RMSD with a threshold of 0.5 Å, yielding a final set of 123 unique conformers.

Given the number of conformations and stoichiometries that needed to be explored during the CSP process, it was determined that further reduction in computational cost

was necessary. Therefore, the focus shifted towards conformations more likely to be experimentally observed, emphasising the crystal structure of neat molecule B.

CANNOL, depicted in Figure 4.3, represents the neat crystal structure of molecule B. Examination of CANNOL shows that molecule B crystallises with its alkyl chain extended [95].

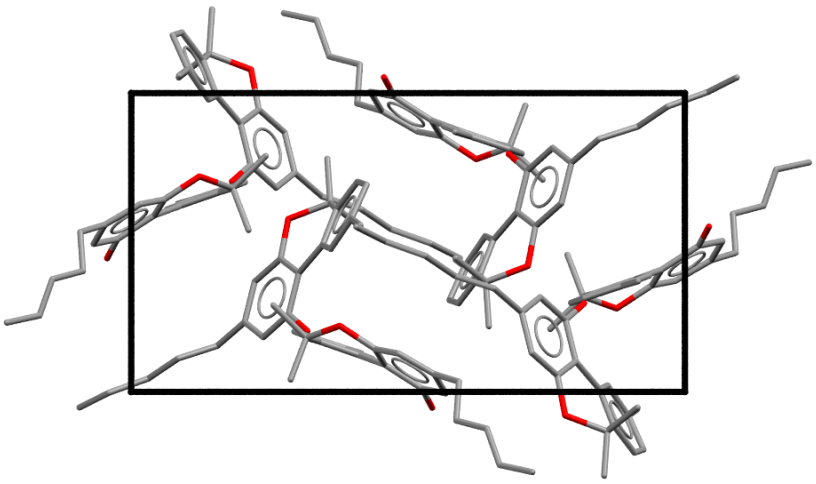


FIGURE 4.3: The crystal structure of molecule B, CANNOL from the Cambridge structural database. Here, the alkyl chains are shown to be extended.

Particular attention was directed towards studying conformers featuring extended alkyl chains. The orientation of the hydrogen atom on the hydroxyl group of molecule B could not be determined with certainty. Analysis of the conformers indicated that they did not exhibit significant variation in hydroxyl orientations. Consequently, it was considered important to perform an OH scan to examine the energies of the molecule at different angles.

For each conformer, such a scan was carried out by fixing all bonds and bond lengths except for the OH group. The results indicated that energy minima were generally found approximately 180° from the original OH orientation in the conformer. It is proposed that future CSP efforts should incorporate these OH scans to enable a more comprehensive exploration of this conformational space.

## 4.2 Crystal Structure Prediction

CSP was performed on the DFT optimised conformers and their respective OH scans in a co-crystal with the geometry optimised molecule A. Three different stoichiometries were investigated, 1:1, 2:1 and 1:2 of molecule A and B respectively, each using different amounts of sampling.

For the 1:1 case, the sampling shown in Table 4.1 was used.

| Space group | Number of valid structures |
|---|---|
| P 1 $2_1$ / c 1 | 50000 |
| C 1 2 / c 1 | 50000 |
| P 1 $2_1$ 1 | 20000 |
| P b c a | 20000 |
| P $2_1$ $2_1$ $2_1$ | 10000 |
| P -1 | 10000 |

TABLE 4.1: Number of valid crystal structures generated using the crystal landscape generator for target XXX for the 1:1 stoichiometry. The number of crystal structures correlates with how frequently each space group is observed within the crystallographic structural database.

In the 2:1 and 1:2 case, the sampling of space group P 1 $2_1$ / c 1 was doubled. This space group is among the most frequently observed symmetries for organic molecular crystals and is therefore sampled more extensively. Ideally more sampling of these structures would be conducted, however due to time constraints, we found that it was more appropriate to investigate a larger number of conformers.

## 4.3 Optimisation

In each of the CSPs, crystal structures within 10 kJ mol$^{-1}$ were optimised using DFTB+. This was chosen over VASP due to time constraints imposed by an impending deadline. The structures were then optimised using DMACRYS.

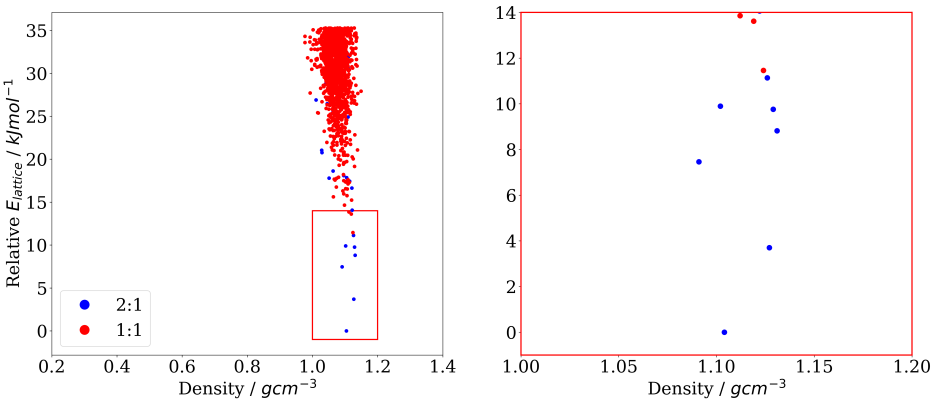



FIGURE 4.4: Crystal landscape of the lowest 1500 crystal structures generated for target XXX in the Blind Test 2021. Shown in red is the magnified low energy region of the crystal landscape corresponding to 14 kJ mol$^{-1}$ above the global minimum.

We predicted that the co-crystals formed with stoichiometries 1:2 (A:B). Both 1:1 and 1:2 were more stable than neat forms of molecule A and B. The 2:1 stoichiometry did not appear to form stable structures.

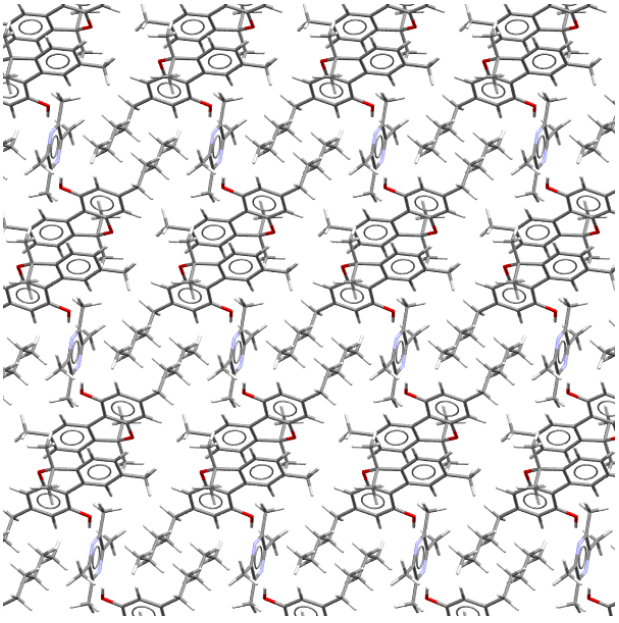


FIGURE 4.5: Lowest energy structure for target XXX of the blind test obtained from performing **KEY:** Grey – carbon; white – hydrogen; blue - nitrogen; red – oxygen.

## 4.4 Conclusions and Future Work

The complete workflow employed for the prediction of target XXX has been presented. As the experimental structure is not yet available, direct comparison for accuracy is not possible. However, a list of the lowest-ranked crystal structures has been submitted to the CCDC for review. The publication relating to this work has not yet been released. Though this study has highlighted several areas warranting further exploration.

Firstly, the identification of conformers from different starting points emerged as a novel idea that deserves additional investigation. It was observed that initiating searches from varied starting positions led to the discovery of many new conformers not previously identified as discussed in Chapter 5.

Secondly, for one of the targets, a PXRD pattern was available that could potentially have been incorporated into the calculations. However, standard methods for integrating such data into the CSP process were lacking. Including this information might improve the accuracy of future CSP predictions which is to be discussed in Chapter 6.

Thirdly, packing rigid conformers alone may not be enough to determine experimentally observable crystal structures. As performed in the case of molecule B, sampling around each conformer may enable improved structure prediction as discussed in Chapter 8.

# Chapter 5

# Conformer Search Methods

In this chapter, conformational search methods are explored in the context of their application to CSP schemes. Previous work has employed LMCS as a primary tool for sampling conformational space, owing to its superior performance compared with other search methods available at the time [96]. However, recent years have seen the emergence of new conformational search techniques that may offer improved determination of conformations when exploring molecular conformational space. Notably, a recent publication by Pracht et al. describes the use of MD to sample low-energy conformational space within the CREST program [81].

In this work, the commonly used LMCS method is benchmarked against CREST, with a focus on exploring parameters potentially suitable for CSP applications. Subsequently, the development of a new conformer search methodology is investigated. This involves examining how multiple conformer search methods can be combined to achieve a more comprehensive sampling of the PES.

## 5.1 Clustering Methods

Before conducting investigations it is essential to establish a method for comparing conformations. For any ensemble of conformations generated by a given method, it is desirable to determine whether two conformations correspond to the same local minimum on the PES. Therefore, a clear and well-defined comparison technique must be outlined to enable effective bench marking.

Even if conformations occupy the same minima, their geometries and specific coordinates might differ slightly due to numerical error. Therefore, it is necessary to employ a method to eliminate duplicate conformations. While this problem may initially seem trivial, it is crucial to ensure that the comparison process does not affect nearby minima when evaluating different molecular conformations. This issue, known as over-clustering, poses a challenge in clustering molecular geometries. This concept can also be applied to comparing two different ensembles of structures generated by distinct methodologies. Here, methods of comparing conformations are described.

### 5.1.1 Clustering Using Root Mean Square Deviation (RMSD)

RMSD-based clustering is an efficient method for grouping molecular conformations. This approach calculates distances between equivalent atoms in different structures, producing a score that reflects the similarity of the conformations. The use of the root mean square emphasises larger interatomic deviations, so even a few mismatched atoms can significantly influence the overall RMSD value.

$$\mathrm{RMSD} = \sqrt{\frac{1}{N}\sum_{i=1}^{N}\left|\mathbf{r}_i^{(A)} - \mathbf{r}_i^{(B)}\right|^2}, \tag{5.1}$$

where $\mathbf{r}_i^{(A)}$ and $\mathbf{r}_i^{(B)}$ are the position vectors of the $i^{th}$ atom in molecule or structure A and and B respectively and $N$ is the total number of atoms in one molecule or structure.

Although two conformers may be chemically identical, their calculated RMSD after structural superposition is rarely exactly zero in practice. Minor residual differences often arise due to numerical precision limits and floating-point rounding errors during coordinate alignment, resulting in subtle deviations in atomic positions. Consequently, practical RMSD thresholds are used and below which, conformers are considered practically identical.

Values below approximately 0.2 Å usually indicate virtually identical structures, differing only due to numerical noise. RMSDs between 0.2 and 0.5 Å often reflect minor differences, such as variations in hydrogen atom positions, yet may still correspond to

the same conformer. RMSDs exceeding 0.5 Å typically suggest a significant geometric or conformational difference [97]. The precise threshold employed depends on factors such as the molecule's size, flexibility, and the specific purpose of the comparison, whether for clustering in crystal structure prediction or distinguishing unique conformers [98]. These are shown in Figure 5.1.

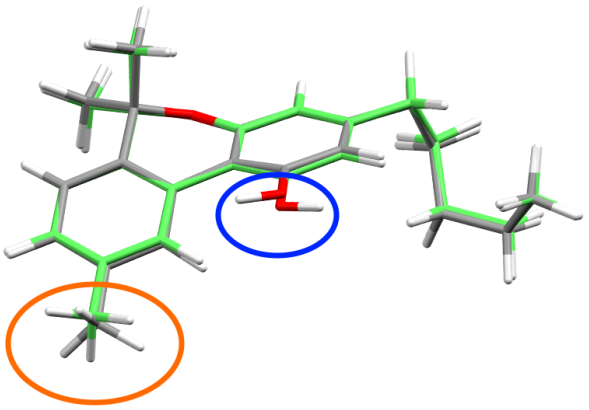


FIGURE 5.1: Overlay of two molecules yielding an RMSD of 0.402 Å. The differing orientation of the O–H functional group is highlighted in blue, while the mismatched methyl group is shown in orange. Although the methyl group's orientation is largely inconsequential, the O–H group can engage in directional hydrogen bonding, significantly influencing the intermolecular interactions within a crystal lattice. **KEY:** Grey – carbon; white – hydrogen; red – oxygen; green – carbons belonging to the second molecule.

Using RMSD alone, it may be difficult to resolve these differences, which may cause distinct conformations to be grouped together and potentially omitted from consideration when exploring crystal landscapes.

The threshold for clustering conformations could be lowered further to mitigate these issues; however, doing so may result in the inclusion of more geometrically and energetically similar conformations, thereby increasing the computational cost of the CSP calculations unnecessarily. Adding a check to see if the energy between conformers is small is better, but might be insufficient as two conformer may hold similar energies whilst being geometrically distinct when it comes to directional functional groups. Therefore an alternate method of ensuring conformers are clustered, whilst not removing distinct conformations which hold directional functional groups, is necessary.

### 5.1.2 Torsional Clustering

To tackle this problem, code has been written to cluster conformations based on torsion angles. This approach compares equivalent torsion angles between conformations and allows for resolving mismatched atoms such as in the example above. In addition, a Root Mean Square (RMS) torsion angle can be used which can identify rotamers allowing for higher resolution when clustering.

This code has been written to allow the user to cluster a series of conformations in the xyz format; a typical format for molecule structures in computational chemistry. The method removes duplicate molecular conformations using a list of user defined torsions and for two conformations to be considered identical, two conditions must be met:

- The difference in the angle between any pair of equivalent torsions is within a specified threshold.
- The root mean square difference of all pairs of torsion angles are lower than a specified threshold.

The rationale for these constraints is to detect large torsional differences in molecular conformations, whilst allowing for some molecular flexibility in the clustering. In principle it could be sufficient to cluster simply based on angular differences between any pair of equivalent torsions. However, this may over-cluster conformations particularly those in which multiple pairs of torsions are close to the angular limit. The use of a RMS difference adds another criterion to be met in order for two conformers to be unique. This reduces potential over-clustering as if a pair of conformers had many torsion angles that were close to the torsion angle difference limit, they would be able to be treated as unique.

Flexible torsions are automatically detected for any conformation; however, it is also possible to manually specify which torsions to use for clustering. This allows the user to deliberately exclude certain conformational changes that are not significant for their investigation.

To ensure geometric accuracy when calculating torsional angles, these were defined as directional (clockwise or anticlockwise) relative to the torsion under study. This approach prevents molecular conformations from being incorrectly classified as identical when, in fact, their torsions are geometrically distinct despite having the same numerical values.

To address challenges associated with molecules exhibiting symmetry, a method was required to identify equivalent torsional environments. For example, in the rotation of a C–N bond in a $NO_2$ group, multiple torsion angles may be geometrically equivalent. Therefore, the rotational symmetry for such torsions was determined by identifying the

maximum value of $n$ for which a rotation of the torsion angle by $360/n$ results in an RMSD between two rotamers of less than 0.1 Å. When comparing conformers, differences were then assessed while accounting for this symmetry.

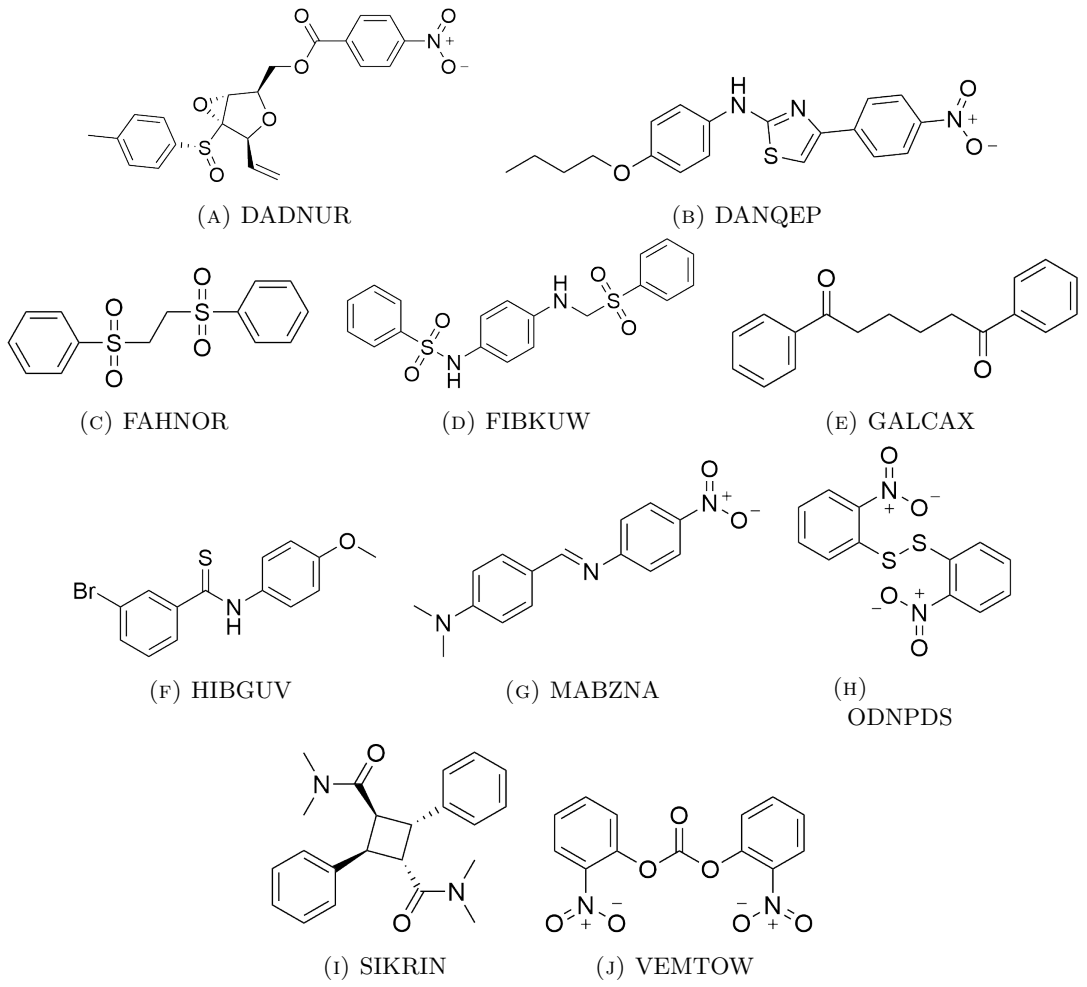
To ensure efficient storage and account for High Performance Computing (HPC) efficiency, conformations were stored inside sqlite3 databases [99]. This means that conformers can be easily clustered and fed into other applications without difficulty.

The code benefits from reduced computational cost compared to utilising an RMSD between conformations. This is due to the torsion angles being defined on a molecular basis. Atom labelling within the xyz file format is used to identify atoms which make up each flexible torsion. The molecular graph between two conformations does not need to be calculated unlike RMSD methods which allow for a significant speed up when making conformational comparisons.

## 5.2 Conformer Searches

### 5.2.1 Test Molecules

The molecules shown in Figure 5.2 were studied by performing a conformational search using LMCS and CREST methods. These molecules were chosen as they possessed similar features to drug molecules which would be under future investigation which include flexibility, functional groups, atom types and molecular sizes.



(A) DADNUR (B) DANQEP (C) FAHNOR (D) FIBKUW (E) GALCAX (F) HIBGUV (G) MABZNA (H) ODNPDS (I) SIKRIN (J) VEMTOW

FIGURE 5.2: Molecular diagrams of the molecular unit of crystals according to REFCODE to be used to test the Low Mode Conformer Search (LMCS) and the Conformer Rotamer Ensemble Sampling Tool (CREST) using iMTD-GC and iMTD-sMTD. Molecules are referred to as the name of their REFCODE present within the Cambridge Structural Database for more compact labelling of molecular structures.

To address the state of conformational search methods currently available, a set of conformers obtained from published works by Thompson 2014 [24] were used. In this work, the LMCS method [100] was applied to each of the molecules from Figure 5.2 where the resultant conformations exist as minima with MMFF PES [70].

### 5.2.2 Single CREST Search

A single CREST search was performed for each molecule in Figure 5.2. To do this, molecular conformations were extracted from each crystal with the corresponding REFCODE as starting positions. Where there were multiple conformational geometries, the lowest energy conformation was chosen. We performed both the iMTD-GC and iMTD-sMTD procedure on molecules. The PES was explored up to 25 kJ $mol^{-1}$ above the global minimum. Conformer search details can be found in Appendix A.

### 5.2.3 Comparison of Methods

Each ensemble was clustered according to similarity in torsion angles using our torsional clustering procedure in subsection 5.1.2. Conformers were deemed identical if the maximum difference in torsion angles was $< 10^{\circ}$ and the RMS difference $< 5^{\circ}$. These angles seemed appropriate for distinguishing conformers.

The two sets of conformers produced by CREST using the iMTD-GC and iMTD-sMTD methods were compared by analysing their torsion angles to identify conformations present in both sets. As DFT-optimised conformers are typically used as inputs for CSP methods, geometry optimisation was performed using Gaussian with the 6-311G(p,d)/B3LYP level of theory incorporating a GD3BJ dispersion correction. This approach meant that any minima unique to the GFN2 landscape were removed, leaving only those present on the DFT PES. The results are shown in Figure 5.3.

For conformations generated using the LMCS method, geometry optimisation was also performed using the same settings as previously mentioned.

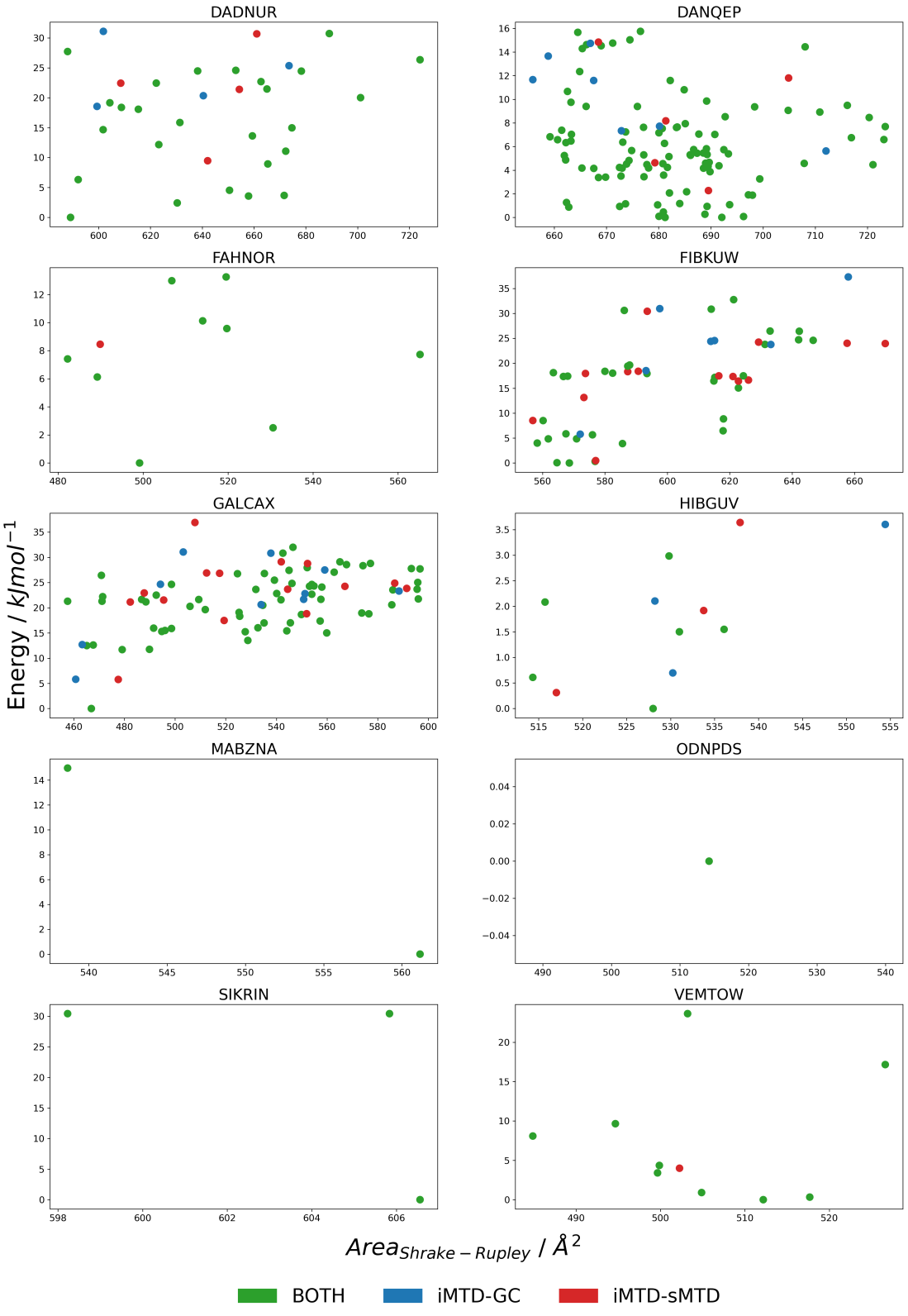



FIGURE 5.3: Relative distribution of conformational energies produced by the iMTD-sMTD and iMTD-GC algorithms in conformer rotamer ensemble sampling tool (CREST) methods following density functional theory optimisation. Areas are calculated using the Shrake-Rupley method. Many conformers were identified by both algorithms, although some conformers were missed by each method.

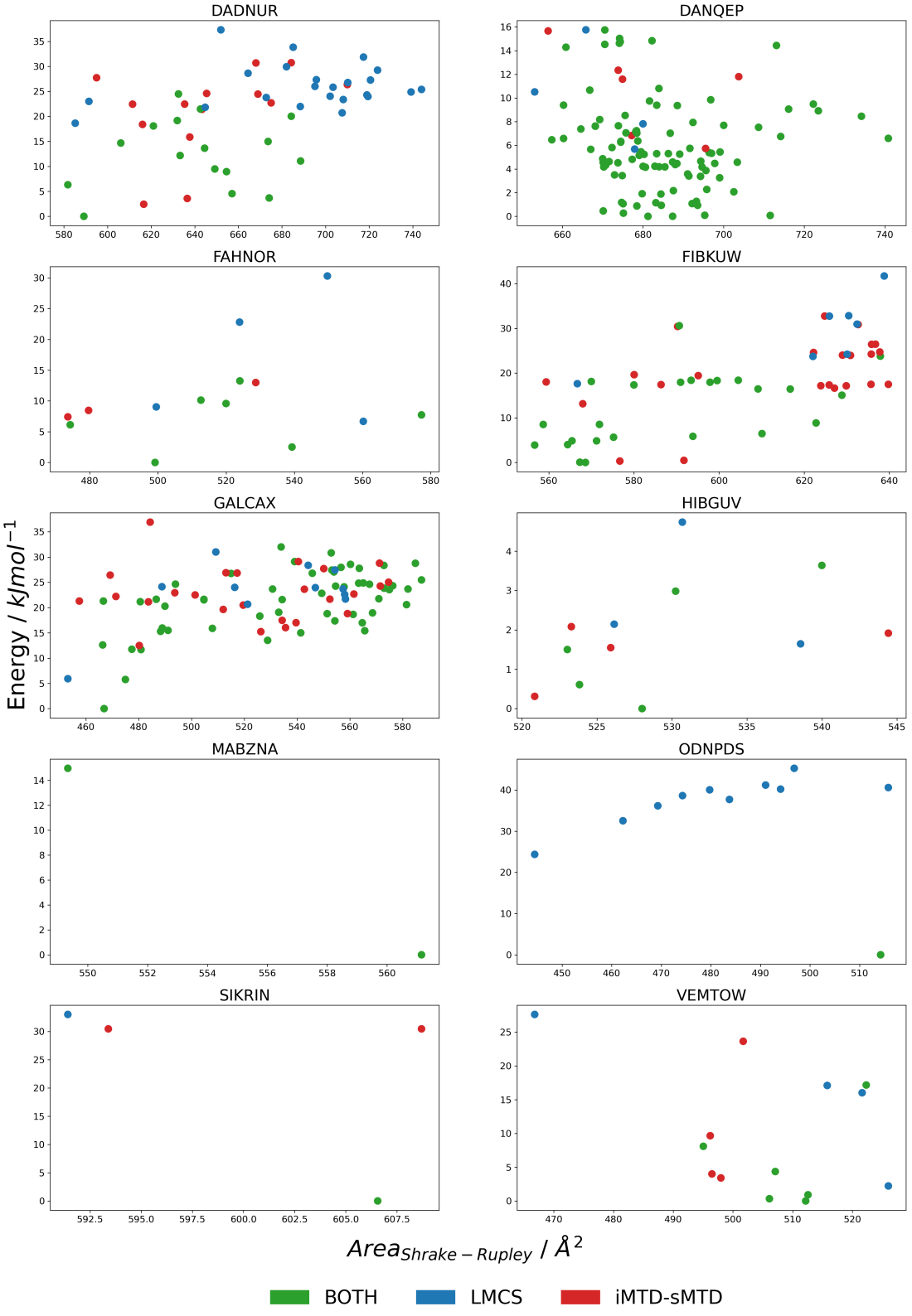



FIGURE 5.4: Relative distribution of conformational energies produced by the Low Mode Conformer Search (LMCS) algorithm and iMTD-sMTD algorithm in Conformer Rotamer Ensemble Sampling Tool (CREST) methods following density functional theory optimisation. Areas are calculated using the Shrake-Rupley method. Significant numbers of low energy conformations were missed for the LMCS.

In most subplots, the green points, representing shared results, tend to cluster in low-energy regions, indicating that both methods generally identify similar favourable conformations. The iMTD-sMTD algorithm often finds lower energy conformations compared to the iMTD-GC method (e.g., in the subplots for DADNUR, DANQEP, and FIBKUW), suggesting that iMTD-sMTD is slightly more effective at locating the lowest energy conformers. Both methods are able to find high energy conformations which the other method does not.

In Figure 5.4, LMCS shows a broader range of energy values, occasionally reaching higher levels (e.g., in DADNUR, DANQEP, and FIBKUW), while iMTD-sMTD sometimes finds lower energy conformations but less consistently. This observation is likely due to the sampling in the CREST methodology being only up to 25.1 kJ $mol^{-1}$ (6 kcal). Most structures not found by LMCS appeared to be much higher in conformational energy. There are many shared conformations in lower energy regions for both methods, indicating significant overlap. Neither LMCS nor iMTD-sMTD consistently outperforms the other. LMCS has broader coverage but does not reliably find the lowest energies, whereas iMTD-sMTD is more effective in specific cases.

## 5.3 Improving Upon a CREST Search

### 5.3.1 Utilising varying starting positions

It has been shown that there is discrepancy between the LMCS and CREST search methods. Neither method is clearly superior but rather produce different ensembles of conformations for a single molecule. This may be due to the algorithm or the energy model used in PES exploration.

It is possible that a single CREST search does not produce a complete ensemble of conformers despite the energy lid for calculations being sufficient to explore the landscape due to the energy barriers between conformers. This results in certain conformations being inaccessible from an initial geometry as conformations would need to surpass the energy limit in order to explore it further.

One solution is to increase the energy lid; however, a challenge arises in that the height of the required energy barrier is unknown. Although it is possible to raise this barrier substantially, doing so could impose significant computational costs. This issue is particularly pronounced for large molecules, as the number of degrees of freedom increases this cost exponentially.

If such an energy barrier exists, it is proposed that CREST searches be performed using geometries located on different sides of the barrier, with each search conducted independently. However, this introduces a new challenge in determining the appropriate starting positions for a given molecule. Ideally, this selection process should be automated, requiring minimal user input.

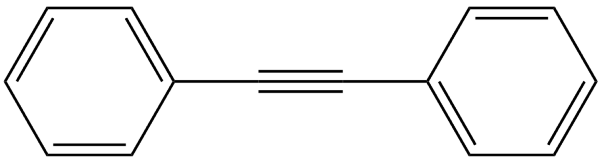


FIGURE 5.5: Diphenylethyne

Diphenylethyne shown in Figure 5.5 adopts a low energy state when both aromatic rings are in the same plane due to aromatic effects as shown in Figure 5.6. Single point energy calculations in both conformations (6-311G**/PBE0 with GD3BJ) reveal an energy barrier of 3.57 $kJmol^{-1}$, where the orthogonal conformation was lowest in energy.

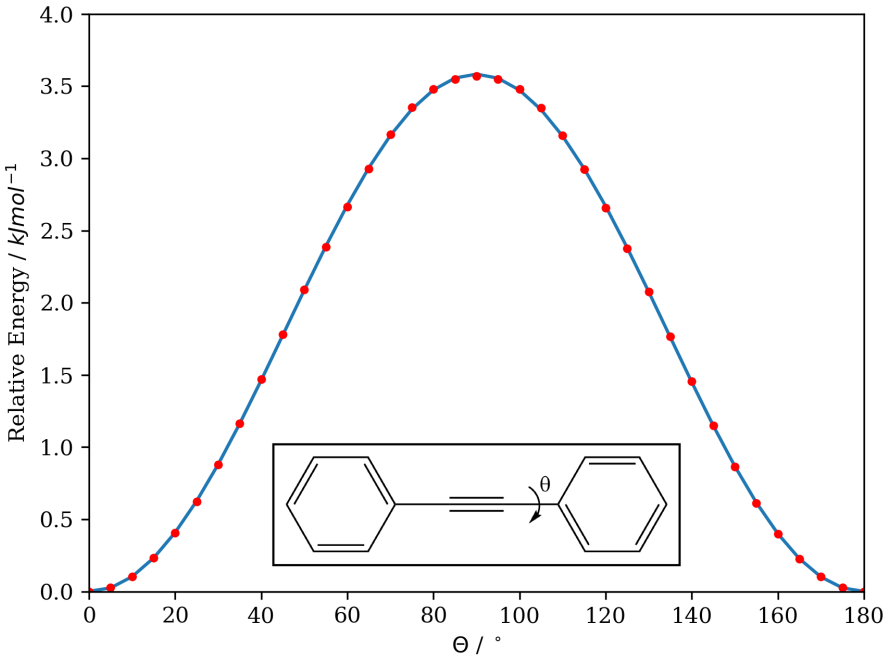



FIGURE 5.6: Rotational energy barrier of the phenyl group in diphenylethyne, calculated by rotating around the C–C bond using density functional theory with the 6-311G**/PBE0 basis set and GD3BJ dispersion correction. At 0°, the phenyl rings are parallel to each other, while at 180° an energy maximum is observed.

### 5.3.2 Determining CREST Starting Positions

Presented here is the workflow termed mCREST, introduced as an improved method for generating conformers.

To determine an appropriate starting position, it is necessary to consider the factors that contribute to large energy barriers. A relatively cheap method of generating many conformations is through distance geometry. Here, for each molecule, 10,000 molecular conformers were generated using RDKit described in section 3.3.1.

To analyse the conformational PES, torsion angles were measured from the resulting conformers. Initially, PCA was performed on all molecular torsions. However, due to the circular nature of torsion angles, meaningful data was not able to be obtained and a method capable of accounting for their periodic characteristics was required.

Geodesic PCA [101] is performed, and the resulting principal components are clustered to identify conformations that are geometrically similar to one another. The number of principal components used is limited to ensure that at least 90% of the total variance is captured. This approach reduces the dimensionality of the data by filtering out information from non-flexible torsion angles. For illustration, the first two principal components are shown in Figure 5.7.

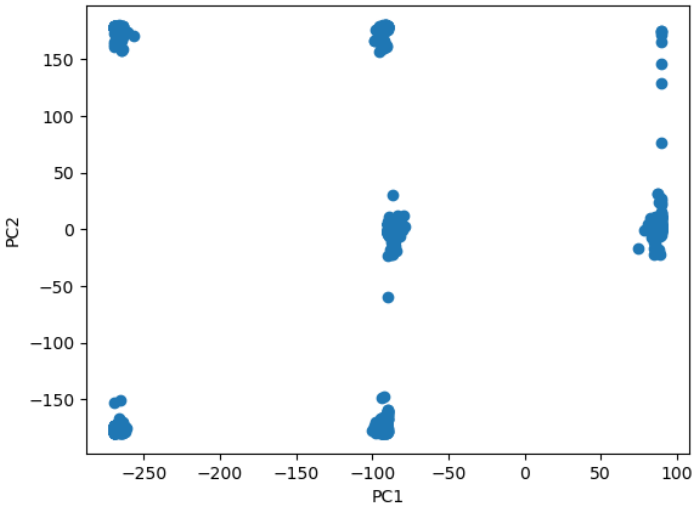
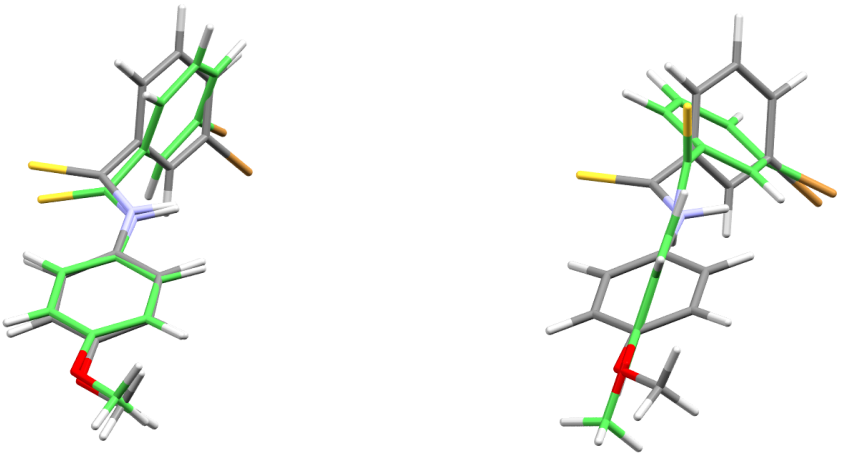



FIGURE 5.7: Geodesic principal components of the torsion angles for DADNUR. RDKit was used to generate molecular conformations, from which torsion angles were calculated. Clustering of conformations is observed, suggesting the presence of similar geometries.

By viewing the first two Principal Component (PC)s, conformers which possess similar geometries are clustered together. Selecting geometries from different clusters would result in conformers which were largely distinct in their conformations such as those shown in Figure 5.8. The idea here being that structures belonging to separate clusters may be separated by larger energy barriers such as those shown in Figure 5.6.



(A) Same cluster (B) Different Cluster

FIGURE 5.8: Overlay of two molecules using the smallest root mean square distance between them from different clusters identified by Geodesic Principle Component Analysis. Molecules within the same cluster exhibit similar geometries, with only minor variations in torsion angles throughout the molecule. In contrast, molecules from different clusters display significantly different geometries, characterised by large differences in torsion angles. **KEY:** Grey – carbon; white – hydrogen; red – oxygen; yellow – sulphur; orange – bromine; green – carbons belonging to the second molecule.

To enable automation of the process, k-means clustering is employed to identify clusters of molecular conformations with similar geometries, facilitating the exploration of specific regions of conformational space. To automatically determine the appropriate number of clusters, a silhouette score is calculated for values of $k$ ranging from 1 to 14 shown in Figure 5.9.

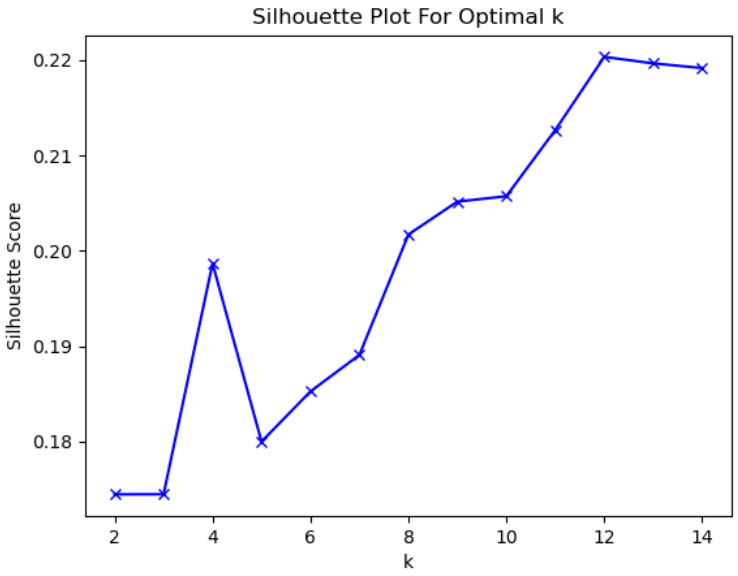



FIGURE 5.9: Example silhouette scores for varying the number of clusters $k$. A higher silhouette score indicates better clustering of data points. In this example, the first maximum occurs at $k = 4$, although high silhouette scores can also be achieved at larger values of $k$.

While the objective is to maximise the silhouette score, it is also essential to consider the computational cost associated with performing conformational searches from numerous starting points. Therefore, efforts are made to limit the number of starting locations to reduce computational expenses. To accomplish this, the first local maximum in the series of silhouette scores is selected. In the example provided, the first maximum occurs when four clusters are identified. The lowest-energy conformer from each cluster is then extracted, and a CREST search is performed on each of these starting conformations.

### 5.3.3 Comparison to Other Methods

The conformer search is improved by utilising starting positions in very different locations on the PES. One conformation was not found by mCREST but was found by LMCS. However, given conformations found elsewhere we claim this method is an improvement over LMCS. For CSP applications, it is unlikely that the conformation missed would yield observable crystal structures due to its relatively middling surface area coupled with its very high energy compared to the global energy minimum conformation.

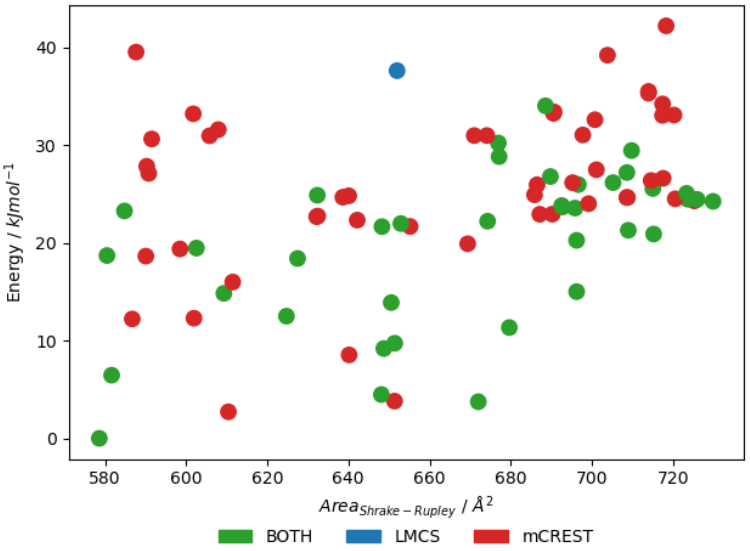



FIGURE 5.10: Comparison of conformational search methods (Low Mode Conformer Search (LMCS) and multiple Conformer Rotamer Ensemble Sampling Tool (mCREST) using the iMTD-sMTD search algorithm. Overall, mCREST performs better than LMCS; however, one conformation is missed by mCREST but identified by LMCS.

## 5.4 Conclusions and Future Work

A routine has been developed for CREST that offers a greater yield of conformations on the DFT PES. One significant advantage of using CREST is its ease of integration into the CSPy program. LMCS is performed within the program Macromodel, which requires a licence. Utilising CREST therefore was beneficial to ensure CSPy become more freely available to users.

mCREST is comparatively more computationally demanding than CREST, as it requires multiple CREST runs. A single CREST search typically consumes around 8–12 Central Processing Unit (CPU) hours for smaller molecules with fewer torsion angles. Most molecules required approximately 24 hours, while the largest molecule took up to 70.1 hours.

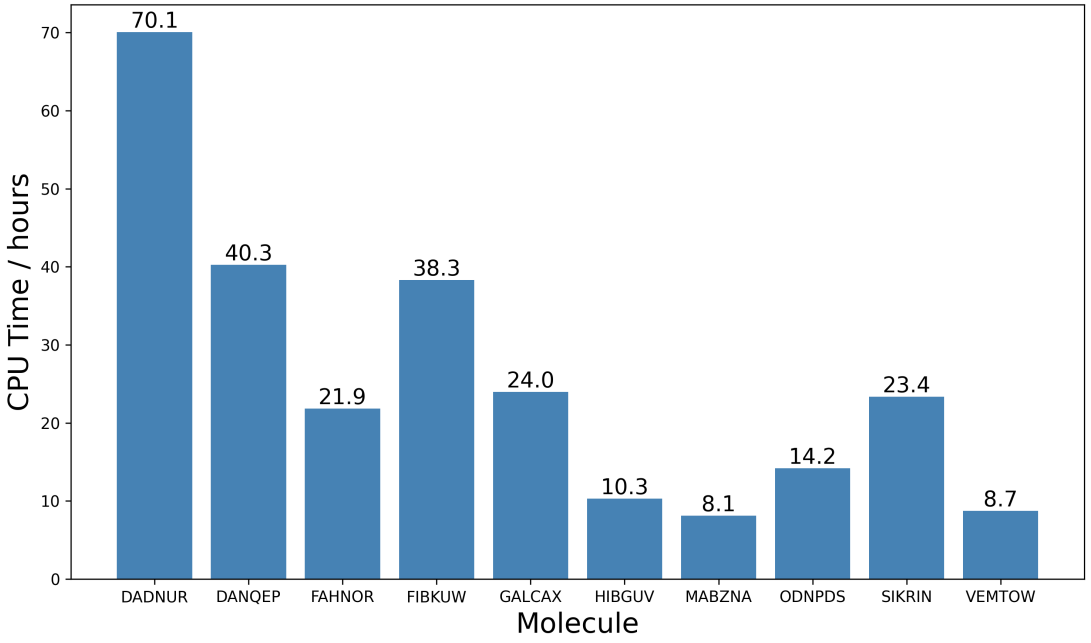



FIGURE 5.11: Computer Processing Unit time taken to perform iMTD-sMTD conformer search using the Conformer Rotamer Ensemble Sampling Tool for test molecules.

The computational cost of mCREST depends on the number of clusters identified after PCA; for instance, a search involving four clusters would result in a total cost equal to four times that of a single CREST run. Nonetheless, the improved performance of mCREST may justify the increased cost, as conformer generation remains considerably less expensive than generating crystal landscapes within the CSP workflow. As CSPy moves towards open-source availability, identifying an alternative to LMCS is essential to enable automated conformer generation and to streamline CSP for flexible molecules.

One drawback of this approach arises during the application of Geodesic PCA, as molecular symmetry is not accounted for. This issue is evident from the symmetry observed

in the distribution of data points along PC1 and PC2. It is recognised that this may result in the starting positions generated for the searches being overly similar.

Future applications of this method should attempt to remove these phenomena such as where there is internal symmetry present within clusters or the geodesic components.

Instances of poor clustering when determining starting positions, such as observed for VEMTOW, are likely attributable to the intrinsic symmetry of the molecule, wherein certain torsions are equivalent.

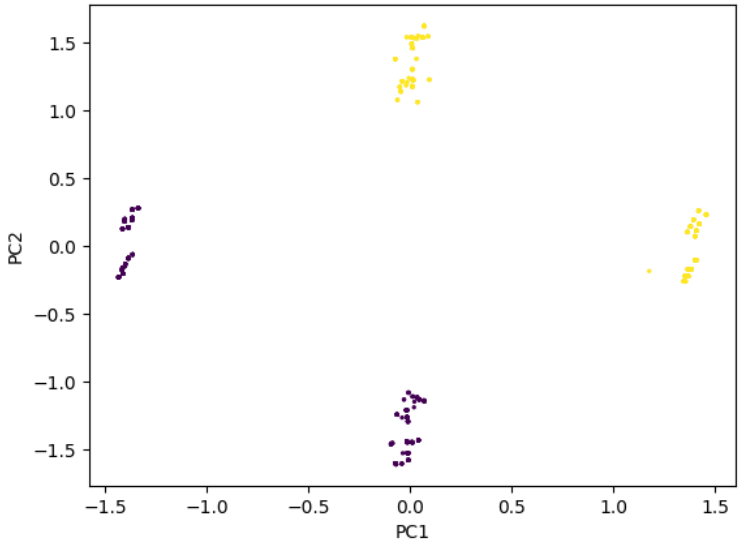
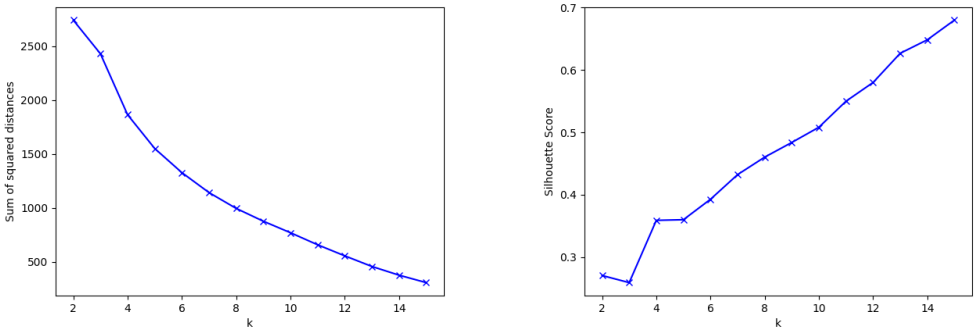



FIGURE 5.12: GeoPCA distribution for VEMTOW. Internal symmetry within the clusters and across PC1 and PC2 is observed, leading to multiple clusters where relatively few would be expected.

In Figure 5.12, no clearly identifiable elbow point is observed. Furthermore, examination of the silhouette scores reveals a maximum at $k = 4$. Although the silhouette score continues to increase beyond $k = 5$, this trend suggests that the data may not be well-suited to clustering.




(A) Elbow plot (B) Silhouette Plot

FIGURE 5.13: Analysis of conformer selection for mCREST searches on VEMTOW. No easily identifiable elbow point is observed. A maximum occurs at $k = 4$, and silhouette scores continue to increase beyond $k = 5$.

While the improvements in the search algorithm nonetheless outweigh this issue, it may still be possible to significantly reduce computational cost by eliminating these redundancies.

# Chapter 6

# Monte Carlo Simulated Annealing

In this chapter, a new approach to CSP is introduced that incorporates experimental data to guide and enhance the prediction process. Specifically, a MCSA procedure is employed that integrates commonly available experimental data to steer CSP towards identifying observable structures with greater confidence. The method simulates PXRD or NMR data, comparing them with their experimental counterparts. By perturbing trial structures and annealing them with temperature, crystal structures are deduced which match experimental data.

CSP typically aims to find the most stable crystal structures of a given molecule using purely computational techniques, primarily through minimisation of total energy, often with limited or no experimental input during the prediction.

In principle, experimental methods alone can determine crystal structures. For example, structure determination is frequently achieved via SCXRD; however, obtaining suitable single crystals can be challenging, requiring significant time and optimisation of crystallisation conditions. Often, organic molecules crystallise as fine crystalline powders, which necessitate characterisation through PXRD or solid state Nuclear Magnetic Resonance (ss-NMR).

Methods exist for determining crystal structures directly from utilising PXRD data alone. Harris and Johnston notably developed a GA to determine structures by fitting calculated powder patterns to experimental PXRD data [102]. The approach evolves a population of candidate structures, optimising molecular orientations and positions to best match experimental PXRD data. The work was later added to by combining R-factor with lattice energy to build a ”hybrid hypersurface” for structure solution [103]. However, the success of methods such as this depend critically on accurate indexing to determine the unit cell parameters from a powder diffraction pattern, which are required to obtain the R-factor. Indexing, usually being performed using software that analyses peak positions to find a lattice consistent with observed diffraction angles. While

automated programs exist, indexing can be time-consuming and may require manual intervention if the pattern is complex or shows peak overlap [104]. For organic crystals, indexing can be particularly challenging because of peak broadening, low symmetry, and preferred orientation effects. Other approaches aim to integrate experimental PXRD data directly into CSP methods [105]. Even partial PXRD information, such as approximate unit cell dimensions or space group symmetry, can significantly constrain the search space, guiding CSP towards physically relevant solutions. Furthermore, simulated PXRD patterns generated from hypothetical CSP predicted structures can be compared directly with experimental PXRD patterns, effectively validating and refining predictions [106–108].

Another data-driven CSP method include MCSA, as demonstrated by Balodis et al., who employed experimental ss-NMR data to guide their algorithm towards correct crystal structures [109]. Similar success has been achieved using electron diffraction [110, 111] and ss-NMR constraints [112–115].

## 6.1 Monte Carlo Simulated Annealing

This work focuses on the use of more accessible PXRD data, as it is often more widely available than NMR data. Similar to other methods, a generated PXRD pattern is compared to the experimental pattern and subsequently optimised. The strength of this method lies in the application of cDTW for pattern comparison, eliminating the need for indexing and thereby potentially streamlining the process considerably.

To validate this method, tests have been conducted using both experimental and simulated data, demonstrating that it can be used to predict the crystal structures of various polymorphs accurately and reliably under different pressures. Furthermore, the approach introduces a novel way of matching experimental data with CSP results and is versatile, with potential for adaptation to incorporate other types of experimental data in the future.

## 6.2 Method

A general method was employed that utilises MC moves to minimise the cost function defined by Equation 6.1.

$$E_{\text{pseudo}} = E_{\text{total}} + E_{\text{PXRD}} \,, \tag{6.1}$$

where $E_{\text{total}}$ is the sum of the intermolecular and intramolecular energies of the crystal structure normalised per molecule, $E_{\text{PXRD}}$ is a pseudo-energy term that is attributed to the difference between the crystal structure and experimental PXRD pattern. This term is defined as $E_{\text{PXRD}} = \lambda D_{\text{cDTW}}$, where $D_{\text{cDTW}}$ is the cDTW dissimilarity measure between the simulated PXRD pattern of a trial crystal structure and the experimental PXRD pattern. $\lambda$ is a scaling factor with the units kJ mol$^{-1}$. Selecting the value of $\lambda$, allows the user to determine the relative weight of the cDTW distance term compared to crystal energy. This term steers the optimisation towards structures whose simulated PXRD agrees with the experimental diffraction pattern.

Each crystal generated is a single trajectory. The pseudo energy of the initial trial structure is calculated before a random change is attempted on one of the parameters defining the crystal structure. The types of move allowed are molecular translation, molecular rotation changes in unit cell lengths, angles, and volume. For flexible molecules, changes in torsion angles around selected bonds are also included. The magnitude of this change is selected randomly between specified upper and lower limits which are shown in Table 6.1. Space group symmetry is preserved throughout the simulation, so that molecular moves (translation, rotation and torsion angles) are applied only to the asymmetric unit.

| Parameter | Change / ± | Degrees of Freedom |
|---|---|---|
| Flexible torsion | 3.5° | T |
| Unit cell volume | 25 Å$^3$ | 1 |
| Unit cell length | 0.5 Å | 3 |
| Unit cell angle | 0.5° | 3 |
| Molecular translation | 0.5 Å | 3 |
| Molecular rotation | 0.05° | 3 |

TABLE 6.1: Upper and lower boundaries for move types in the Monte Carlo simulated annealing (MCSA) protocol. $T$ denotes the number of flexible torsions in the asymmetric unit.

MC move sizes were selected to produce small geometric changes in the crystal structure. The relative change in pseudo energy, i.e. the difference before and after the MC move, is calculated. The move is then accepted based on a probability, $P_{\text{acc}}$.

$$P_{acc} = \exp\left(-\frac{\Delta E_{\text{pseudo}}}{RT_n}\right) , \tag{6.2}$$

where $\Delta E_{\text{pseudo}}$ is change in pseudo energy, $R$ is the universal gas constant, $T_n$ is the temperature at step $n$.

The temperature of the system is defined by the user through a starting and a final temperature. The temperature at the first step is equal to the starting temperature, while the temperature at the final step matches the final temperature. If a step is accepted, the structure is annealed by progressively reducing the temperature towards the final value. This temperature reduction continues until the target number of steps has been completed. The decrease in temperature follows a linear profile, as described further in section 6.5.4.

A trajectory was also permitted to terminate before reaching the maximum number of accepted steps if 120 consecutive MC steps were rejected.

An overview of the workflow is depicted in Figure 6.1.

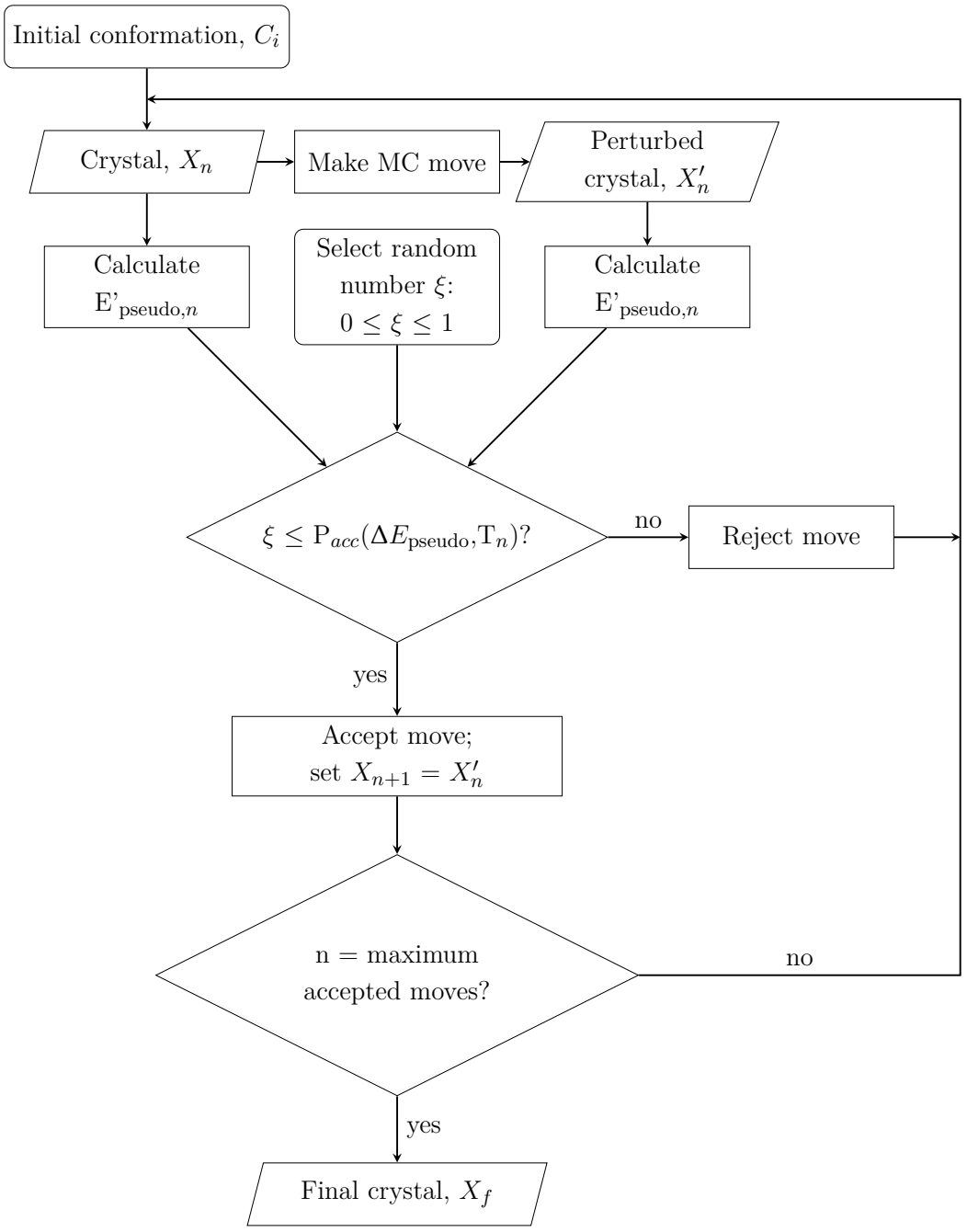
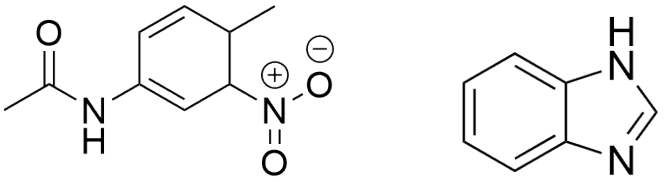



FIGURE 6.1: Monte Carlo simulated annealing (MCSA) workflow for a single trajectory to identify matches with experimental data. In this process, a conformer $C_i$ is used to generate an initial crystal structure $X_0$ via the crystal landscape generator. The pseudo energy at step $n$, $E_{\text{pseudo},n}$, of the crystal is calculated as defined in Equation 6.1. The crystal is then perturbed via a Monte Carlo (MC) move to form $X'_n$, and the pseudo energy is recalculated as $E'_{\text{pseudo},n}$. The change in pseudo energy, $\Delta E_{\text{pseudo}}$, is determined, and the move is accepted based on the probability $P_{\text{acc}}$ defined by Equation 6.2. If the move is rejected, a different MC move is attempted. If accepted, the perturbed crystal becomes the structure of crystal at the next step, such that $X_{n+1} = X'_n$. This process is repeated until a predetermined number of MC steps have been accepted.

## 6.3 Experimental Data

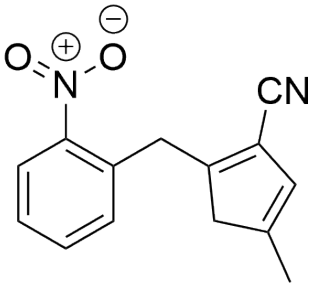
Three molecules shown in Figure. 6.2, were studied in the development of the method:

- N-(4-methyl-2-nitrophenyl)acetamide
- Benzimidazole
- 5-methyl-2-((2-nitrophenyl)amino)thiophene-3-carbonitrile which is often referred to as ROY due to its red, orange and yellow polymorphs.




(A) N-(4-methyl-2-nitrophenyl)acetamide

(B) benzimidazole




(C) ROY

FIGURE 6.2: Chemical diagrams of molecules used in the Monte Carlo simulated annealing procedure. N-(4-methyl-2-nitrophenyl)acetamide and ROY contain multiple freely rotatable bonds, whereas benzimidazole is rigid with no freely rotatable bonds.

The molecules were selected due to their availability of multiple PXRDs. Experimental data was obtained from Lunt et al. who completed an automated syntheses and PXRD diffraction of benzimidazole and ROY [116]. PXRD data from this work was taken for each of the 8 samples of benzimidazole and ROY which were then used to guide MCSA. In this thesis, these will be referred to as the PXRD patterns of experimental samples 1-8 for both molecules. Each of the 8 PXRD patterns for benzimidazole resemble the alpha polymorph, whilst the 8 PXRD patterns for ROY mostly resemble the monoclinic orange needle (ON) polymorph. It has been identified that in sample 4, the monoclinic yellow (Y) polymorph may also be present as well as the ON polymorph. Different polymorphs of benzimidazole were investigated by obtaining PXRD patterns for each form. The PXRD patterns of BZDMAZ02, BZDMAZ03, and BZDMAZ07 were simulated to represent the alpha, beta, and gamma polymorphs, respectively [117–119]. For the monoclinic (ON) polymorph of ROY, QAXMEH01 [120] was used. Additionally, the PXRD

pattern of the monoclinic amber polymorph of N-(4-methyl-2-nitrophenyl)acetamide was simulated using MNIAAN02 from the CSD.

### 6.3.1 Experimental Matches

COMPACK, as described in section 3.4.4, was employed to determine whether the generated structures matched the experimental structure.

## 6.4 Initial Results

Prior to testing the procedure, an assessment was conducted to determine whether matches to the MNIAAN02 system could be identified without relying on any PXRD information. This served as a control to evaluate whether the method performed differently if such data was available. The MCSA procedure was carried out using N-(4-methyl-2-nitrophenyl)acetamide as the input molecule, employing the parameters listed in Table 6.2 for this testing.

| Parameter | Value |
|---|---|
| Initial Temperature | 2500 K |
| Final Temperature | 100 K |
| Total Accepted MC Steps | 4000 |
| Total Trajectories | 1000 |

TABLE 6.2: Initial parameter set used for Monte Carlo simulated annealing (MCSA) during preliminary testing on the MNIAAN02. The settings span a wide thermal window and include many Monte Carlo steps. A total of 1000 crystal structures are generated, each using a single trajectory.

The set of parameters was selected to ensure a sufficiently high temperature to overcome energy barriers between minima. A suitable number of MC steps was targeted to enable the simulation to reach the global minimum and thoroughly explore the landscape, in line with the chosen MC move size.

The MCSA procedure was carried out starting from 1000 QR crystal structures. This approach was employed to demonstrate that the methodology could function without incorporating PXRD data, relying exclusively on the minimisation of $E_{\text{total}}$ by setting $\lambda = 0$ kJ mol$^{-1}$.

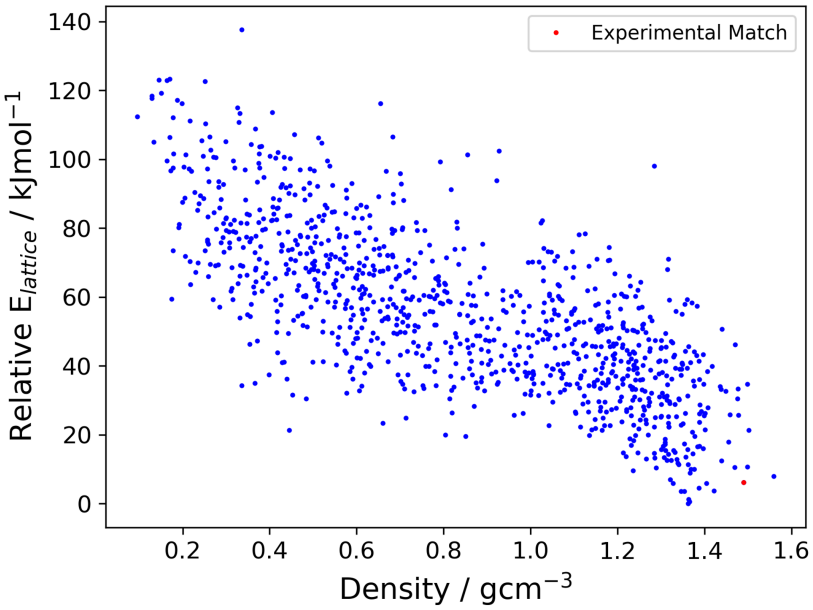
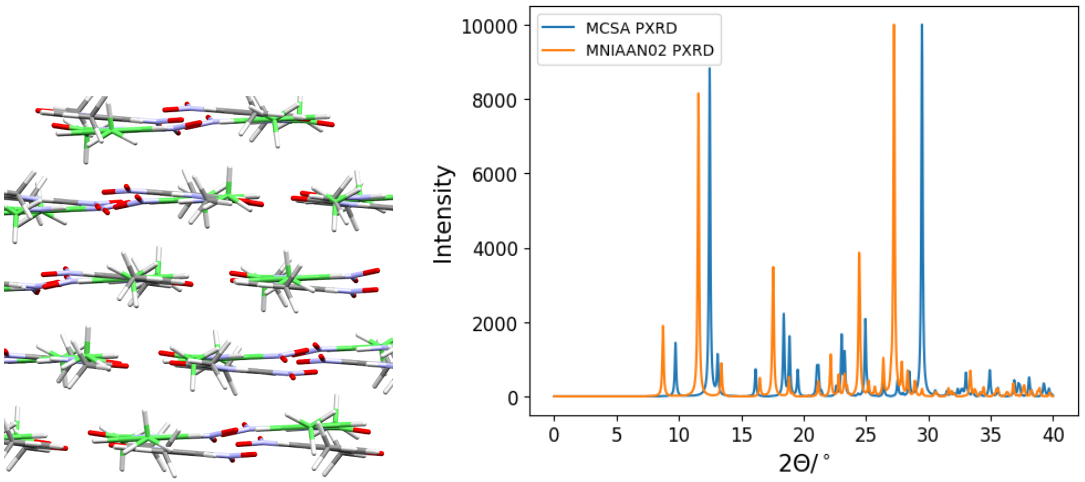



FIGURE 6.3: Crystal landscape for the search of MNIAAN02 using Monte Carlo simulated annealing with $\lambda = 0$ kJ mol$^{-1}$. Each data point represents the final structure of a single trajectory. Points in red are crystal structures which match with the experimentally observed structure, whilst points in blue did not match.

The final structure of a single trajectory matched with the structure of MNIAAN02 [121]. However, the accuracy of the match was poor overall and there was a significant mismatch in the generated PXRD patterns of either structure shown in Figure 6.4.




(A) Geometric overlay (B) PXRD Pattern

FIGURE 6.4: Comparison of experimental matches from Monte Carlo simulated annealing during initial testing targeting MNIAAN02. The settings used include 4000 steps and $\lambda = 0$ kJ mol$^{-1}$. There is reasonable overlap between crystal structures, and the packing is mostly similar. Some discrepancies are observed in the alignment of molecules. The PXRD patterns show that the peaks are shifted, though they still display some resemblance. **KEY:** Grey – carbon; white – hydrogen; red – oxygen; blue – nitrogen; green – carbons belonging to the second crystal structure.

The same methodology was again performed, setting the $\lambda = 20$ kJ mol$^{-1}$, to identify whether altering the value of $\lambda$ provided sufficient guidance for the simulation. The generated PXRD pattern of MNIAAN02 from the CSD was used to guide the MCSA procedure. It is hypothesised that the addition of PXRD to the system will improve the calculation as both $E_{\text{total}}$ and $E_{\text{PXRD}}$ are optimised.

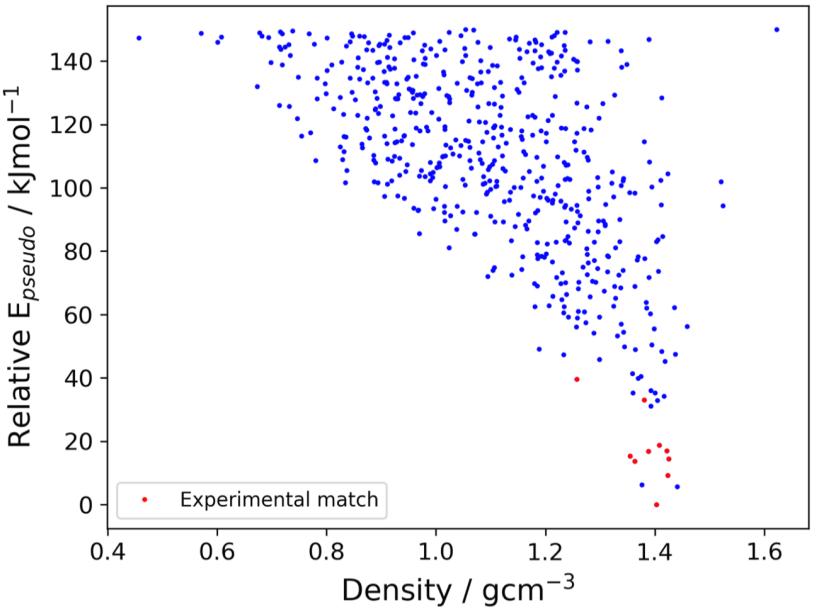



FIGURE 6.5: Crystal landscape for the search of MNIAAN02 using Monte Carlo simulated annealing with $\lambda = 20$ kJ mol$^{-1}$. Each data point represents the final structure of a single trajectory. Points in red are crystal structures which match with the experimentally observed structure, whilst points in blue did not match. 10 experimental matches have been found which all exist in the low energy region of the crystal landscape.

By increasing the parameter $\lambda$, significantly more matches were found to the target crystal. The accuracy of the match according to $E_{\text{pseudo}}$ also improved, ranking the structure lowest in energy on the landscape. In addition, there was also an improvement in the match according to the lattice parameters and RMSD summarised in Table 6.3.

| Determination | a / Å | b / Å | c / Å | $\alpha$ / ° | $\beta$ / ° | $\gamma$/ ° | $\rho$ / g cm$^{-3}$ | RMSD$_{30}$ |
|---|---|---|---|---|---|---|---|---|
| MNIAAN02 | 10.158 | 11.635 | 8.041 | 90.000 | 94.550 | 90.000 | 1.362 | - |
| MCSA $\lambda = 0$ | 9.089 | 11.362 | 8.385 | 90.000 | 91.968 | 90.000 | 1.490 | 0.672 |
| MCSA $\lambda = 20$ | 10.005 | 11.484 | 8.005 | 90.000 | 91.381 | 90.000 | 1.403 | 0.260 |

TABLE 6.3: Crystal lattice parameters for the lowest root-mean-square deviation crystal structure compared to experimental crystal obtained after performing Monte Carlo simulated annealing (MCSA) in the prediction of MNIAAN02 for different values of $\lambda$. For comparison, the lattice parameters of MNIAAN02, which did not undergo any relaxation, are also shown. The calculation using MCSA with $\lambda = 20$ kJ mol$^{-1}$ produced lattice parameters that aligned more closely with the experimental crystal structure than those from MCSA with $\lambda = 0$ kJ mol$^{-1}$.

## 6.5 Parameterisation of MCSA

This section describes the process used to parametrise the method. Parameterisation is crucial to ensure optimal performance and versatility across a wide range of different molecules. The workflow was parametrised by scanning each parameter across a series of sensible values across a range of systems.

### 6.5.1 Optimisation of Lambda

The value for $\lambda$ should be selected so that when an MC move is made the relative change in energy of crystal $\Delta E_{total}$ and $\Delta E_{PXRD}$ are approximately equal to ensure that neither term dominates the cost function, Equation 6.1.

Due to the change in the topology of any one pseudo energy landscape for any given crystal, the relative weights of $\Delta E_{total}$ and $\Delta E_{PXRD}$ may differ with each MC step. Therefore it is difficult to identify the effect of $\lambda$ through individual MC steps and that a full simulation needs to be performed for testing. A suitable value is identified by conducting simulations and scanning across different values of $\lambda$.

The MCSA workflow was performed using the alpha polymorph of benzimidazole and the monoclinic amber polymorph of N-(4-methyl-2-nitrophenyl)acetamide. Calculations were run for each of different values for $\lambda$ and utilised the hit rate as a metric to identify an ideal value whereby a high hit rate provides a greater probability to find an experimental match from any single starting position. Changing the parameter $\lambda$ has informed how much the simulation steers towards either total energy or PXRD energy. Testing indicates that the amount in which the PXRD energy influences a trajectory is system dependant and also depends on the quality of the experimental PXRD. In a blind study, testing all possible values for $\lambda$ may not be feasible; however, sensible values between 0 - 40 kJ $mol^{-1}$ were selected.

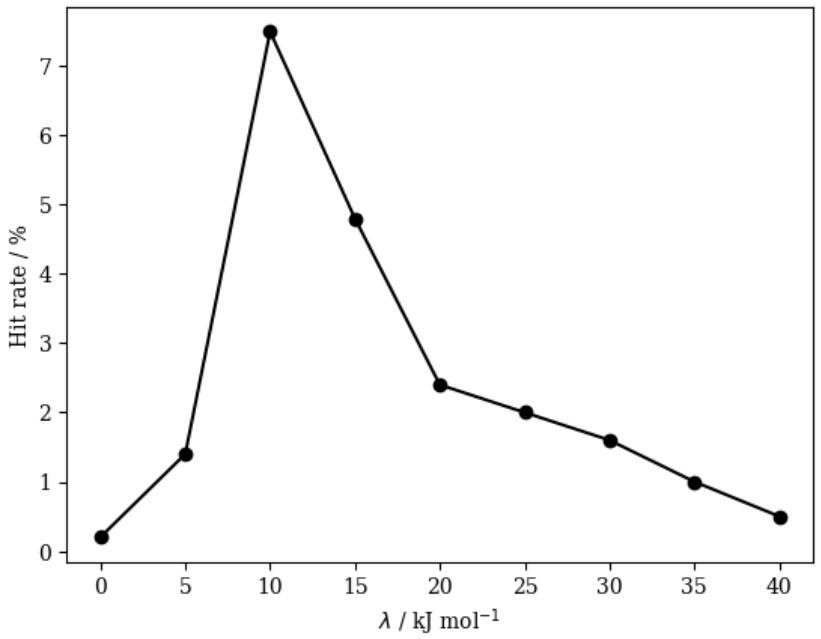



FIGURE 6.6: Hit rate for Monte Carlo simulated annealing using different values of $\lambda$, which dictates the influence of $E_{\mathrm{PXRD}}$ on the system targeting BZDMAZ02, a crystal containing an asymmetric unit with no flexible torsions. A value of $\lambda = 10$ kJ mol$^{-1}$ was found to provide the best hit rate, indicating the probability of finding the experimental structure from any starting position.

For benzimidazole, a value of $\lambda = 10$ kJ mol$^{-1}$ yielded the highest number of matches as shown in Figure 6.6.

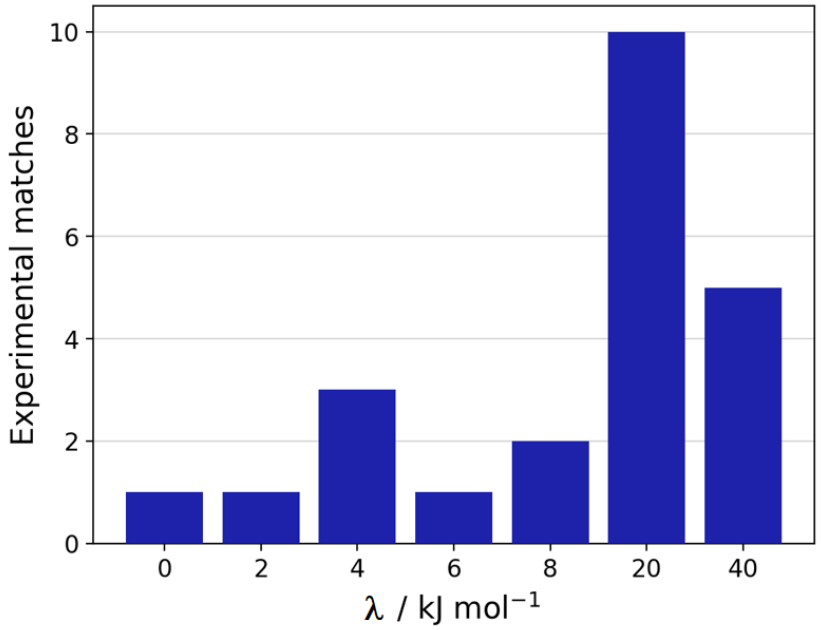



FIGURE 6.7: Number of matches for Monte Carlo simulated annealing using different values of $\lambda$, which dictates the influence of $E_{\text{PXRD}}$ on the system targeting MNIAAN02, a crystal containing an asymmetric unit with multiple flexible torsions. A value of $\lambda = 20$ kJ mol$^{-1}$ provided the greatest number of experimental matches across 1000 trajectories.

For N-(4-methyl-2-nitrophenyl)acetamide, a value of $\lambda = 20$ kJ mol$^{-1}$ yielded the highest number of matches as shown in Figure 6.7.

The use of cDTW to calculate PXRD energy reliably produced matches to experimental structures across a range of different values for beta but found the most experimental matches between a range of 10 - 20 kJ mol$^{-1}$ for rigid and flexible molecules.

### 6.5.2 Adaptive and Static Move Styles

Two different approaches were tested for when making MC steps, static or adaptive move sizes. In the previously tested static case, the scale of the move sizes are fixed and are selected between upper and lower limits which remain constant throughout the simulation. It was observed that towards the end of a trajectory, the acceptance rate of MC moves decreased significantly, leading to an increase in computational cost.

An alternative adaptive approach could be used which allows the size of the MC moves to adapt within the procedure based upon the recent history of its acceptance rate. A target acceptance rate of 0.5 was established, defined as the ratio of accepted moves to the total number of attempted moves. To achieve this target, the scale of the MC moves was adjusted. Specifically, the move scale was increased if the acceptance rate fell below the target value, and decreased if it exceeded the target.

Both move types were tested on three crystal polymorphs of benzimidazole, and experimental matches were identified at the global energy minimum for all methods. The adaptive method demonstrated greater efficiency for the alpha and gamma polymorphs but performed less effectively for the beta polymorph as shown in Figure 6.8.

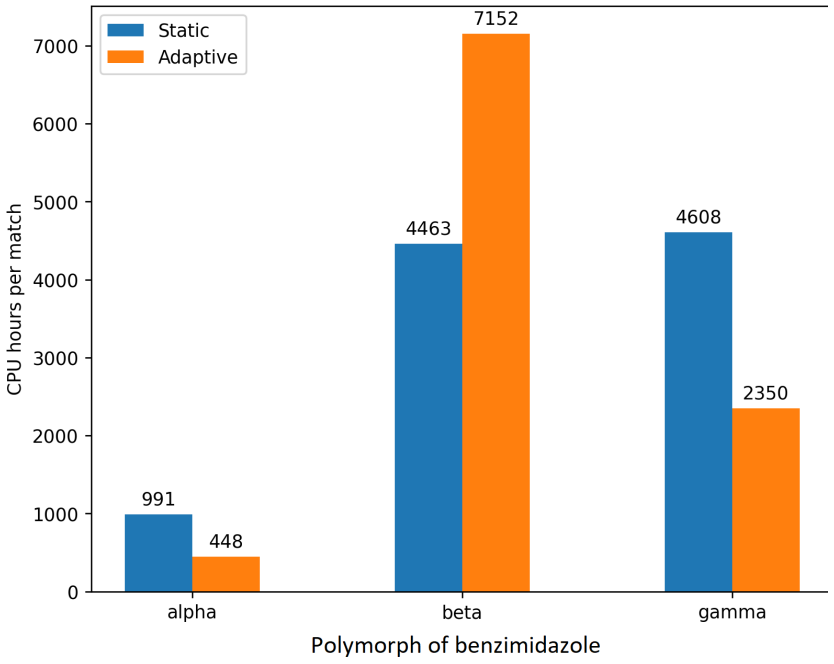



FIGURE 6.8: Cost efficiency of static and adaptive move types in Monte Carlo simulated annealing targeting the different polymorphs of benzimidazole. For the alpha and gamma polymorphs, utilising the adaptive move type significantly reduced computational cost; however, for the beta polymorph, the adaptive move type significantly increased computational cost.

Part of the success of the adaptive move type is attributed to its capacity for small adjustments, refining structures to align with the PXRD pattern. The final move scales of matching structures were approximately ten times smaller than the initial starting move scale, ranging from 0.062 - 0.135. This significant reduction resulted from fewer overall MC attempts. Although it remains unclear which method is ultimately superior, the adaptive move style has been selected for continued use.

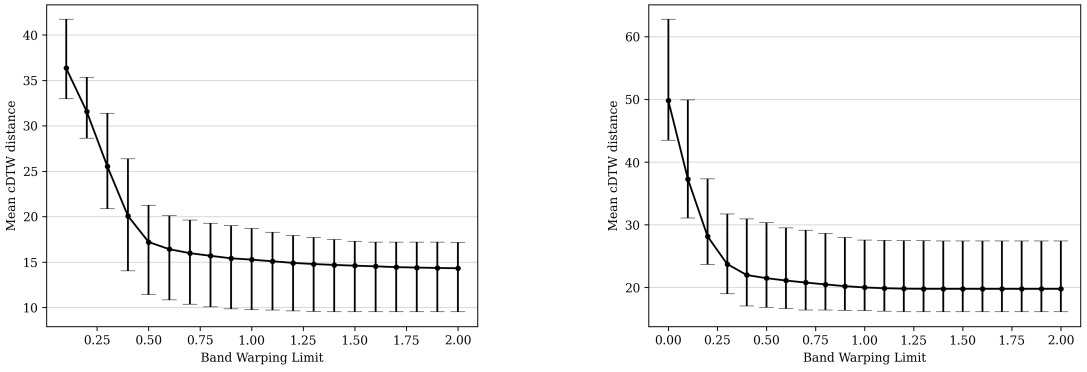
### 6.5.3 Band Warping Limit

The constrained dynamic time warping distance is employed to measure the similarity between two PXRD patterns [122–124]. The warping limit permits slight alterations in the PXRD pattern without significantly affecting the value of $E_{\mathrm{PXRD}}$ in the cost function. This flexibility is necessary because PXRD patterns often do not align perfectly due to physical and practical factors, yet they can still correspond to the same crystal structure. The warping effect allows minor shifts in peak positions to be accommodated, ensuring

a more accurate comparison of diffraction data. The degree of warping can be adjusted, effectively shaping the cost landscape associated with $E_{\mathrm{PXRD}}$, and providing flexibility when calculating the cDTW. A high warping limit can make it challenging to distinguish between good and poor matches, as peaks can be mismatched, effectively broadening the basin associated with $E_{\mathrm{PXRD}}$ and potentially causing unrelated patterns to appear similar. Conversely, an extremely small warping limit may prove too restrictive, resulting in a narrow basin that fails to account for realistic experimental variations and thus offers limited benefit for guiding the comparison.

The level of constraint should be parametrised to be suitable for systems including both simulated and experimental data. When comparing PXRD patterns using cDTW, it is necessary to account for physical phenomena that influence PXRD patterns. For instance, preferred orientation within experimental crystals arising from growth conditions or mechanical processing influence the relative peak intensities observed which mean that some peaks could become invisible and peak shape can be effected by grain size.

To determine the optimal value for the band warping limit, the cDTW distance was calculated between each of the eight experimental patterns for benzimidazole and ROY and the corresponding PXRD patterns of the experimental crystals from the CSD. The aim of this was to identify how similar or different PXRD patterns from the automated synthesis were from the confirmed single crystal structure. This would enable us to capture the required amount of warping sufficient to match a structure to the experimental pattern.




(A) Benzimidazole (B) ROY

FIGURE 6.9: The effect of band-warping limits on the constrained dynamic time warping (cDTW) distance for benzimidazole and ROY. For each molecule 8 sets of experimental powder X-ray diffraction patterns (PXRD)s are compared against a simulated PXRD of the corresponding crystal in the Cambridge Crystallographic Database. Error bars indicate the range of distances across the set of patterns. In both molecules, an identifiable elbow point is observed at 0.5 for benzimidazole and between 0.25 and 0.50 for ROY.

The cDTW distances in Figure 6.9 show that while increasing the band warping limit lowered the cDTW distance between any pair of patterns, an elbow point was identified at 0.50 for benzimidazole. Whilst less identifiable, a similar value of between 0.25 and 0.50 was also identified for ROY. These values were of interest as they allowed for enough warping to account for small differences between the patterns, but insufficient to allow for false matches to be present. A band warping limit 0.5 was used for PXRD comparisons hereafter.

### 6.5.4 Temperature

The probability of accepting an MC move which increases its pseudo energy is related to the current temperature (Equation. 6.2) and therefore the selection of temperatures is an important feature to consider during the procedure. A high temperature means that MC moves are more likely to be accepted when the pseudo energy is increased, whilst lower temperatures reduce this probability as shown in Figure 6.10.

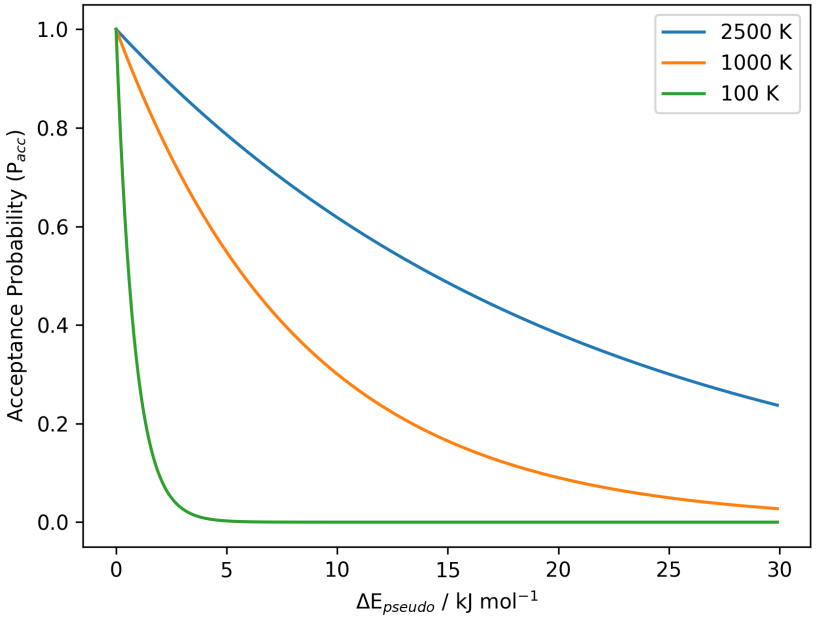



FIGURE 6.10: Effect of temperature on the acceptance probability for a step in the Monte Carlo simulated annealing procedure. Low temperatures permit smaller positive changes in the pseudo energy, $E_{\text{pseudo}}$, compared to higher temperatures.

#### 6.5.4.1 Starting and Finishing Temperatures

The starting and finishing temperatures for the MCSA simulations are parameters users can specify in the MCSA settings. These parameters act as fixed points in the simulation and cannot change.

The aim of the MCSA procedure is to identify the global minimum on the pseudo energy landscape. As is typical with any simulated annealing process, beginning at a high temperature facilitates exploration of the hypersurface by overcoming energy barriers, since the acceptance probability for a positive change in pseudo energy remains high. As MC moves are made, the energy is gradually lowered, making moves that increase the pseudo energy less likely to be accepted. This process aims to focus on structures exhibiting low pseudo energies.

A temperature of 100 K was initially chosen to be a reasonable final temperature as it allowed for small changes in the pseudo energy. The MCSA procedure was performed using 100 K as a final temperature. However it was found that this led to many structures getting close to the experimental structure, but not close enough to be an experimental match.

Lowering the temperature further to 0 K allowed the trial structure to more closely align to the experimental powder pattern and therefore be a closer match geometrically. These results are shown in Table 6.4

| Final Temperature / K | Experimental Matches | Total CPU Time (hrs) |
|---|---|---|
| 100 | 72 | 12,455 |
| 0 | 82 | 10,340 |

TABLE 6.4: Comparison of experimental matches and computational cost for different final temperatures targeting BZDMAZ02 using $\lambda = 10$ kJ mol$^{-1}$ in Monte Carlo simulated annealing for 1000 trajectories. Utilising a final temperature of 0 K results in a lower computational cost and yields a greater number of experimental matches.

Ideally, all trajectories should lead to the experimental structure. Examining the rate at which matches are identified for a given number of trajectories enables identification of the parameters most effective in directing starting positions towards a global minimum. Doing so would not only give greater confidence but also allow us to perform fewer trajectories, decreasing computational cost.

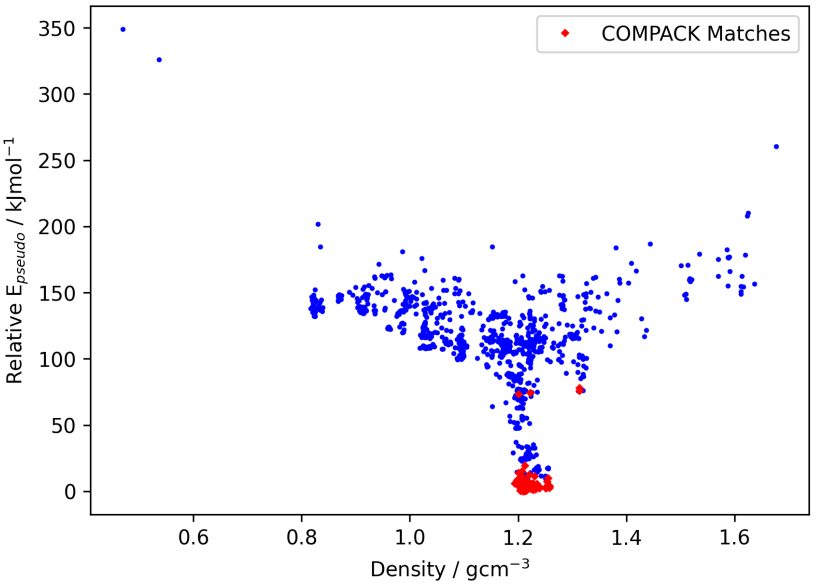



FIGURE 6.11: Crystal landscape for the search of BZDMAZ02 using Monte Carlo simulated annealing with $\lambda = 10$ kJ mol$^{-1}$ with a final temperature of 0 K. Each data point represents the final structure of a single trajectory. Points in red are crystal structures which match with the experimentally observed structure, whilst points in blue did not match. Many experimental matches have been found which exist in the low energy region of the crystal landscape, some structures match in higher energy regions.

#### 6.5.4.2 Temperature Profiles

Previously, a linear temperature profile was explored, in which the change in temperature between accepted MC steps remains constant.

$$T_n = T_i - n \cdot \frac{T_i - T_f}{n_{max}} , \tag{6.3}$$

where $n$ in the step number, $T_n$ is the temperature at step $n$, $T_i$ and $T_f$ are the initial and final temperatures and $n_{max}$ is the maximum number of steps. The linear profile allows for change in the temperature at each step but this could be optimised further using other temperature profiles [125, 126] as part of the MCSA procedure.

Another alternative profile is exponential where the temperature changes exponentially with step number.

$$T_n = T_i \left(\frac{T_f}{T_i}\right)^{\frac{n-1}{n_{max}}} , \tag{6.4}$$

The exponential temperature profile allows the algorithm to spend more steps exploring at medium and low temperatures, while spending fewer steps at high temperatures

as shown in Figure 6.12. The profile is naturally cooler, having a mean temperature throughout procedure lower than that of a linear profile for the same starting and final temperatures.

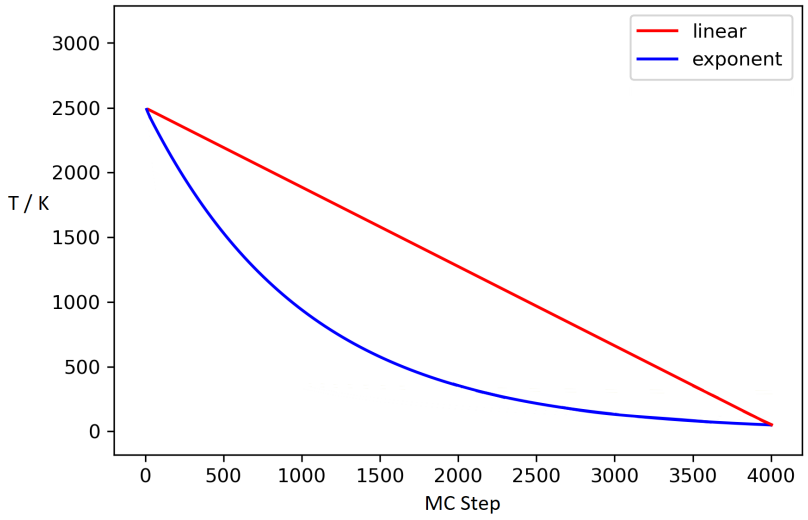
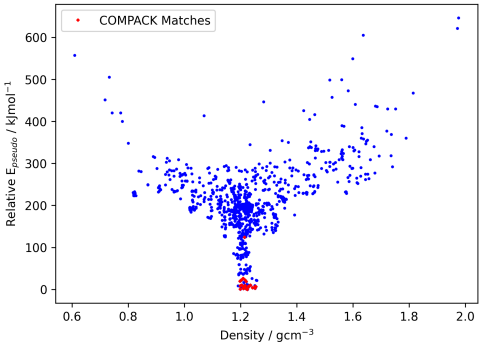
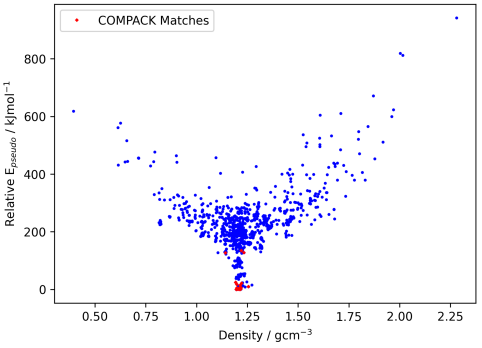



FIGURE 6.12: Comparison of linear and exponential temperature profiles that can be used for the Monte Carlo simulated annealing procedure. The various profiles determine how many steps should be spent in higher or lower temperature regions, given a starting temperature of 2500 K and a final temperature of 100 K. Some profiles do not allow for specifying both a fixed starting and finishing temperature.

Two MCSA runs were completed each with the linear and exponential profiles using the BZDMAZ02 PXRDs with $\lambda = 20$ kJ mol$^{-1}$.




(A) Linear, 24 experimental matches




(B) Exponential, 18 experimental matches

FIGURE 6.13: Crystal landscape for the search of BZDMAZ02 using Monte Carlo simulated annealing with $\lambda = 20$ kJ mol$^{-1}$ with a final temperature of 100 K for linear and exponential profiles. Each data point represents the final structure of a single trajectory. Points in red are crystal structures which match with the experimentally observed structure, whilst points in blue did not match. Many experimental matches have been found which exist in the low energy region of the crystal landscape, some structures match in higher energy regions. Using a linear profile 24 experimental matches were found compared to the exponential temperature profile in which 18 matches were found.

The exponential temperature profile performed worse than the linear profile, yielding only 18 matches compared to 24 for the linear case. It's possible that higher temperatures were required to better explore the configurational space. Additionally, the exponential profile incurred slightly higher computational costs due to a lower average acceptance rate across all temperatures.

### 6.5.5 Maximum Number of Monte Carlo Steps

Studies have demonstrated that 4,000 accepted MC steps can yield a reasonable number of matches to experimental structures across various molecules. However, this number of steps may be excessive and potentially waste computational resources. Striking a balance between achieving experimental matches and minimising the number of steps is crucial for improving computational efficiency. To explore this balance, the method was tested using different limits for the maximum number of MC steps.

The MCSA procedure was conducted on a range of 50 - 6000 maximum steps. The performance of these calculations are shown in Figure 6.14.

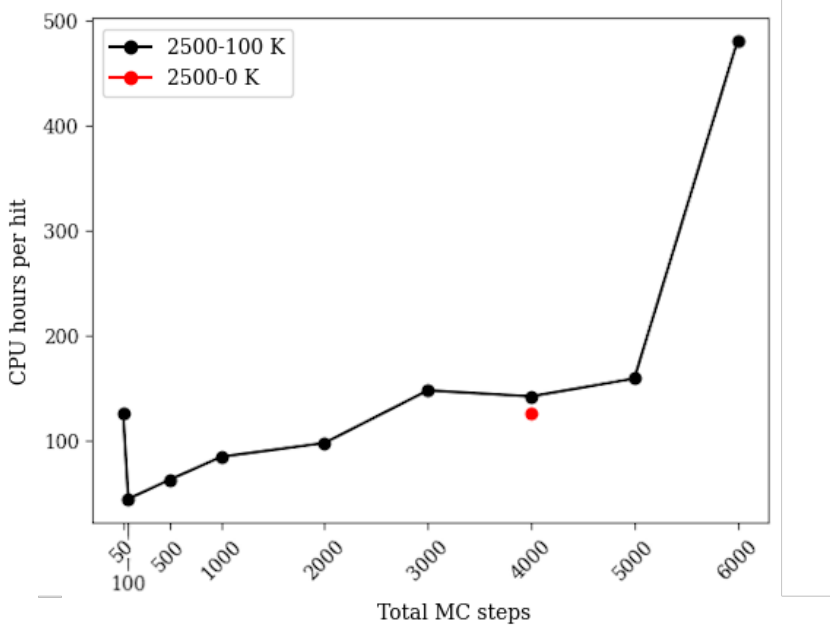



FIGURE 6.14: Effect of maximum number of Monte Carlo (MC) steps for Monte Carlo Simulated Annealing runs for benzimidazole targeting BZDMAZ02 using $\lambda = 10$ kJ mol$^{-1}$.

Only 50 MC steps are required to identify matches to the experimental structure, although 100 MC steps offer optimal CPU efficiency. Both step counts successfully reproduce the landscapes observed using 4000 steps while maintaining a good ranking of the experimental structure match in terms of pseudo energy. To verify that the results were

not obtained merely by chance, additional tests confirmed reproducibility using different random seeds for the MC moves.

Although optimal efficiency is achieved at 100 MC steps, it is noteworthy that the hit rate remained at 0.6%. Consequently, generating a larger number of QR structures would be advisable to confidently identify experimental matches. For systems featuring a greater number of degrees of freedom, such as those with rotatable bonds, 100 MC steps may prove insufficient. Therefore, the maximum number of steps should be carefully adjusted to accommodate system complexity. For systems of this size, a compromise could involve employing 500 MC steps, which delivers good CPU efficiency alongside a reasonably high hit rate for this system.

### 6.5.6 Pressure

Pressure plays an important role in stabilising crystal structures, so much so that many structures undergo polymorphic changes under high pressures [127]. At ambient temperatures (∼1 atmosphere), the effect of the work done from pressure is negligible such that a change in volume of 0.1 g $cm^{-3}$ corresponds to around 0.8 J $mol^{-1}$ of work done. As a result, the presence of pressure and its effect on stability has been neglected so far.

At high pressures however, the amount of work done on stabilising a crystal is significant and can cause significant changes to the crystal landscape. When investigating crystal structures known to be observed at high pressures, the effect of pressure can be added to the cost function which reflects this work done.

$$E_{\text{pseudo}} = E_{\text{total}} + E_{\text{PXRD}} + PV\,, \tag{6.5}$$

where $P$ is the pressure and $V$ is the volume normalised per molecule.

An MCSA run with benzimidazole using the simulated PXRD of BZDMAZ07 (high pressure gamma polymorph) was used as input. The revised cost function was used, leading to the landscape shown in Figure 6.15.

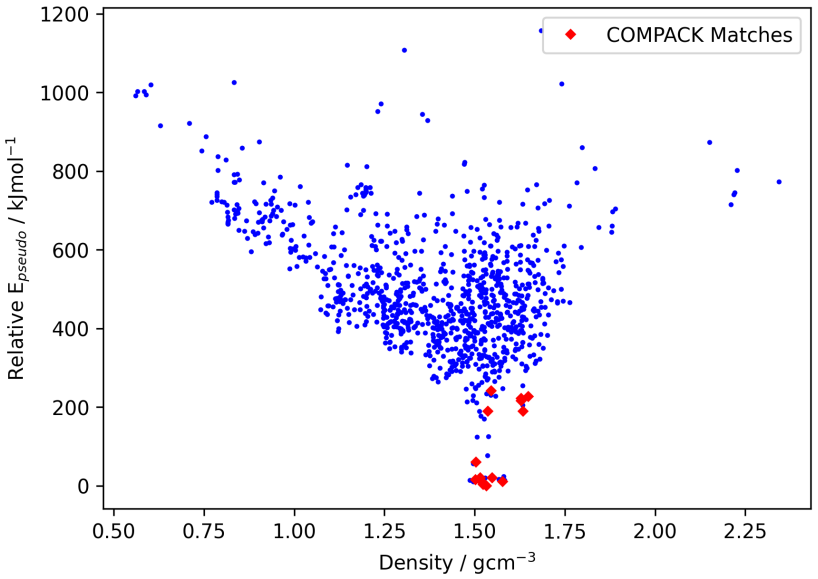



FIGURE 6.15: Crystal landscape for the search of BZDMAZ07 using Monte Carlo simulated annealing including pressure term with $\lambda = 20$ kJ mol$^{-1}$ with a final temperature of 100 K for the linear profile. Each data point represents the final structure of a single trajectory. Points in red are crystal structures which match with the experimentally observed structure, whilst points in blue did not match. Many experimental matches have been found which exist in the low energy region of the crystal landscape, some structures match in higher energy regions.

It is noted that, to identify an experimental match for the high-pressure gamma polymorph, the inclusion of pressure is not strictly necessary. However, incorporating pressure enhances the efficiency of the search, as the procedure preferentially identifies structures of higher density. Consequently, the search yields more matches for the same computational cost.

## 6.6 Experimental PXRD Patterns

Thus far, the methodology has been tested on PXRD patterns generated from experimental structures in the CSD. As these PXRD patterns originate from the structures used for matching, they serve as the "ideal" PXRD. The methodology will now be applied to real-world experimental PXRD data and compared against structures within the CSD.

### 6.6.1 Benzimidazole

MCSA was performed on the eight experimental PXRD patterns for the benzimidazole alpha polymorph.

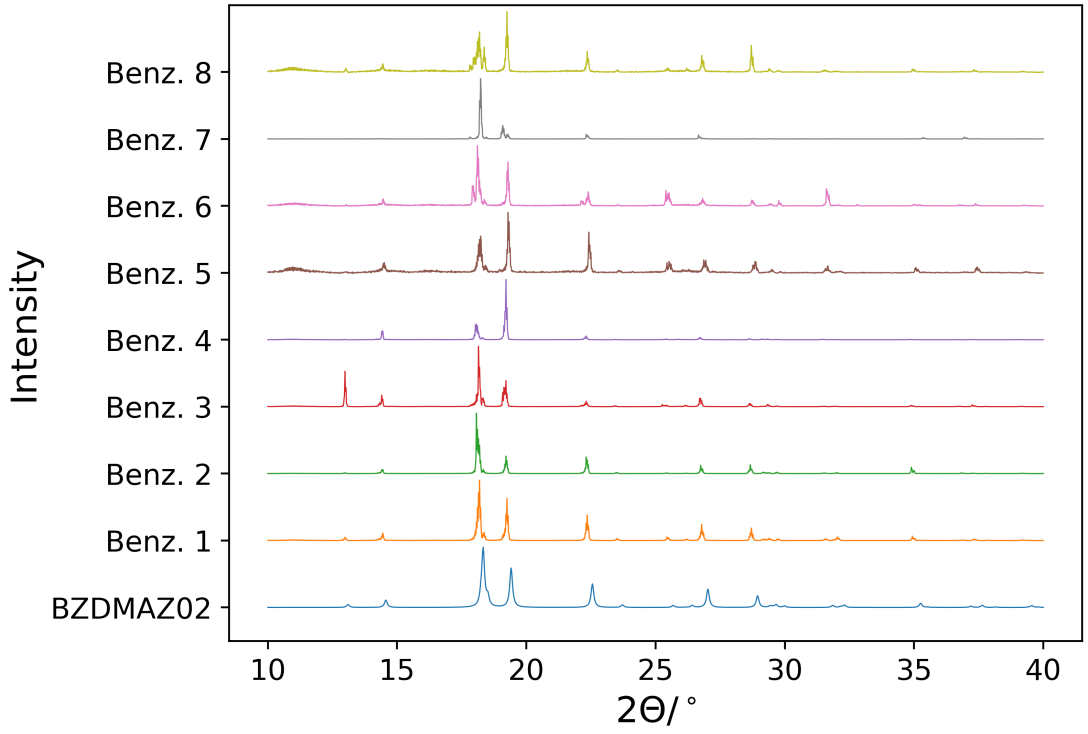



FIGURE 6.16: Experimental powder X-Ray diffraction patterns of benzimidazole. Each benzimidazole sample was synthesised through an automated process, and powder X-Ray diffraction data were collected for each sample. For comparison, the simulated powder X-Ray diffraction pattern of the alpha polymorph from BZDMAZ02 is also shown. The data suggest that the experimental powder X-Ray diffraction patterns correspond to the alpha polymorph. The characteristic peaks at approximately 13° and 24° are absent in benzimidazole 7. Peaks for other samples are present but with significantly reduced intensity.

Figure 6.16 shows that across the eight different samples of benzimidazole, there are significant differences in the PXRD patterns likely due to preferred orientation. The challenge remains that powders subject to this might be less effective when determining

crystals structures in our MCSA approach. Whilst peak positions are unaffected by this phenomena, the relative peak sizes can change substantially, resulting in inaccurate cDTW distances between simulated and experimental patterns. The experimental PXRDs have therefore been compared to the alpha polymorph in these studies.

MCSA calculations were conducted using each of the eight experimental PXRD patterns of benzimidazole to guide the search. The calculations were performed for 4000 accepted steps, employing a temperature range of 2500–100 K, and utilising the adaptive move type with $\lambda = 20$ kJ mol$^{-1}$. Results obtained using a generated PXRD pattern from the experimental structure for comparison under the same settings are also provided. These results are presented in Table 6.5.

| Sample | Experimental matches | Experimental match at global minimum? |
|---|---|---|
| BZDMAZ02 | 24 | yes |
| Benzimidazole 1 | 32 | yes |
| Benzimidazole 2 | 15 | yes |
| Benzimidazole 3 | 4 | yes |
| Benzimidazole 4 | 0 | no |
| Benzimidazole 5 | 7 | yes |
| Benzimidazole 6 | 4 | yes |
| Benzimidazole 7 | 0 | no |
| Benzimidazole 8 | 8 | yes |

TABLE 6.5: Results of Monte Carlo simulated annealing performed using 4000 accepted steps, a temperature range of 2500–100 K, and the adaptive move type. The number of experimental matches identified based on 1000 trajectories for each of the different powder x-ray diffraction patterns (PXRD)s.

In a blind context, it is assumed that structures with the lowest pseudo energy correspond to the experimental structure. For most systems, the global minimum was found to match the experimental structure. Notably, strong success was achieved using Sample 1 of benzimidazole, which yielded more experimental matches than even the simulated BZDMAZ02 pattern. Although this result is unexpected, it may not be atypical given the inherent randomness of the procedure.

Samples 4 and 7 proved ineffective for determining crystal structures, preventing the algorithm from effectively utilising PXRD data to guide the procedure.

### 6.6.2 ROY

MCSA was performed on the eight experimental PXRD patterns for the ROY ON polymorph (QAXMEH01).

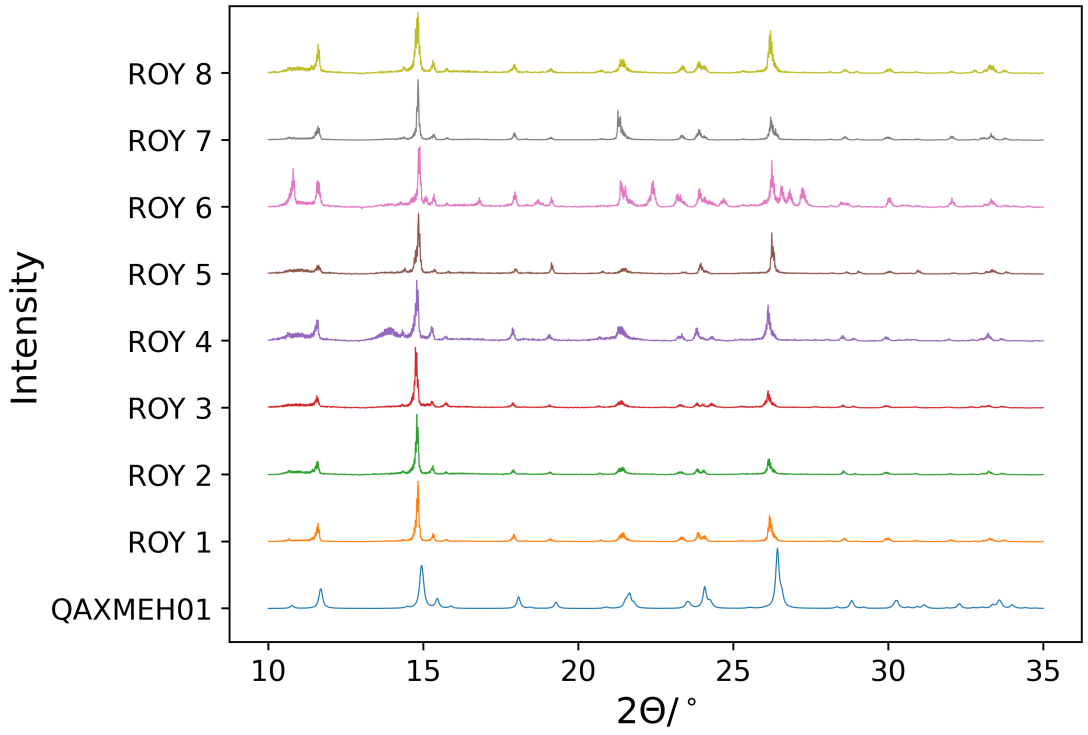



FIGURE 6.17: Experimental powder X-ray diffraction patterns (PXRD)s of ROY. Each ROY sample was synthesised through an automated process, and PXRD data were collected for each sample. For comparison, the simulated PXRD pattern of the alpha polymorph from QAXMEH01 is also shown. The data suggest that the experimental PXRD patterns correspond to QAXMEH01 within the Cambridge Structural Database. Notably, ROY 6 exhibits additional peaks, which could indicate the presence of multiple polymorphs within the sample.

Figure 6.17 shows that PXRD patterns have good resemblance to QAXMEH01 polymorph. Sample 6 of ROY possessed a significant peak at around 11°, suggesting the presence of a second polymorph QAXMEH.

MCSA calculations were conducted using each of the eight experimental PXRD patterns of ROY to guide the search. The calculations were performed for 4000 accepted steps, employing a temperature range of 2500–100 K, and utilising the adaptive move type with $\lambda = 20$ kJ mol$^{-1}$.

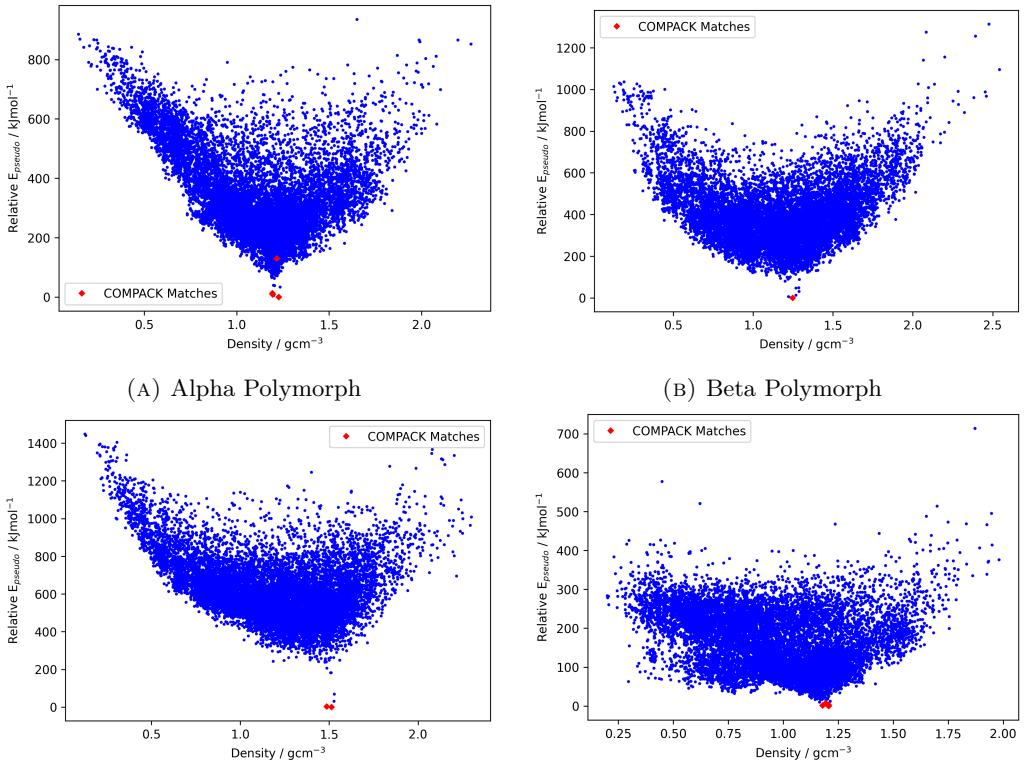
No matches to experimental structures were identified using ROY. Parameters were chosen for use with benzimidazole, a rigid structure. A new set of parameters should be calculated for more flexible molecules. This may be due to insufficient MC steps being

utilised throughout the simulation to account for the increase in the configurational space.

### 6.6.3 Blind study

The methodology was validated in a blind context by performing analyses without prior knowledge of the space group. Inclusion of up to the top 25 most common space groups was permitted, mirroring the approach used in CSP, since 99% of all structures crystallise within these groups.

An attempt was made to predict all three polymorphs of benzimidazole. In the absence of experimental data, PXRD patterns were simulated from CSD structures BZDMAZ02 ($\alpha$), BZDMAZ03 ($\beta$), and BZDMAZ07 ($\gamma$) using PLATON. The number of accepted MC steps was reduced to 500 to lower computational costs.




(A) Alpha Polymorph

(B) Beta Polymorph

(C) Gamma Polymorph

(D) Alpha Polymorph (experimental)

FIGURE 6.18: Crystal landscape for the search of benzimidazole polymorphs in a blind test using a maximum of 500 MC steps, with $\lambda = 20$ kJ mol$^{-1}$ and a final temperature of 100 K for the linear profile. Each data point represents the final structure of a single trajectory. Points in red correspond to crystal structures that match the experimentally observed structure, whilst points in blue do not. The alpha, beta, and gamma polymorphs are generated from their corresponding CSD structures, whereas the experimental alpha polymorph uses the experimental PXRD pattern of benzimidazole 1.

All of the tests managed to find matches to the experimental crystal structure at the global minimum. The experimental PXRD pattern appears to have more competing structures according to pseudo energy most likely due to the noise within the pattern compared to its simulated counterpart.

### 6.6.4 NMR

Rather than employing PXRD data to guide CSP, the methodology was tested using simulated NMR data for BZDMAZ02. The data were generated with ShiftMLv2. The equation for the cost function is as follows:

$$E_{\text{pseudo}} = E_{\text{total}} + E_{\text{NMR}}\,. \tag{6.6}$$

$E_{NMR}$ is defined to represent the difference in the $^{1}$H chemical shift values within the crystal structure.

$$E_{NMR} = \epsilon \times \sqrt{\frac{\sum_{i=1}^{n} \left(\delta_{i,trg} - \delta_{i,\text{shiftML}}\right)^2}{n}} \tag{6.7}$$

Similar to the role of $\lambda$, the parameter $\epsilon$ can be used to adjust the importance of NMR data within the procedure. A single run was performed utilising NMR data and the cost function (Equation 6.6), applying a value of $\epsilon$ equal to 10 kJ mol$^{-1}$ ppm$^{-1}$.

The methodology was applied using the simulated NMR data of MNIAAN02.

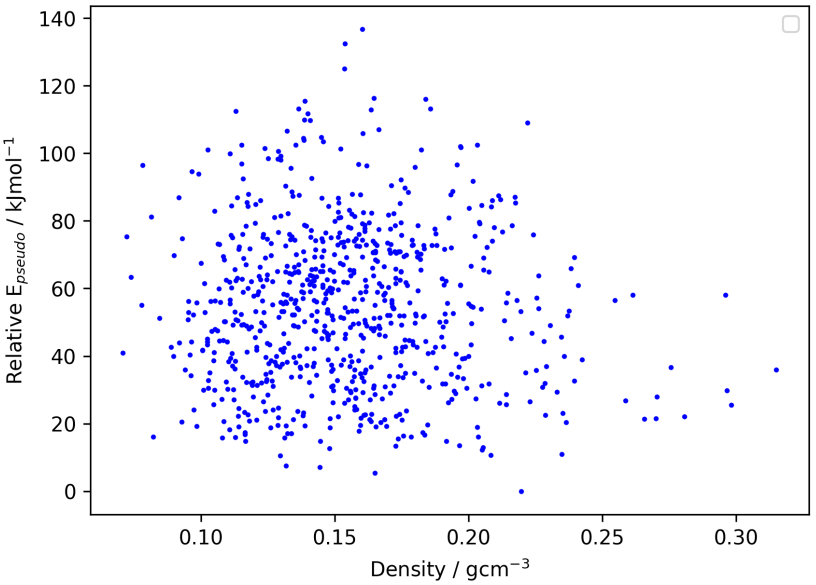



FIGURE 6.19: Crystal landscape for the search of MNIAAN02 using a maximum of 4000 Monte Carlo steps, with $\epsilon = 10$ kJ mol$^{-1}$ and a final temperature of 100 K for the linear profile. Nuclear Magnetic Resonance data was used to guide the search. Each data point represents the final structure of a single trajectory. No experimental structures were found during the search.

As seen in Figure 6.19, no matches to the experimental structure were identified, which may once again be attributed to the number of degrees of freedom within this crystal system. Consequently, further parameterisation is required.

## 6.7 MC Refinement

The use of MCSA was explored to match crystal structures from an existing CSP dataset to an experimental PXRD pattern. Instead of QR crystal structures, hypothetical structures derived from CSP methods, obtained from previous work [116, 128], were employed. The initial and final temperatures were both set to 0 K, ensuring that only MC moves reducing the pseudo energy were accepted. Eight experimental PXRD patterns were utilised for each of benzimidazole and ROY. Additionally, the impact of employing a 'perfect' PXRD pattern, simulated from the experimental crystal structures of the alpha polymorph for both benzimidazole and ROY, was investigated.

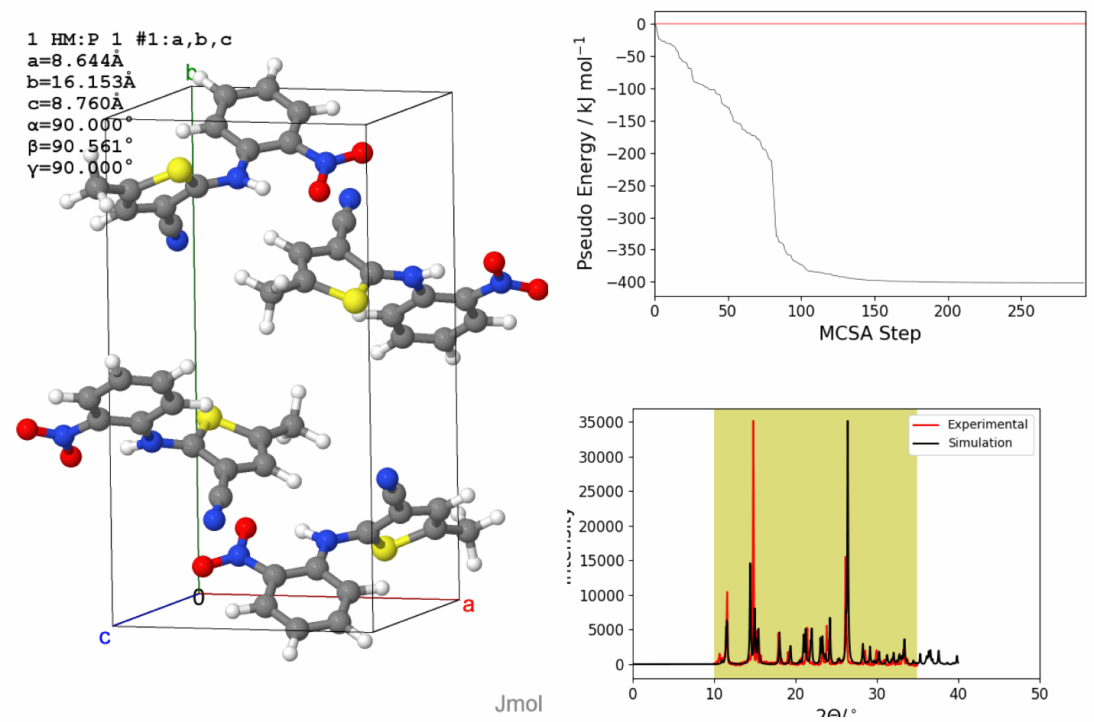



FIGURE 6.20: Screenshot monitoring the change in pseudo energy throughout the Monte Carlo refinement process for a single trajectory of ROY. After approximately half the total number of steps, the change in the pseudo energy is very small. **KEY:** Grey – carbon; white – hydrogen; red – oxygen; blue – nitrogen; yellow - nitrogen

Figure 6.20 shows that the change in pseudo energy becomes very small towards the end of the procedure.

Initial structures were anticipated to lie close to minima on the pseudo energy landscape; therefore, when employing the adaptive move type, simulations were initiated using move sizes ten times smaller than those applied in MCSA from QR structures. The procedure minimised both total energy and PXRD energy, allowing an increase in one of these energies provided a corresponding decrease occurred in the other. In this manner, the MCSA procedure leverages both low total and PXRD energies to identify potential

candidates. It is assumed that candidates matching the experimental structure exhibit both of these characteristics.

| Sample | CSP Rank | | CSP + MC Refinement Rank | | | CSP+MCSA | |
|---|---|---|---|---|---|---|---|
| No. | Total | Distance | Total | Distance | Pseudo | Distance | $\Delta E_{12}$ / kJ mol$^{-1}$ |
| BZDMAZ02 | 9 | 1 | 7 | 1 | 1 | 2.19 | |
| BZDMAZ03 | 6 | 1 | 1 | 1 | 1 | 2.23 | |
| BZDMAZ07 | - | - | - | - | - | - | - |
| Benz. 1 | 9 | 37 | 1 | 1 | 1 | 7.97 | 45.33 |
| Benz. 2 | 9 | 2 | 3 | 2 | 1 | 8.63 | 13.85 |
| Benz. 3 | 9 | 3 | 1 | 9 | 4 | 10.20 | 5.05 |
| Benz. 4 | 9 | 48 | 5 | 316 | 184 | 13.53 | 5.23 |
| Benz. 5 | 9 | 1 | 1 | 1 | 1 | 19.21 | 25.23 |
| Benz. 6 | 9 | 1 | 5 | 2 | 1 | 14.18 | 10.53 |
| Benz. 7 | 9 | 54 | 12 | 185 | 100 | 11.38 | 4.99 |
| Benz. 8 | 9 | 1 | 1 | 2 | 1 | 17.89 | 27.32 |
| ROY 1 | 1 | 47 | 127 | 2 | 2 | 14.64 | 18.97 |
| ROY 2 | 1 | 35 | 101 | 9 | 4 | 18.87 | 25.73 |
| ROY 3 | 1 | 34 | 135 | 6 | 2 | 18.81 | 37.93 |
| ROY 4 | 1 | 49 | 132 | 1 | 1 | 23.07 | 61.57 |
| ROY 5 | 1 | 46 | 112 | 3 | 4 | 16.85 | 40.37 |
| ROY 6 | 1 | 35 | 139 | 1 | 1 | 26.78 | 30.72 |
| ROY 7 | 1 | 56 | 133 | 2 | 1 | 17.51 | 6.07 |
| ROY 8 | 1 | 33 | 112 | 1 | 1 | 17.69 | 41.50 |

TABLE 6.6: Relative rankings of experimental structures at 0 K on the Z' = 1 crystal structure prediction (CSP) landscape of benzimidazole and ROY molecules. The CSP rank shows how the structure of the experimental match is ranked from global minimum before any Monte Carlo (MC) refinement in terms of total energy and constrained dynamic time warping (cDTW) distance. CSP + MC Refinement Rank shows how each structure is ranked after the procedure including pseudo energy. CSP+MCSA Distance is the cDTW distance after refinement and $\Delta E_{12}$ is the pseudo energy difference between the second-lowest and the lowest-energy structure by total energy.

The ranking of the benzimidazole experimental structure was generally poor within the initial CSP set, where it ranked ninth. For most systems, the structure was successfully ranked as the global minimum based on pseudo energy. For two structures, the ranking after MC refinement remained notably low, likely due to preferred orientation effects within the crystal, as previously discussed. It is noticed that structures with a large $\Delta E_{12}$ typically were well ranked, which could be used to provide confidence in a match. The small energy difference between the first- and second-ranked structures on the pseudo energy surface suggests that, in a blind study, these structures may not be convincingly well-matched.

The MC refinement process was also carried out using BZDMAZ07; however, no matches to this structure were found. A single structure within the CSP dataset was identified as possessing a similar packing motif to BZDMAZ07. An attempt was made to refine

the structure by applying additional pressure to encourage convergence towards the experimental form. Although greater similarity was observed in packing density, an experimental match could not be achieved.

QAXMEH01 for ROY was initially well ranked energetically within the DFT-optimised CSP dataset; however, the PXRD agreement between the experimental and CSP structures was poor. Following application of the procedure, an improvement in the overall cDTW distance was achieved, but this resulted in a substantial increase in the total energy of matching structures. Consequently, the overall ranking of structures according to $E_{\text{pseudo}}$ deteriorated significantly. Improved results could be found by performing calculations increasing the influence of $E_{\text{PXRD}}$ compared to $E_{\text{total}}$ by increasing $\lambda$.

## 6.8 Basin Hopping Approach

Thus far, the procedure for identifying crystals has involved evaluating the energy of the crystal at each step. An alternative approach is also possible, whereby, instead of using the crystal's immediate energy, the energy of its local minima is considered, such that:

$$E_{\text{pseudo}} = E'_{\text{total}} + E_{\text{PXRD}} \,, \tag{6.8}$$

where $E'_{\text{total}}$ is the sum of the intermolecular and intramolecular energies of the local minimum of the crystal structure normalised per molecule.

This approach effectively flattens the PES, as removing energy barriers between structures allows the trajectory to explore more easily. It is observed that a trajectory may still need to traverse higher-energy configurational space to reach the final structure. Therefore, the MCSA procedure should retain a temperature component to facilitate the search for global minima on the landscape. Employing larger MC move sizes could potentially reduce the computational cost associated with exploring.

To determine $E'_{\text{total}}$, the structure from each step is subjected to geometry optimisation using DFTB+.

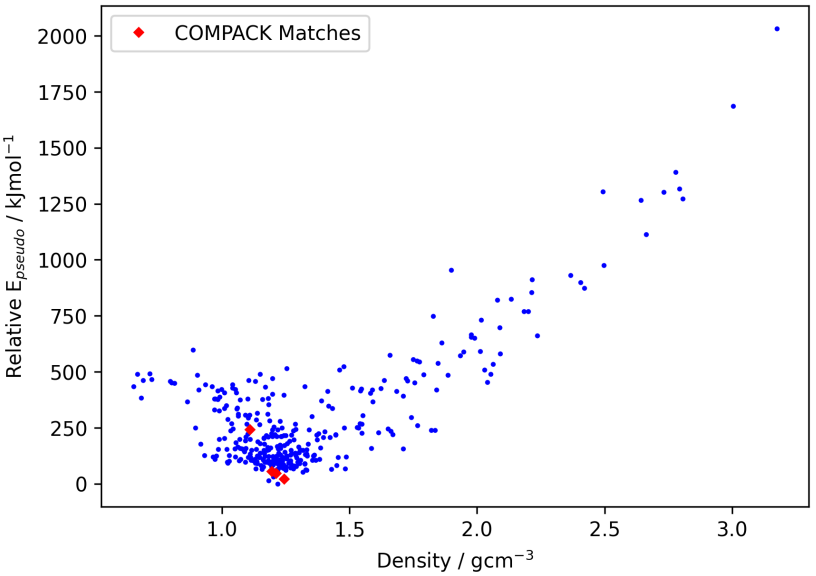



FIGURE 6.21: Crystal landscape for the Monte Carlo simulated annealing with basin hopping of BZDMAZ02 using a maximum of 4000 MC steps, with $\lambda = 10$ kJ mol$^{-1}$ and a final temperature of 100 K for the linear profile. Each data point represents the final structure of a single trajectory. No experimental structures were found during the search. Points in red correspond to crystal structures that match the experimentally observed structure, whilst points in blue do not.

Multiple structures were identified using the basin hopping approach; however, none were located at the global minimum. As energy minimisation must be carried out at every step to identify global minima, the method proves significantly more computationally demanding than MCSA. Further parameter optimisation could be explored, or an alternative energy calculation method, such as a force field, might be adopted to accelerate the process.

## 6.9 Conclusions and Future Work

Both the MCSA procedure and MC refinement procedures appear to provide good accuracy for rigid benzimidazole, although improvements in either case should be made. Efforts here should address preferred orientation and molecular flexibility.

Preferred orientation effects can be simulated using software, where different $h$, $k$, and $l$ values are considered, and the March–Dollase parameter is adjusted to model the degree of orientation. This approach would lead to a significant increase in the computational cost of calculations, especially for flexible molecular systems. Costs could potentially be reduced by investigating the Bravais-Friedel-Donnay-Harker (BFDH) morphology of crystals to predict Miller planes as seen in Figure 6.22.

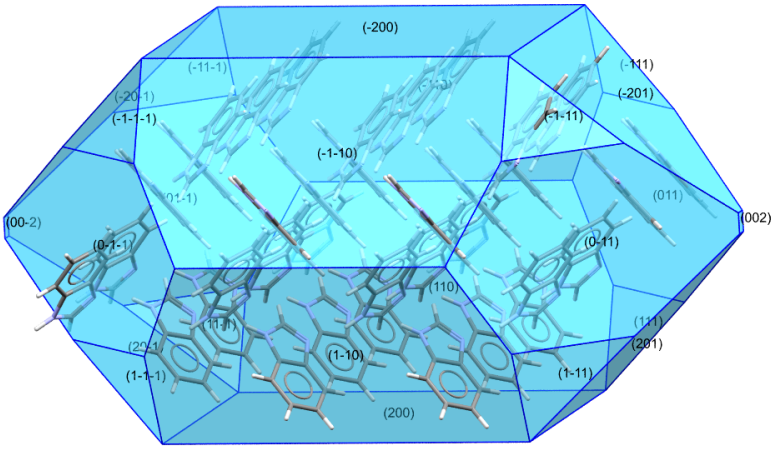



FIGURE 6.22: BFDH Morphology of BZDMAZ02. Each surface indicates its corresponding miller plane.

No matches have yet been obtained using NMR data; however, further parameter testing is warranted. Additional testing with rigid molecules should also be conducted to evaluate whether the methodology is fundamentally sound. Additionally, this could be coupled with PXRD such that the cost function would be:

$$E_{\text{pseudo}} = E_{\text{total}} + + E_{\text{PXRD}} + E_{\text{NMR}} \,. \tag{6.9}$$

This may enable the identification of experimental crystal structures, as NMR can provide insights into molecular conformation, while PXRD offers information about crystal

packing and intermolecular distances. Parameterisation of both $\lambda$ and $\epsilon$ simultaneously would need to be performed for best results.

Flexible molecules continue to present challenges when employing the MCSA procedure. Further parameter testing could be undertaken to identify a more optimal setup for such molecules. Alternatively, modifications to the manner in which the MCSA procedure is executed may prove beneficial, for instance by focusing on torsional moves. This would require analysis of the different move types utilised during the run, for instance how much does each degree of freedom change over the course of the run.

It is also possible that PXRD alone is insufficient for thoroughly exploring configurational space. For MC refinement, this limitation is less problematic, as the optimal arrangements are typically close to the starting configuration, and significant movement of the molecule within the crystal is not required.

Neither the MCSA procedure or MC refinement process when approaching CSP is superior but rather each method allows for different approaches depending on accessible data. Whilst the MC refinement procedure appears effective for both rigid and flexible molecules, it requires a CSP dataset and it fails to perturb structures sufficiently to find experimental structures that are not present on the CSP landscape. However, high pressure polymorphs can be found using MCSA without the the explicit use of high pressure in calculations.

CSP did not succeed in identifying the gamma polymorph of benzimidazole. Even after applying the MC refinement procedure, no match was obtained. However, an MCSA performed on QR structures successfully identified a match, indicating that the MCSA workflows possesses an advantage when it comes to dealing with PXRD data. However, the data here was ”ideal”, as it was simulated from the experimental crystal structure. The procedure should be run on experimental data and their success tested.

# Chapter 7

# Crystal Structure Prediction of Idelalisib and its Solvates

Idelalisib, sold under the brand name Zydelig®, is a medication used to treat chronic lymphocytic leukaemia [129, 130]. While the drug's pharmaceutical applications are significant, they are not the focus of this discussion. Instead, a currently deployed Active Pharmaceutical Ingredient (API) was utilised to demonstrate the CSP method and assess its effectiveness on larger, flexible molecules. This compound was selected due to its considerable molecular size and numerous degrees of freedom, presenting a greater challenge compared to more rigid systems. Despite its complexity, the task remains manageable, as the configurational space is not excessively large.

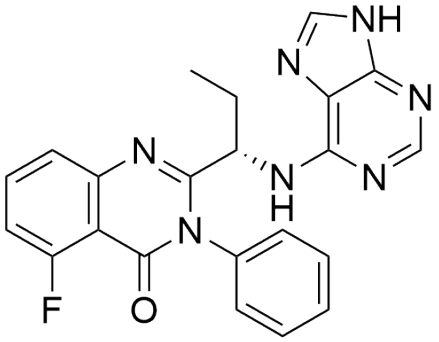
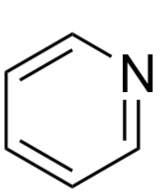


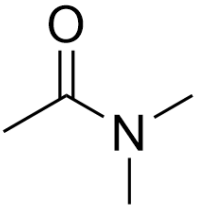

FIGURE 7.1: Idelalisib

Molecules such as idelalisib sit on the frontier of accuracy for CSP methods to date. In addition to predicting the crystal structure of neat idelalisib, the capability to predict several of its solvate forms, including those with dimethylacetamide, pyridine, and acetonitrile, was also demonstrated. These solvates present further challenges due to the requirement to explore regions of the crystal landscape where Z' > 1.




(A) Pyridine


(B) Acetonitrile




(C) Dimethylacetamide

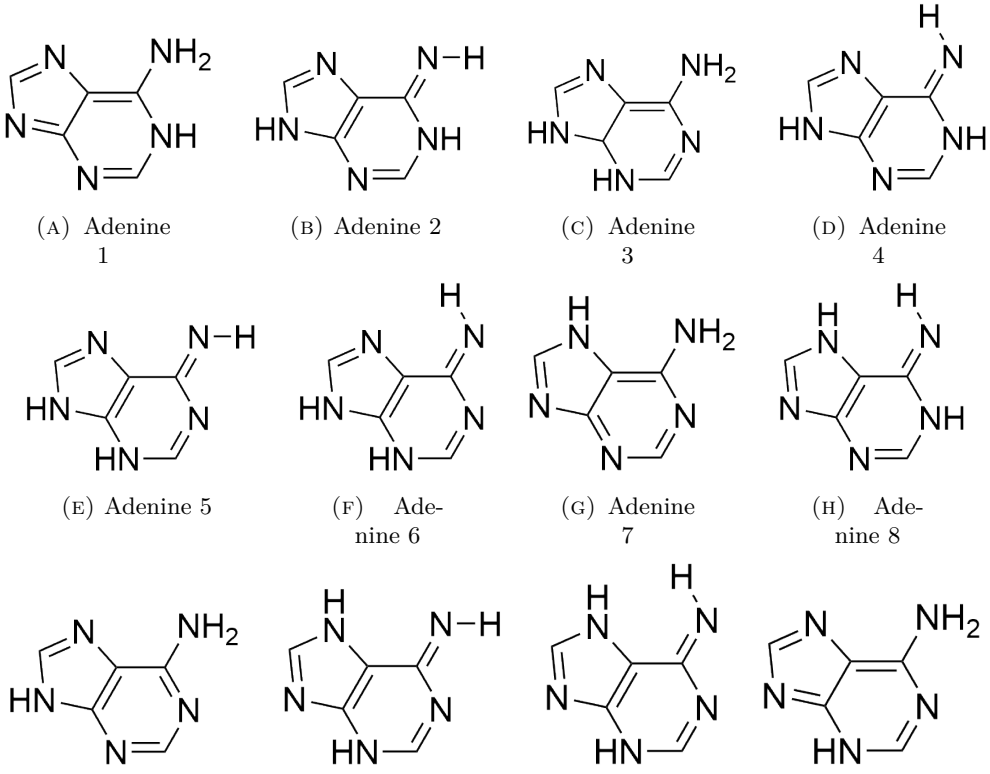
FIGURE 7.2: Solvents used to form solvate structures in the CSP of idelalisib

## 7.1 Search Details

### 7.1.1 Conformer Generation

To tackle such a molecule, a vigorous search of both conformations and tautomers must be conducted. Tautomers are two or more isomers of a compound that readily interconvert by the movement of a proton. Such a search would require conformational searches for each of the tautomeric forms of idelalisib. The combination of all tautomer types and their conformations would result in a considerably expensive study. Therefore, efforts were made to reduce computational cost in various areas without compromising accuracy by concentrating on systems more likely to be observed experimentally.

To identify observable tautomers, sites on the idelalisib molecule capable of accepting protons were investigated. Among these sites, the main areas of interest were susceptible to imine-enamine tautomerism [131]. As idelalisib contains an adenine group, various tautomeric forms of idelalisib were modelled based on those of adenine, as illustrated in Figure 7.3.



(A) Adenine 1 (B) Adenine 2 (C) Adenine 3 (D) Adenine 4

(E) Adenine 5 (F) Adenine 6 (G) Adenine 7 (H) Adenine 8

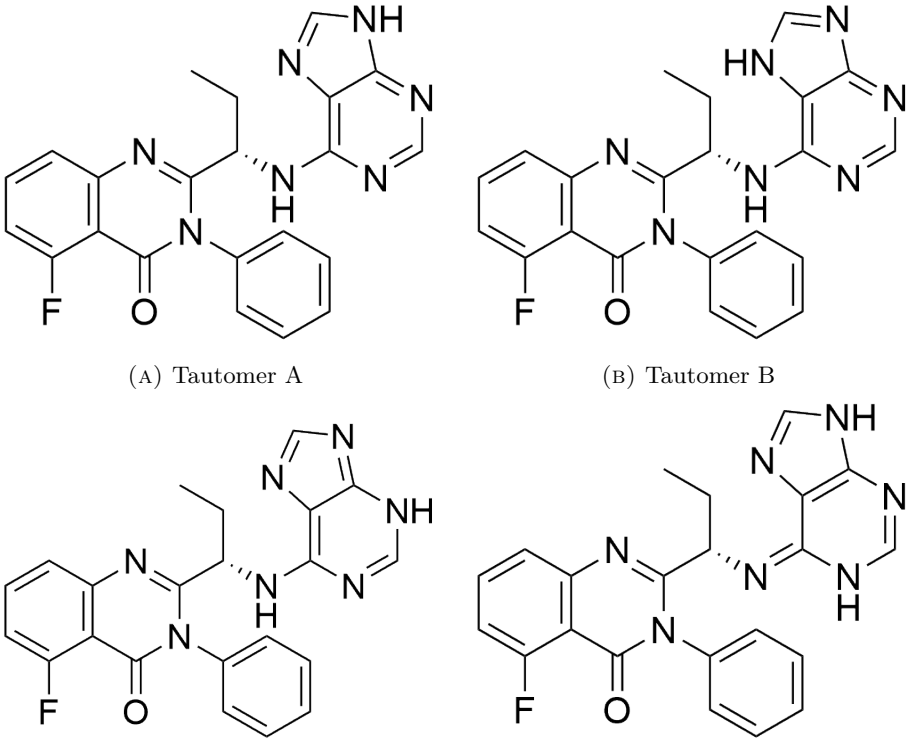
(I) Adenine 9 (J) Adenine 10 (K) Adenine 11 (L) Adenine 12

FIGURE 7.3: Tautomeric forms of adenine to be used as a proxy for the tautomers of idelalisib for crystal structure prediction.

The analysis of adenine tautomers shows that among the various forms, adenine 9 possesses the lowest energy. This suggests that adenine 9 is the preferred tautomeric form in the gaseous state. In contrast, adenine 12 and adenine 7 exhibit higher relative energies, with adenine 12 having a relative energy range of 28.5 - 29.7 kJ $mol^{-1}$, and adenine 7 ranging from 31.4 - 34.7 kJ $mol^{-1}$ depending on molecular symmetry. This indicates that while these forms are less stable than adenine 9, they are still relatively close in energy, making them possible, albeit less favourable, tautomeric candidates.

Even less stable is adenine 5, with an energy of 44.4 kJ $mol^{-1}$. This significant increase in relative energy compared to adenine 9 suggests that adenine 5 is not likely to be observed under crystallisation conditions as there will be a large contribution towards intramolecular energy. All other tautomers exhibit even higher energies, exceeding 59.4 kJ $mol^{-1}$, indicating they are significantly less stable and, therefore, even less probable in terms of their observation.

Based on this energy analysis, four tautomeric forms of idelalisib were proposed. This proposal considers not just the most energetically favourable tautomer but also other tautomers which may be able to interact and stabilise within a crystal structure.



(A) Tautomer A (B) Tautomer B

(C) Tautomer C (D) Tautomer D

FIGURE 7.4: Low energy tautomeric forms of idelalisib.

From stability analyses of adenine, only tautomers with relative energies below 25 kJ $mol^{-1}$ were included as conformations were unlikely to be adopted above this energy

threshold [24]. However, when assessing the stability of tautomeric forms for idelalisib, it is essential to consider not only the energy of each individual tautomer but also the range of conformations that each tautomer can adopt. Unlike rigid or semi-rigid molecules such as adenine, flexible molecules can adjust their torsions to minimise their conformational energy, potentially stabilising tautomers that would otherwise be energetically unfavourable. Therefore, despite initially excluding higher-energy tautomers, both idelalisib A and idelalisib B were considered as starting points for a conformational search to evaluate the relative stabilities of the idelalisib tautomers across different molecular geometries.

CREST software was utilised to perform an extensive conformational search, applying an energy limit of 30 kJ $mol^{-1}$ and using the GFN2-xTB semiempirical tight-binding method. Following the search, the resulting conformers underwent further refinement through geometry optimisation using Gaussian09. The optimisation was carried out with the 6-311G(d,p) basis set and PBE0 functional, along with GD3BJ empirical dispersion corrections. This step ensured that the conformers were optimised at a higher theoretical level, providing more accurate energy rankings on the relative stabilities of conformations.

To reduce redundancy, duplicate conformations were identified and removed using torsional clustering, which groups conformers based on their torsional angles. Where conformations were duplicates, only the lowest energy conformer from each cluster was retained for further analysis, ensuring that only unique conformations were included in subsequent evaluations. This clustering resulted in a total of 51 conformers, for which the SASA was calculated using the Shrake–Rupley algorithm with a probe radius of 1.8 Å.

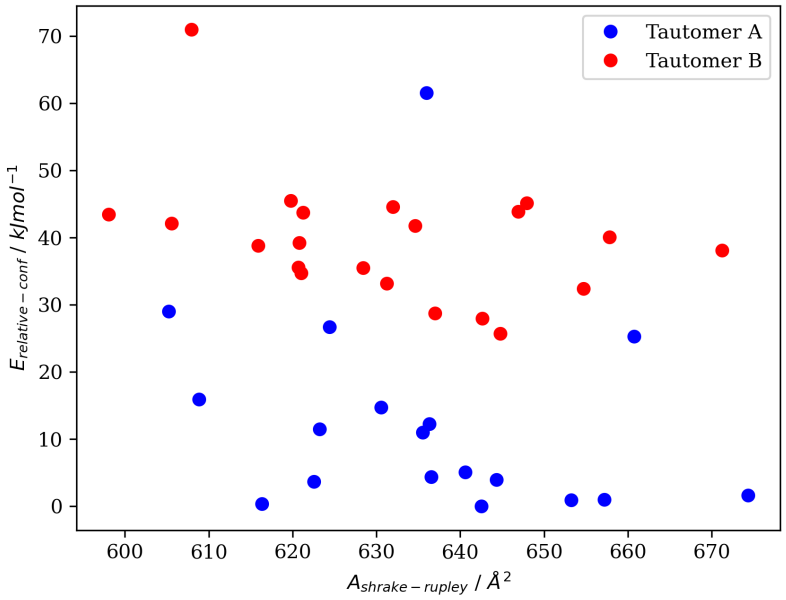
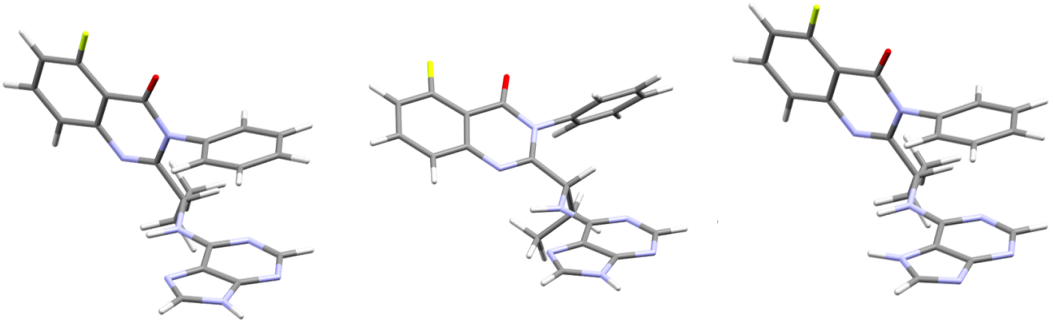



FIGURE 7.5: Gaussian geometry optimisations were performed using the PBE0 functional with the 6-311G(d,p) basis set and GD3BJ dispersion correction, starting from conformations of idelalisib obtained via a CREST search employing the GFN2-xTB method.

As anticipated, the low-energy forms of idelalisib conformers were predominantly idelalisib A. The lowest-energy conformation of idelalisib B exhibited a relative conformational energy of 23.8 kJ mol$^{-1}$ compared to the global minimum conformation, indicating that inclusion of idelalisib B in the CSP process is necessary.



(A) 0.00 (B) 1.55 (C) 38.42

FIGURE 7.6: Molecular conformers of idelalisib after optimisation with the 6-311G(d,p) basis set and PBE0 functional, along with GD3BJ. Shown are the relative conformational energies in kJ mol$^{-1}$. **KEY:** Grey – carbon; white – hydrogen; red – oxygen; yellow – fluorine.

In addition to energy, surface area can be a useful predictor for the formation of low-energy crystal structures. Conformers with larger surface areas may give rise to lower energy crystals, as they allow for more extensive intermolecular interactions within the

crystal lattice. The increased surface contact facilitates greater stabilisation through van der Waals forces, which can significantly influence the stability and formation of the crystal structure. This means that higher energy conformations could be observed if a corresponding increase in surface was sufficiently high.

#### 7.1.1.1 Surface Area Analysis of Conformers

The energy gain per unit of surface area can be modelled using sublimation enthalpies of small molecule aromatic hydrocarbons by recognising the correlation between sublimation energy against molecular surface area [24]. This relationship suggests that the larger the surface area, the greater the intermolecular energies and can indicate which conformations should be prioritised for CSP.

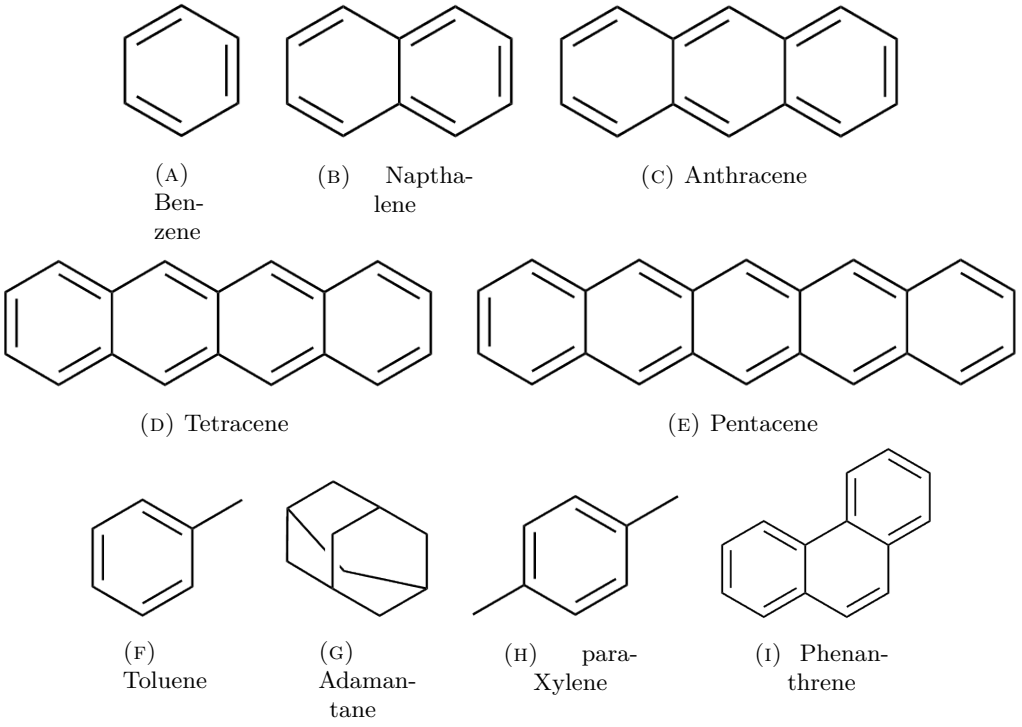



FIGURE 7.7: Molecular structures of small molecule aromatics used to calculate the energy gain per unit of surface area.

The molecular surface areas of molecules, benzene [132–134], napthalene [133, 135–147], anthracene [137, 144, 147–155], tetracene [140, 156–159], pentacene[140, 159], toluene [160], adamantane [161–167], para-Xylene [168], phenanthrene [135, 137, 147] were calculated using Shrake-Rupley method with a probe radius of 1.8 Å. Sublimation enthalpies for each of these molecules was obtained and used as a proxy for the energy stabilisation provided through an increase in surface area.

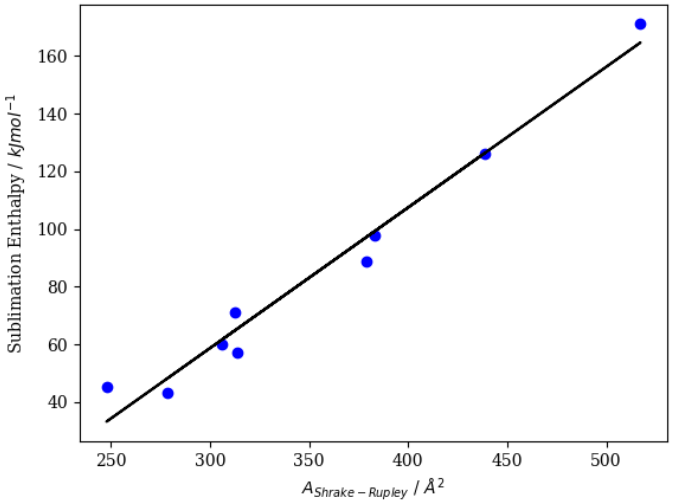



FIGURE 7.8: Variation in measured $H_{\text{sublimation}}$ with molecular $A_{Shrake-Rupley}$ for a set of small rigid hydrocarbon crystal structures. In black is the trend line after least squares regression analysis on the data.

Least squares regression analysis was performed showing the relationship between sublimation enthalpies and surface area shown in Figure 7.8. It was found that there was 0.4876 kJ mol$^{-1}$/Å$^2$ increase in sublimation enthalpy per unit surface area. The gradient allows us to consider the stabilisation energy, providing us with a metric to rank conformations based on surface area and conformational energy shown in Equation 7.1.

$$E_{\text{BIAS}} = E_{\text{CONF}} - E_{\text{SA}} \tag{7.1}$$

where $E_{\text{BIAS}}$ is the biased energy, $E_{\text{CONF}}$ is the conformational energy and $E_{\text{SA}} = 0.4876$ kJ mol$^{-1}$/Å$^2$ which corresponds to the energy stabilisation due to the increase in surface area.

Using $E_{\text{BIAS}}$, it is possible to identify structures which might be good candidates for CSP. Structures that possess higher conformational energy yet exhibit an $E_{\text{BIAS}}$ lower than that of the global minimum conformation can be highlighted by plotting an $E_{\text{BIAS}}$ line relative to the surface area of the global minimum conformation as shown in Figure 7.9.

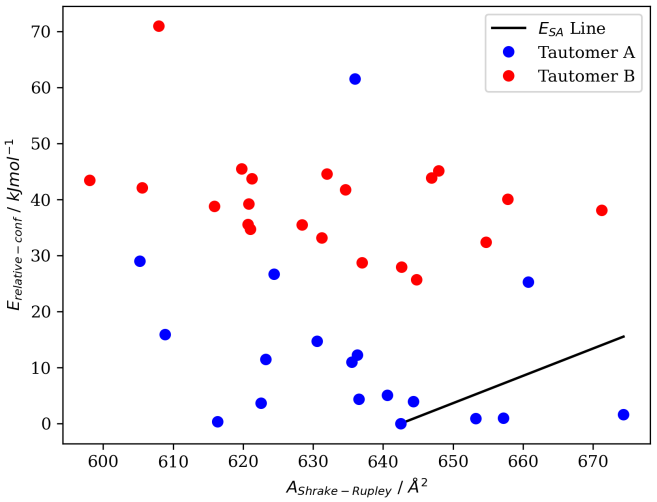



FIGURE 7.9: Gaussian geometry optimised conformers of idelalisib using the 6-311G(d,p) basis set and PBE0 functional, along with GD3BJ dispersion correction obtained from a CREST search employing the GFN2-xTB method. The $E_{\mathrm{SA}}$ line is plotted, indicating that conformers located to the right of this line possess a biased energy lower than that of the global minimum when accounting for the surface area of each conformer.

Conformers below of the $E_{\mathrm{BIAS}}$ line are predicted to have their higher energy compensated by an increase in surface area forming greater intermolecular interaction within the crystal. It is shown however that few conformations exist to the right of this line.

### 7.1.2 Crystal Structure Prediction Sampling

To conduct an effective CSP, only common space groups from the CSD [28] were considered. Owing to the chirality of the molecule, the selection was further restricted to Sohncke space groups, which preserve molecular chirality and lack glide planes or inversion centres. This process was carried out using in-house CLG.

| Space group | Number of valid structures |
|---|---|
| P $2_1$ $2_1$ $2_1$ | 50000 |
| P 1 $2_1$ 1 | 40000 |
| C 1 2 1 | 20000 |
| P 1 | 20000 |
| P $2_1$ $2_1$ 2 | 15000 |
| P $4_1$ | 10000 |
| P $4_3$ $2_1$ 2 | 10000 |
| P $4_1$ $2_1$ 2 | 10000 |
| P $4_3$ | 10000 |
| P $3_2$ | 10000 |

TABLE 7.1: Number of valid crystal structures generated using the crystal landscape generator for neat idelalisib. The number of crystal structures correlates with how frequently each space group is observed within the crystallographic structural database.

The extent of sampling conducted within each space group is proportional to the frequency with which that space group is observed in the CSD. The highest level of sampling is performed in the most common space groups, while less sampling is undertaken in less frequently occurring groups.

## 7.2 Idelalisib

CSP for neat idelalisib was performed. The workflow for this process is illustrated in Figure 7.10.

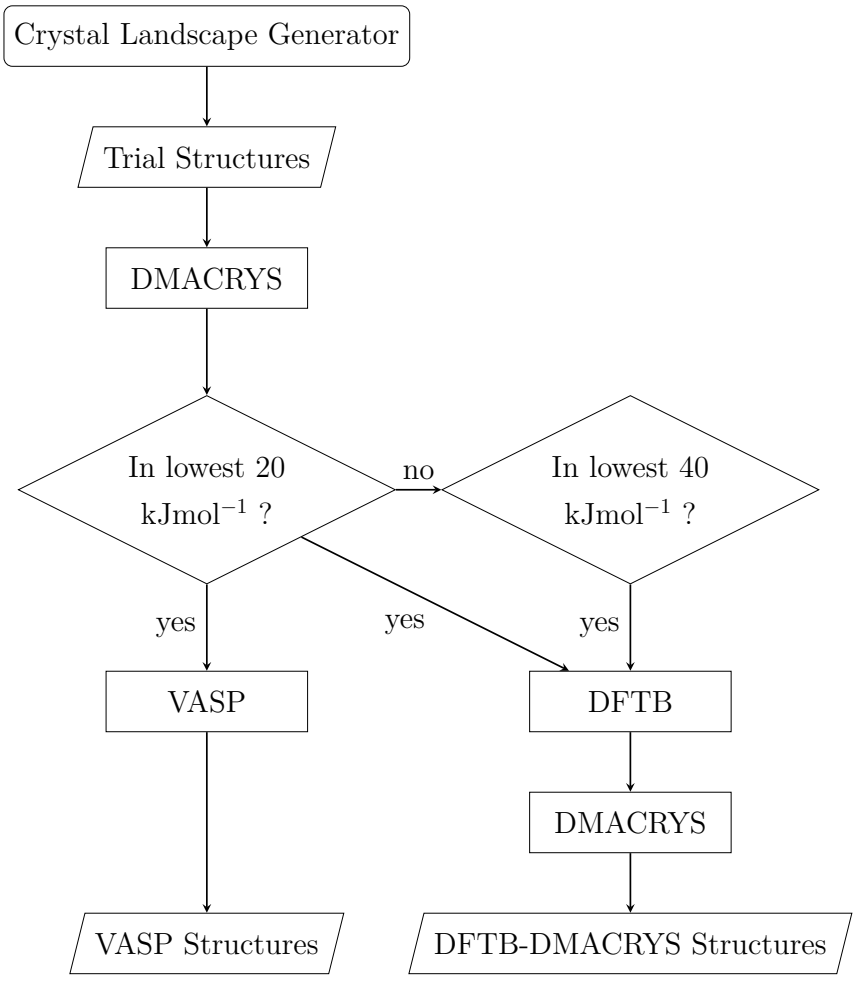



FIGURE 7.10: Workflow for crystal structure prediction of neat idelalisib

CSP was conducted on each conformer using the CLG, which packs rigid conformations into crystal structures. Only packing arrangements resulting in crystal structures with Z' = 1 were considered. These structures were minimised using DMACRYS. A multipole force field was used which did not allow for movement of molecular geometries keeping conformations fixed.

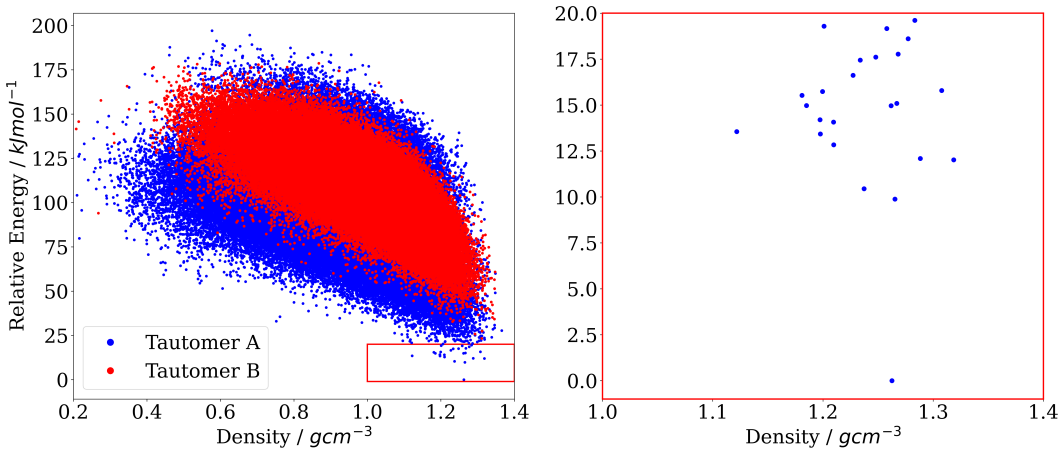



FIGURE 7.11: Crystal landscape of idelalisib after crystal structure prediction. Shown in red is the magnified low energy region of the crystal landscape corresponding to 20 kJ mol$^{-1}$ above the global minimum.

Around 220,000 unique crystal structures were obtained ranging over 200 kJ mol$^{-1}$ above the global minimum crystal structure. The low energy region is dominated by conformers of tautomer A, suggesting that the energy gain of adopting tautomer B was not compensated for by inter-molecular interactions. 4745 structures were found within 50 kJ mol$^{-1}$ of the global energy minimum in which only 188 contained tautomer B.

### 7.2.1 Post Crystal Structure Prediction Optimisation

To construct a comprehensive crystal energy landscape, it is essential to allow the molecular geometries within the unit cell to relax. Two strategies for performing this relaxation have been investigated. One approach employs periodic DFT, which can yield highly accurate crystal structures. However, its significant computational cost makes it impractical to optimise every generated structure. Therefore, only a limited subset of structures from the low-energy region of the landscape is selected for periodic DFT refinement.

Alternatively, a more computationally efficient method uses semi-empirical DFT techniques. These methods help reduce computational demands while still providing reasonable predictions of relative energies. Structures optimised using this approach are then further refined with a multipole force field to improve the accuracy of their energy evaluations.

When examining the size of the lowest energy lid, the number of structures follows a sigmoid distribution shown in Figure 7.12.

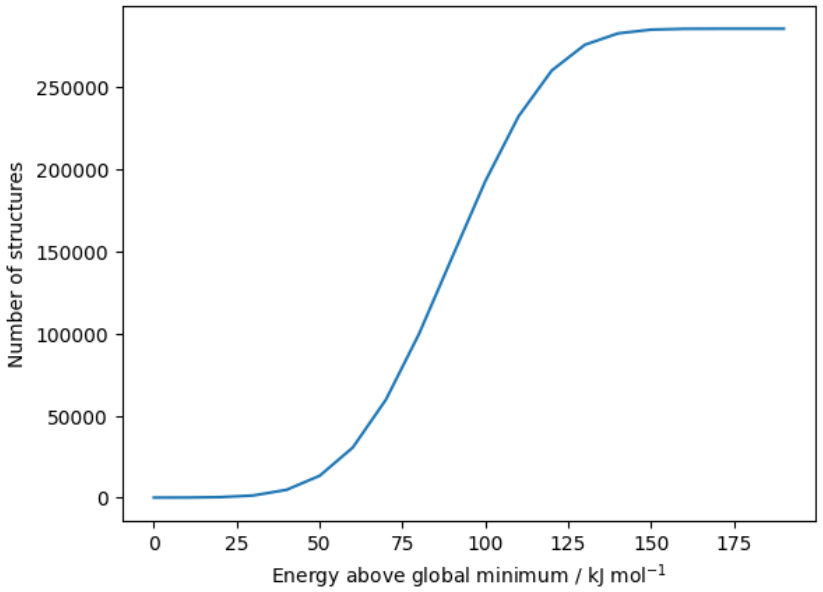



FIGURE 7.12: The cumulative number of structures within select energy windows above the global minimum structure for the generated crystal landscape of idelalisib.

Therefore, to reduce computational cost, only the lower-energy half of the sigmoid curve is likely to be investigated. During post-CSP optimisation, this consideration implies that only narrow energy windows may be selected for re-optimisation using alternative methods.

#### 7.2.1.1 Periodic Density Functional Theory Optimisation

A total of 38 structures were identified within 20 kJ mol$^{-1}$ of the global minimum and subsequently reduced to 24 structures following duplicate removal using COMPACK. VASP minimisation was performed to relax the molecular geometries within each crystal structure. The resulting landscape is illustrated in Figure 7.13.

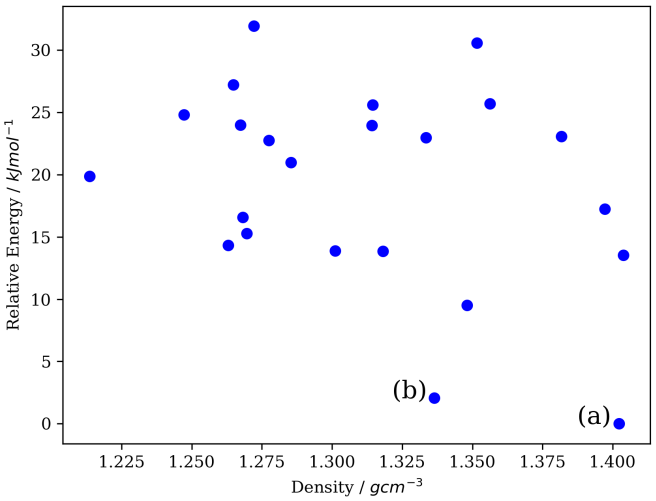
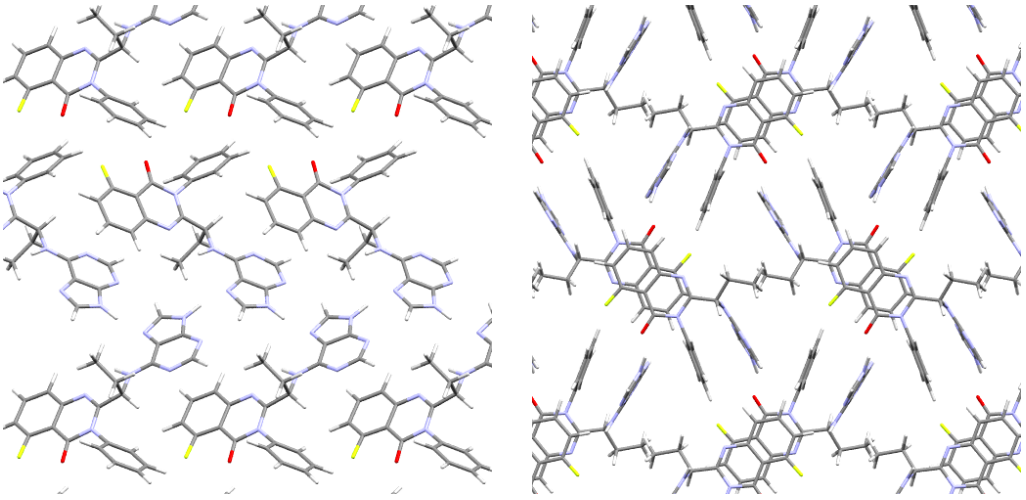



FIGURE 7.13: Crystal landscape for neat idelalisib after VASP optimisation of the lowest 20 kJ mol$^{-1}$. The two lowest energy structures are labelled (a) and (b).

Two crystal structures were identified within 5 kJ mol$^{-1}$ of the global minimum. These structures underwent significant re-ranking, with their new energies rising to over 32.1 kJ mol$^{-1}$. One structure failed to converge during the minimisation process and was consequently excluded from the dataset.



(A) 0.00 (B) 2.06

FIGURE 7.14: Lowest energy VASP optimised crystal structures of idelalisib. Shown are $E_{\text{relative}}$ values for each structure. (a) 0.00, b) 2.06 kJ mol$^{-1}$. **KEY:** Grey – carbon; white – hydrogen; red – oxygen; yellow – fluorine.

#### 7.2.1.2 DFTB+ and DMACRYS

A total of 1056 crystal structures were identified within 40 kJ mol$^{-1}$ of the global minimum and were initially optimised using DFTB+. This optimisation step yields

reasonable and well-defined molecular geometries; however, it often produces inaccurate energy estimates. In particular, DFTB tends to overestimate electron repulsion leading to crystal structures with excessively high calculated densities [169]. To improve both the energy rankings and the predicted densities, further relaxation of the crystal geometries was carried out using DMACRYS. Because DMACRYS is computationally less demanding than VASP, it allows a larger number of crystal structures to be examined.

Since geometry optimisation using DFTB+ permits relaxation of internal molecular coordinates, it is necessary to calculate the new intramolecular contribution of geometries when progressing to DMACRYS optimisation, in order to determine the $E_{\text{total}}$ of the crystal lattice (Equation 2.1). Therefore, single-point calculations were performed on the molecules within the asymmetric unit using Gaussian, employing the 6-311G(d,p)/PBE0 basis set with the GD3BJ dispersion correction.

It was proposed that a standard error exists across all molecular geometries resulting from performing DFT single-point energy calculations on DFTB geometries. Analysis of $E_{\text{total}}$ incorporating the new intramolecular energies led to substantial re-ranking of the structures. It was observed that molecular conformations which were not significantly different still exhibited energies tens of kJ mol$^{-1}$ higher than anticipated.

Part of the intramolecular contribution was accounted for by applying a Polarisable Continuum Model (PCM) to each conformation during the single-point calculation. A dielectric constant of 3.0 was employed for this purpose [170], thus leading to our corrected final energy landscape.

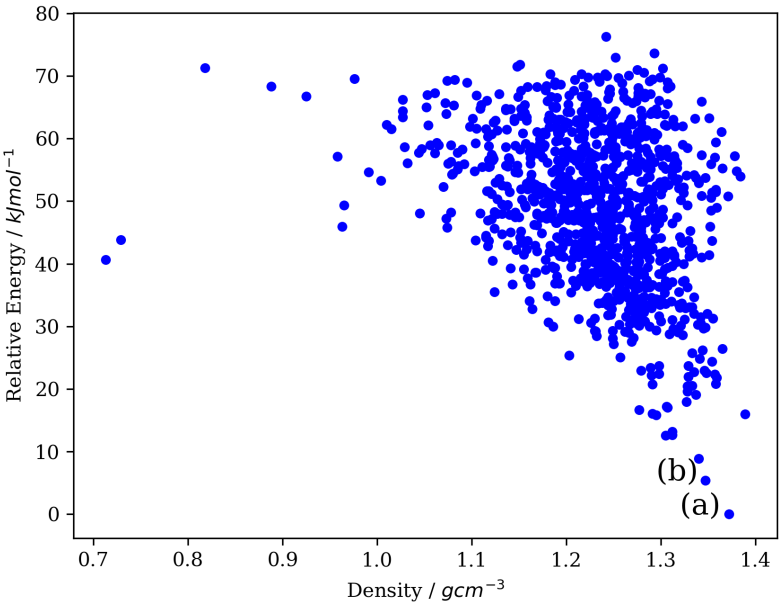
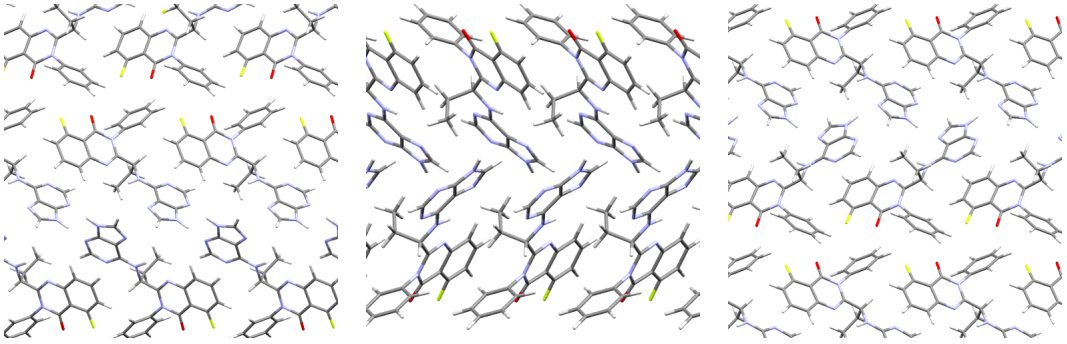



FIGURE 7.15: Crystal Landscape of neat idelalisib after lowest 40 kJ mol$^{-1}$ have been optimised using DFTB+ and DMACRYS. Only structures optimised are shown.

Three crystal structures were found within 10 kJ $mol^{-1}$ of the global minimum.



(A) 0.00 (B) 5.40 (C) 8.84

FIGURE 7.16: Lowest energy DFTB-DMACRYS optimised crystal structures of the idelalisib crystal landscape. Shown are $E_{relative}$ values for each structure in kJ$mol^{-1}$. **KEY:** Grey – carbon; white – hydrogen; red – oxygen; yellow – fluorine.

### 7.2.2 Comparison of Workflows

Two distinct workflows for the CSP of neat idelalisib have been presented.

#### Cost and Accuracy

For a single VASP optimisation, the average total compute time was 112 CPU hours per structure. Due to this high computational expense, only 24 crystal structures were examined. Consequently, if DMACRYS provides poor energy rankings, important low-energy structures might be overlooked.

In comparison, the combined DFTB+ and DMACRYS calculations were significantly less demanding, requiring an average of only 3.8 CPU hours per structure. This means that evaluating 43 times more crystal structures was only slightly more costly overall than performing the limited set of VASP calculations.

A comparison of the two landscapes was performed to identify which structures were identified by either method.

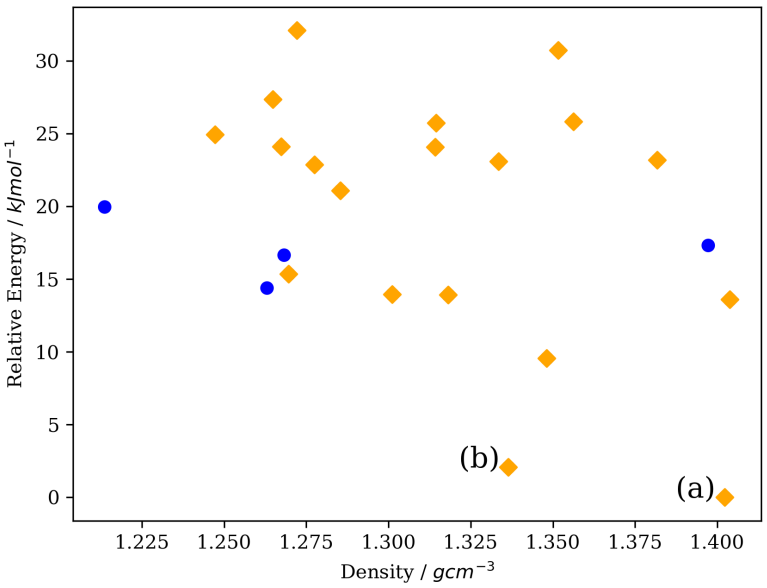



FIGURE 7.17: Structures found by DFTB+ and DMACRYS workflow on the VASP crystal energy landscape. Structures indicated by a diamond are structures which appeared on both VASP and DFTB-DMACRYS landscapes, whereas blue dots indicate structures which did not match.

The structures that were not identified by the DFTB-DMACRYS workflow appear to have relatively high energies, exceeding 13 kJ mol$^{-1}$ above the global minimum. Due to their elevated energy levels, these structures are unlikely to be observed, though they should not be entirely disregarded.

Of the 23 structures generated by the VASP workflow, 19 were matched to those found in the DFTB-DMACRYS landscape. Notably, the workflow successfully identified low-energy crystal structures, while DMACRYS sampled a broader range of configurational space. Both methods identified the same crystal structure as the global minimum which indicates some consistency across both landscapes.

To validate these findings, the computed PXRD patterns for structures a, b, and c from both the VASP and DFTB-DMACRYS workflows were compared to experimental PXRD data of neat idelalisib. However, no convincing matches were observed across these crystal structures, suggesting discrepancies between the theoretical models and experimental results.

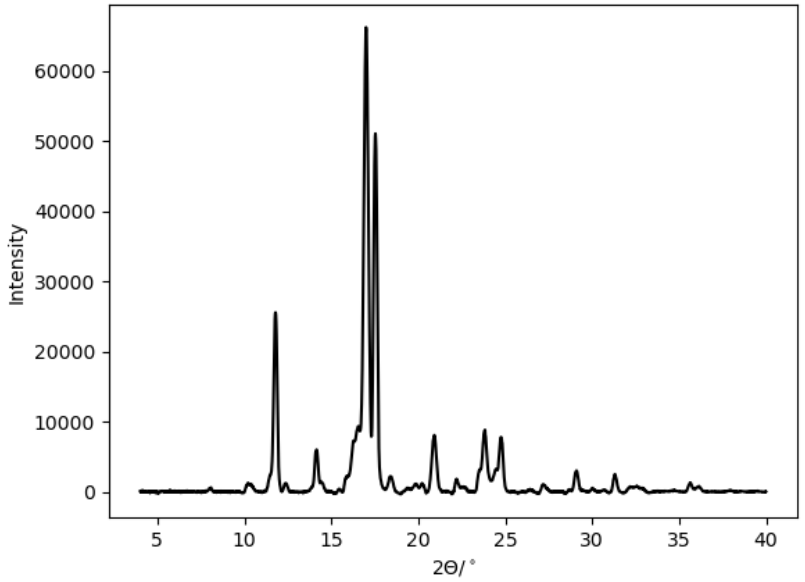



FIGURE 7.18: Powder X-ray diffraction pattern of the experimentally observed polymorph of neat idelalisib.

All generated crystal structures were compared to the experimental crystal structure using cDTW. Distances were evaluated across a range of band-warping limits to identify structures that most closely match the experimental PXRD pattern.

The better rank indicates a shorter cDTW distance. A good candidate will be observed to have a low rank across all band-warping limits. It is therefore possible to eliminate poor candidates which do not reach a low rank. Candidate crystal structures that failed to achieve a rank of at least 50 across all bandwidths were excluded. Good candidates could be optimised further at DFT level to produce better matches to experimental structures.

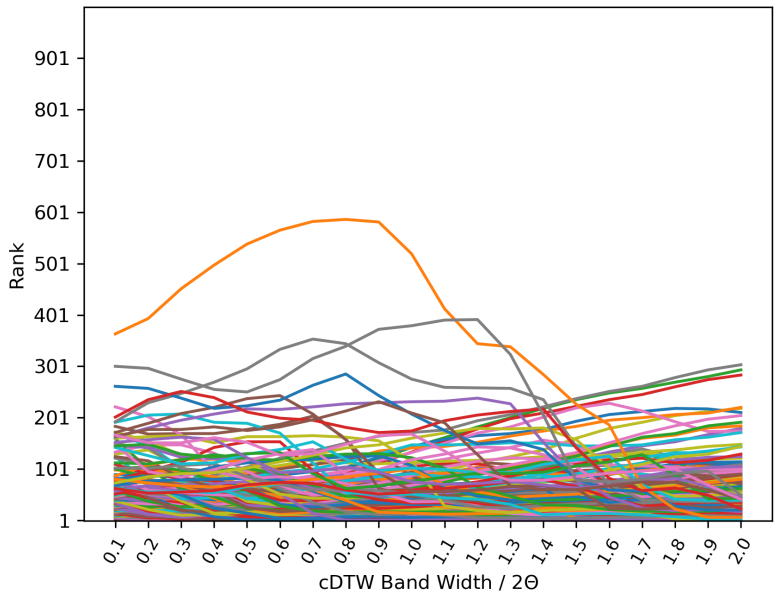



FIGURE 7.19: Ranking of computed powder X-ray diffraction patterns to experimental crystal structure across a range of bandwidths.

A visual comparison of the generated PXRD patterns was conducted. None of the structures appeared to provide a convincing match to experimental data.

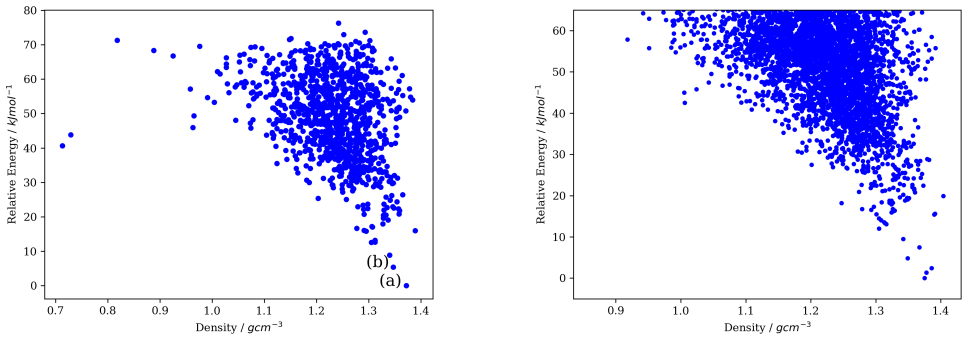
### 7.2.3 Extended Sampling

The sampling of the CSP was extended as detailed in Table 7.2.

| Space group | Number of valid structures |
|---|---|
| P $2_1$ $2_1$ $2_1$ | 100000 |
| P 1 $2_1$ 1 | 80000 |
| C 1 2 1 | 40000 |
| P 1 | 40000 |
| P $2_1$ $2_1$ 2 | 30000 |
| P $4_1$ | 20000 |
| P $4_3$ $2_1$ 2 | 20000 |
| P $4_1$ $2_1$ 2 | 20000 |
| P $4_3$ | 20000 |
| P $3_2$ | 20000 |

TABLE 7.2: Number of valid crystal structures generated using the crystal landscape generator for neat idelalisib in the extended sampling. The number of crystal structures correlates with how frequently each space group is observed within the crystallographic structural database.

It was found that increasing the sampling in the CSP allows for the identification of additional crystal structures within the low-energy region. Therefore, the energy window for post-CSP optimisation was expanded to include structures up to 50 kJ mol$^{-1}$ above the global minimum. As shown in Figure 7.20.




(A) Initial Sampling (B) Extended Sampling

FIGURE 7.20: Crystal Landscape of neat Idelalisib after optimised using DFTB+ and DMACRYS with increased sampling. For initial sampling crystals that were up to 40 kJ mol$^{-1}$ above the global minimum were taken from the initial crystal landscape. For the extended sampling 50 kJ mol$^{-1}$ was taken. Only structures optimised are shown.

Many more structures were subjected to optimisation, revealing a substantially larger number of low-energy structures in the low-energy region than had been previously identified. The PXRD patterns for each of these structures were simulated using PLATON

and visually compared with the experimental diffraction pattern.

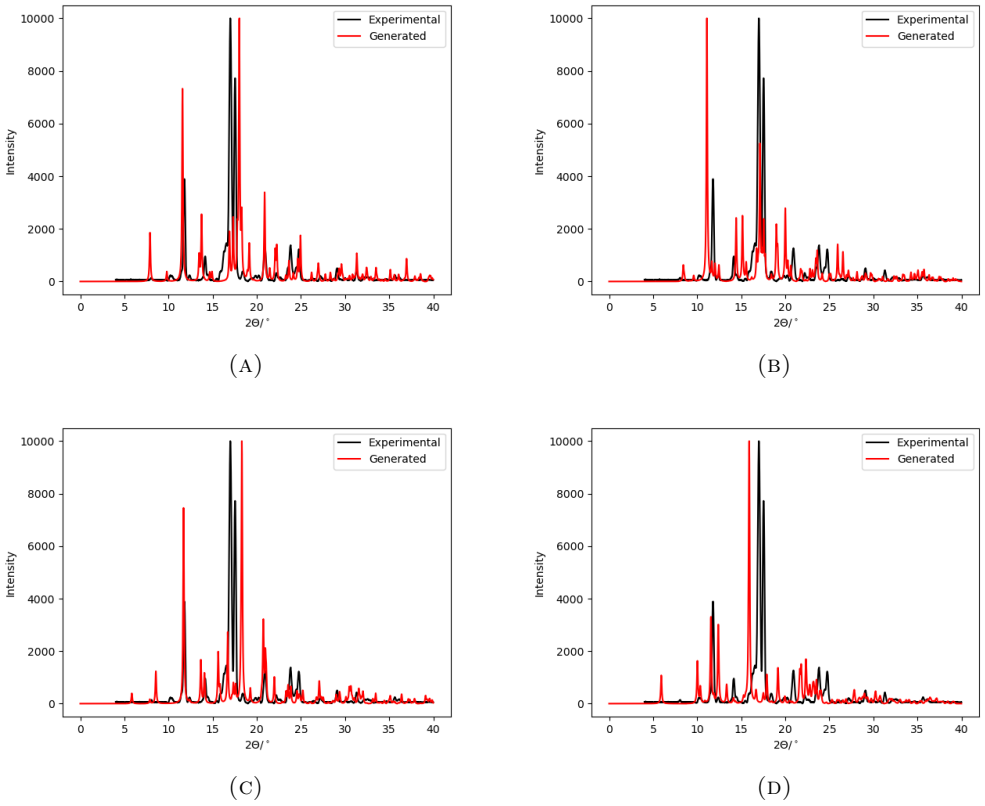



FIGURE 7.21: Overlay of powder X-ray diffraction patterns against the experimental pattern. Patterns shown show the closest resemblance across the dataset after re-optimisation with DFTB+ and DMACRYS for idelalisib.

Figure 7.21 shows that some PXRD patterns exhibit similarities to the experimental crystal structure PXRD; however, there remains uncertainty as to whether the experimental structure has been identified at this stage. Optimisation of promising candidates using VASP may yield improved matches. Additionally, the new flexible CSP approach described in Chapter 8 will also be explored.

### 7.2.4 Structure Re-ranking

For future CSP efforts, determining the appropriate energy range for further optimisation is essential. Here, the energies of structures before and after subsequent optimisation were examined. In systems exhibiting significant re-ranking, a wider energy window is required to ensure that no potentially low-energy structures are overlooked.

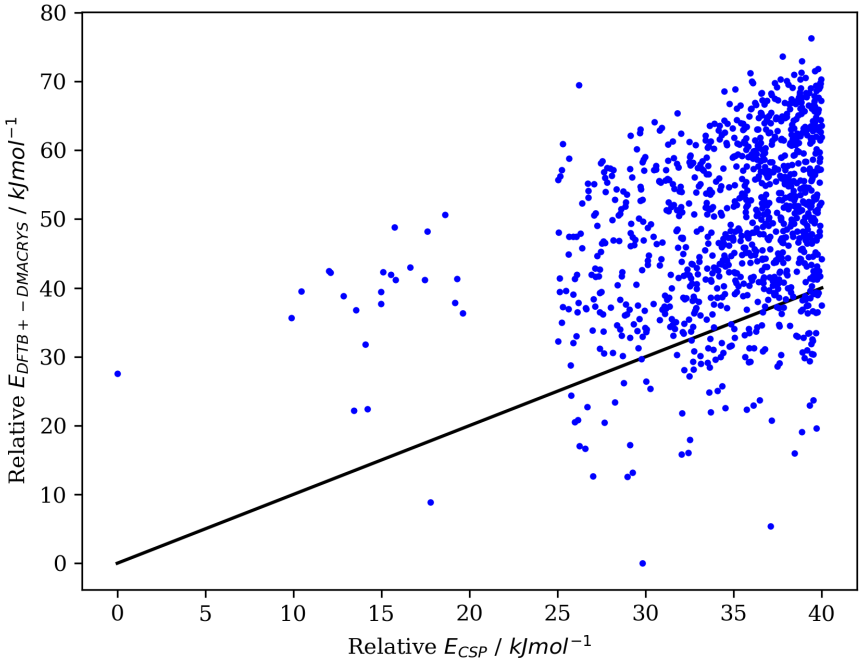



FIGURE 7.22: Re-ranking of structures relative to the global minimum energy for before and after DFTB+ and DMACRYS re-optimisation for neat idelalisib. The black line shows where Relative $E_{DFTB-DMACRYS}$ = Relative $E_{CSP}$. Structures below this line are structures which have decreased in energy relative to the global minimum and structures above have increased in energy.

There is a significant amount of re-ranking before and after DFTB+ and DMACRYS re-optimisation. The global minimum energy structure was ranked around 30 kJ mol$^{-1}$ above the global minimum energy structure before re-optimisation. For these flexible molecules, relaxation of molecular geometry can lead to significant changes in $E_{total}$. An energy window at least 30 kJ mol$^{-1}$ should be used to account for these changes.

For the post-CSP optimisation using VASP, the global minimum identified by VASP was located within 20 kJ mol$^{-1}$ of the CSP global minimum. This indicates that different initial structures converged to the same final predicted crystal structure during optimisation.

One major source of structure re-ranking lies within the energy surface used to describe the intramolecular energies within each crystal structure. Energies are calculated at DFT level whilst the conformations within the crystal use a DFTB+ level of theory. The discrepancy between these two energy surfaces can lead to artificial strain energy being added to the system. Here, it is assumed that the energy gain resulting from this effect will be balanced, as the gain should be similar across all systems. However, once intramolecular interactions were considered, poorer ranking was observed.

### 7.2.5 Using Experimental Data

*This section was in collaboration with Robert Caroll and Edd Bilbe. Caroll performed the growth of the single crystal and performed SCXRD on the idelalisib crystal. Bilbe performed PXRD on a sample of powdered idelalisib.*

In light of insufficient data to confidently determine the crystal structure of idelalisib, a sample was purchased from SelleckChem [171]. On this sample, PXRD and SCXRD was performed.

#### PXRD

PXRD data had already been obtained from a sample of anhydrous idelalisib, as discussed previously. Given that idelalisib is known to possess at least two polymorphs, it was essential to confirm that the data corresponded to the same crystal structure.

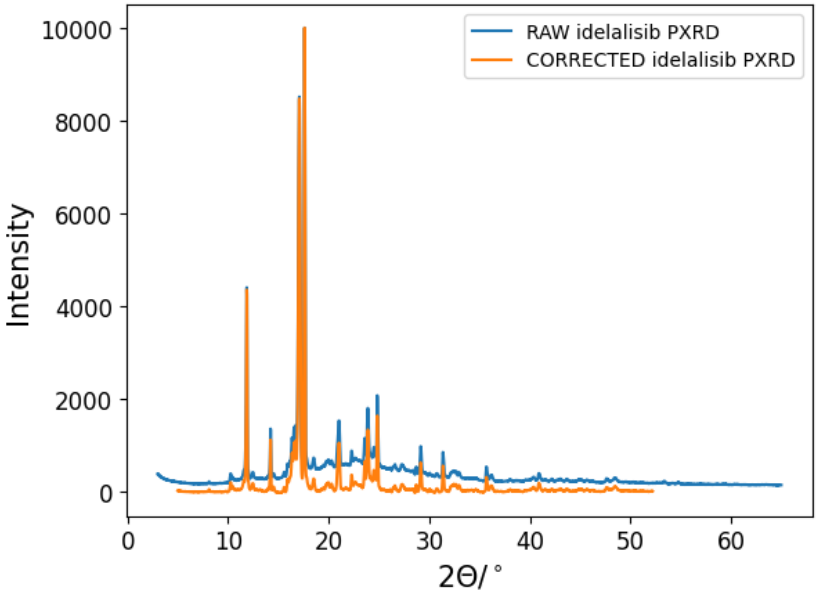



FIGURE 7.23: Comparison of powder X-ray diffraction patterns before and after background correction for powdered idelalisib.

The PXRD of the store-bought product was measured. Background correction was performed, and the data was truncated at 5° using DASH to eliminate noise and the shallow peak observed near 3° [28]. A comparison between corrected and non-corrected PXRDs is shown in Figure 7.24.

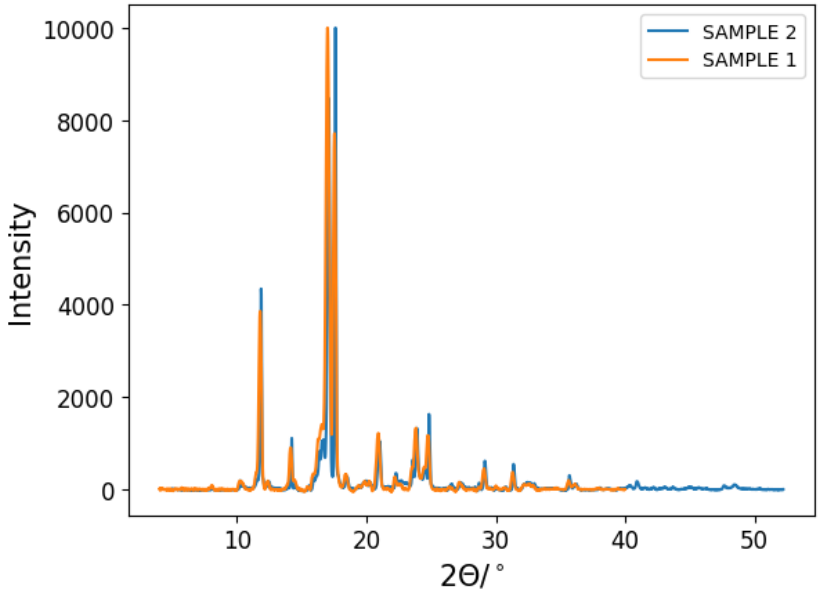



FIGURE 7.24: Powder X-Ray diffraction patterns of two different samples of idelalisib obtained from SelleckChem and communication from Johnson Matthey [172].

Good agreement was observed between the PXRD patterns providing confidence that the structures are identical.

### SC-XRD

Attempts were made to grow single crystals using five solvents: dichloromethane (DCM), tetrahydrofuran (THF), acetonitrile, chloroform, and nitrobenzene. Single crystals of idelalisib were successfully obtained only through slow evaporation from a saturated DCM solution. A suitable crystal with block morphology was selected and mounted on the diffractometer for SCXRD data collection. The experiment was performed on a Rigaku 007HF diffractometer using Cu-K$\alpha$ radiation ($\lambda$ = 1.54184 Å), equipped with Varimax confocal mirrors, a UG2 goniometer, and HyPix Arc-100 detectors. The crystal was maintained at 100.00(10) K during data collection. The structure was solved with SHELXT [173] using Intrinsic Phasing and refined with SHELXL [174] via Least Squares minimisation, employing Olex2 [175].

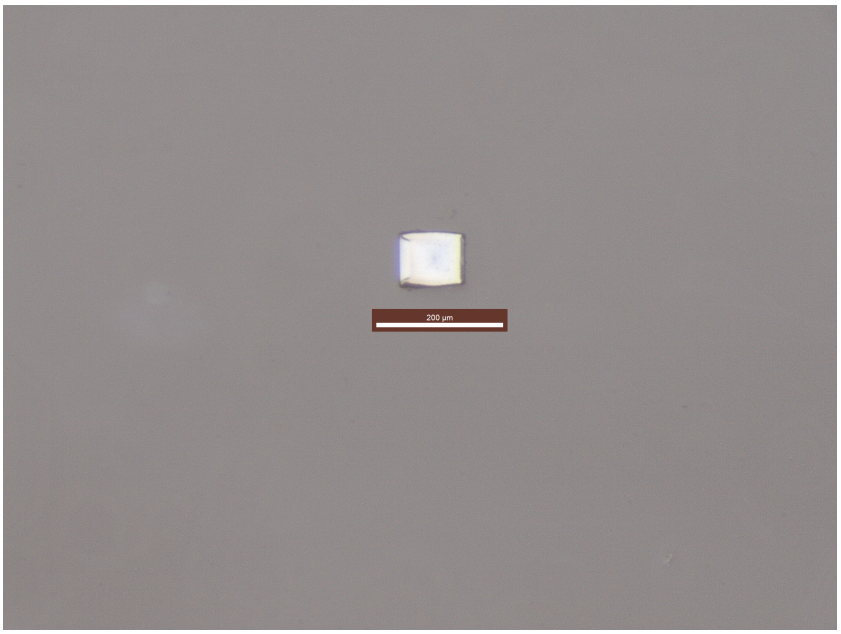



FIGURE 7.25: Single Crystal of Idelalisib obtained from slow evaporation of dichloromethane solution.

The analysis of our experimental results suggests that the structure predominantly adopts the less energetically favourable tautomer B. This conformation facilitates hydrogen bonding interactions with adjacent molecules.

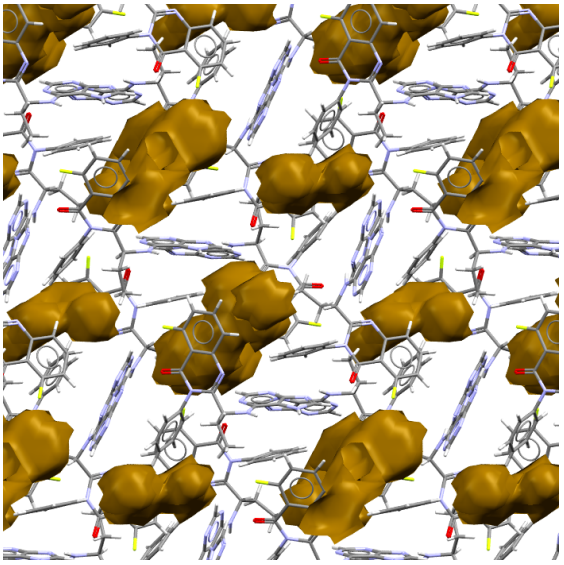


FIGURE 7.26: Crystal structure of the grown single crystal of idelalisib from the slow evaporation of dichloromethane

As seen in Figure 7.26, further investigation using PLATON software revealed a significant amount of void space within the structure, estimated to be approximately 560 $Å^3$. However, these voids exhibit minimal electron density, corresponding to roughly 0.5 DCM molecules per unit cell. The presence of both the higher-energy tautomer and the

seemingly stable porous framework indicates the potential formation of a kinetic product rather than the thermodynamically most stable system. This observation also implies the formation of a solvate, despite the lack of direct visualisation of solvent molecules within the structure.

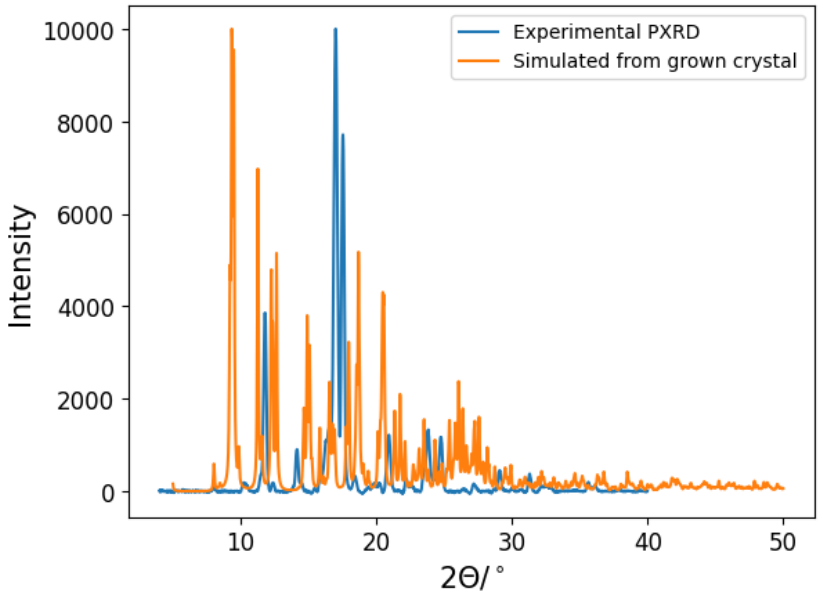



FIGURE 7.27: Powder X-ray diffraction (PXRD) patterns of two different samples of idelalisib. The experimental PXRD patterns were obtained from powdered idelalisib samples sourced from SelleckChem. For comparison, the simulated PXRD pattern was generated using PLATON from the crystallographic information file derived by solving the structure of a grown single crystal.

Comparison of PXRD patterns generated from the Crystallographic Information File (CIF) file is shown in shown in Figure 7.27. The overlay indicates that the structure under investigation does not correspond to the one obtained in the previous PXRD experiment as a substantial mismatch between the two structures has been observed. Consequently, the investigation of this particular crystal form has been discontinued.

### 7.2.6 MC Refinement

A MC refinement as described in section 6.7 was subsequently performed on the 32 lowest-energy structures. Five structures, all belonging to space groups P $4_1$ $2_1$ 2 and P $4_3$ $2_1$ 2, failed to meet the criteria for completion. However, the structures from the last accepted MC step were carried forward for analysis.

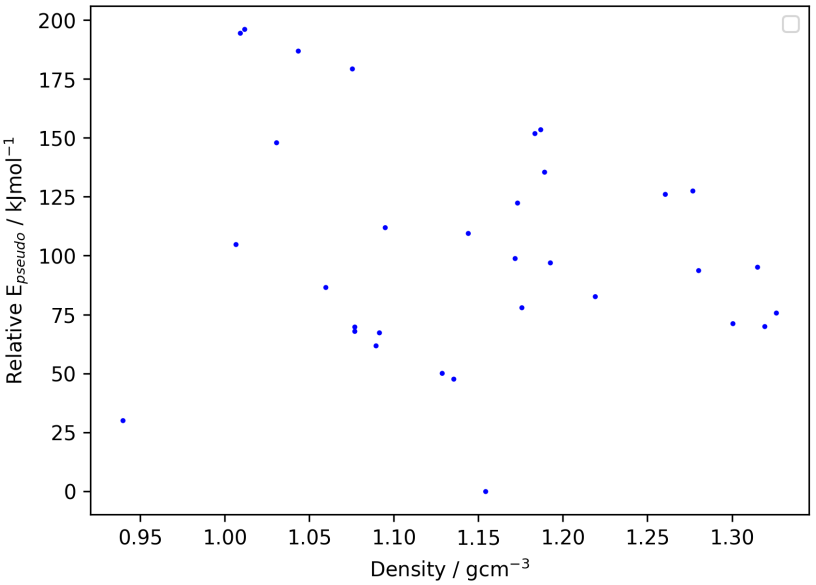



FIGURE 7.28: Crystal landscape for the search of idelalisib using a maximum of 4000 Monte Carlo steps, with $\lambda = 20$ kJ mol$^{-1}$ and a final temperature of 100 K for the linear profile. Each data point represents the final structure of a single trajectory. The experimental powder X-ray diffraction pattern for idelalisib was used to guide the search.

Figure 7.28 reveals a significant energy gap of 30.08 kJ mol$^{-1}$ between the first- and second-ranked structures, suggesting that the global minimum stands out distinctly from the other structures and providing some confidence in this candidate structure being that experimentally observed. Nonetheless, it remains uncertain whether the structure is a true experimental match.

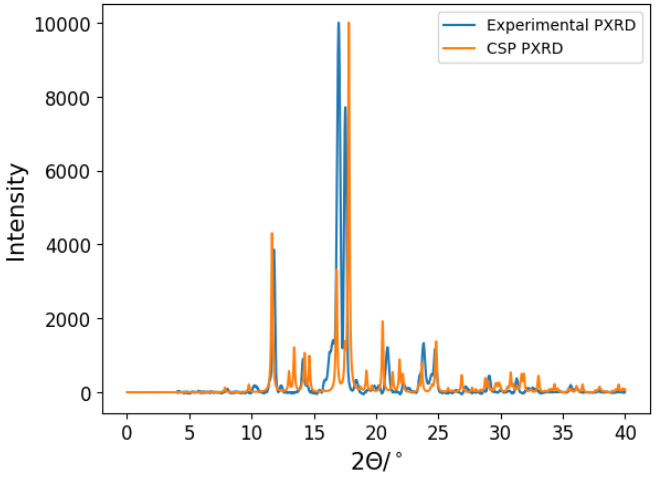



FIGURE 7.29: Powder X-ray diffraction (PXRD) patterns of two different samples of idelalisib. The experimental PXRD patterns were obtained from powdered idelalisib samples sourced from SelleckChem. For comparison, the simulated PXRD pattern was generated using PLATON from structure with the lowest pseudo energy after Monte Carlo refinement.

Comparison of PXRD patterns in Figure 7.29 shows reasonable agreement between structures. However there is an additional peak at around 13° which is unaccounted for by the experimental pattern.

## 7.3 Idelalisib Solvates

Three solvate structures were investigated using a methodology similar to that applied in the neat idelalisib workflow. For each solvate crystal, three stoichiometries were considered, corresponding to 1:1, 2:1, and 1:2 ratios of idelalisib to solvent. The DFTB–DMACRYS workflow described in the previous section was evaluated for its performance in modelling these solvated systems.

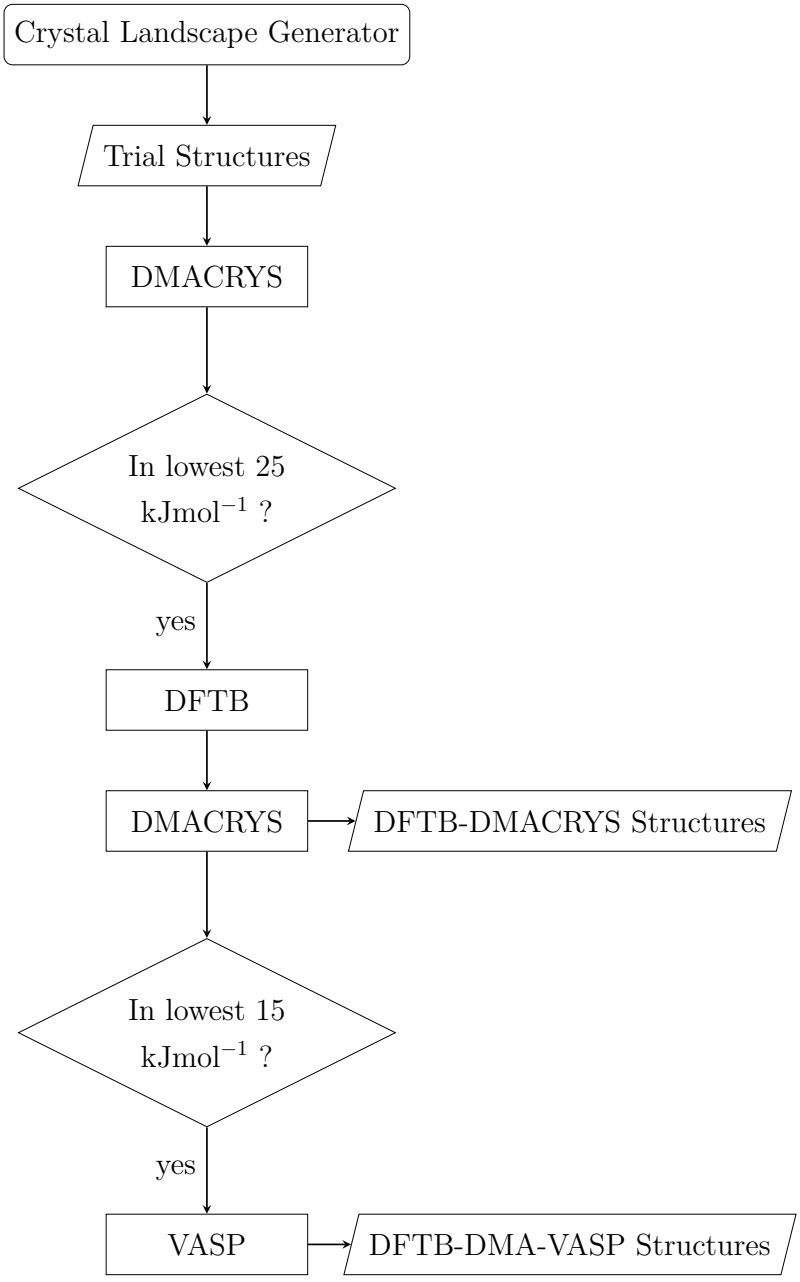



FIGURE 7.30: Workflow for crystal structure prediction of idelalisib solvates.

As idelalisib is chiral, the most 10 most common Schonke space groups for Z' > 1

were selected for CSP. These space groups contain no mirrors, inversion centres, roto-inversions.

| Space group | Number of valid structures |
|---|---|
| P 1 $2_1$ 1 | 50000 |
| P $2_1$ $2_1$ $2_1$ | 40000 |
| P 1 | 20000 |
| C 1 2 1 | 20000 |
| P $2_1$ $2_1$ 2 | 15000 |
| P $4_1$ | 10000 |
| P $3_1$ | 10000 |
| P $3_2$ | 10000 |
| P $4_3$ | 10000 |
| P $4_3$ $2_1$ 2 | 10000 |

TABLE 7.3: Number of valid crystal structures generated using the crystal landscape generator for each idelalisib solvate. The number of crystal structures correlates with how frequently each space group is observed within the crystallographic structural database.

The extent of sampling varies between space groups to account for their respective rates of occurrence in any given crystal structure. Consequently, more extensive sampling is conducted for the space group P 1 $2_1$ 1 compared to P $4_3$ to accurately reflect their differing frequencies.

### 7.3.1 Idelalisib:Pyridine

CSP was performed on an idelalisib–pyridine solvate using the sampling scheme detailed in Table 7.3.

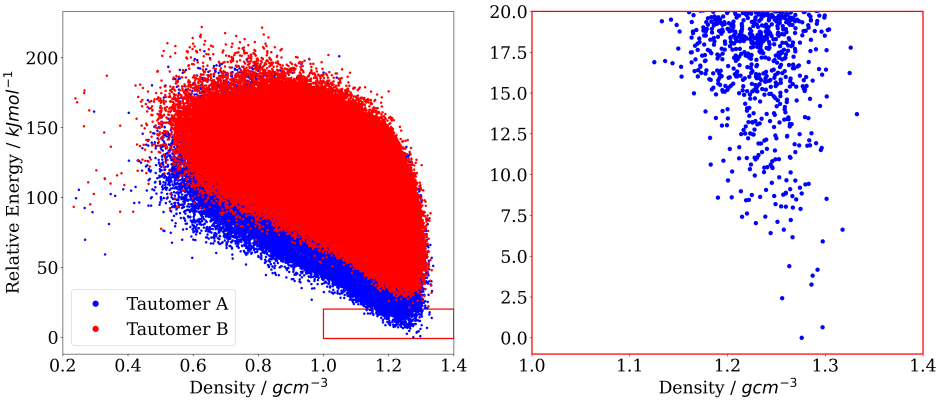



FIGURE 7.31: Crystal landscape of idelalisib:pyridine 1:1 search after CSP. Shown in red is the magnified low energy region of the crystal landscape corresponding to 20 kJ mol$^{-1}$ above the global minimum.

Around 3 million crystal structures were found across all identified conformations of idelalisib. Again, the low energy region was dominated by tautomer A.

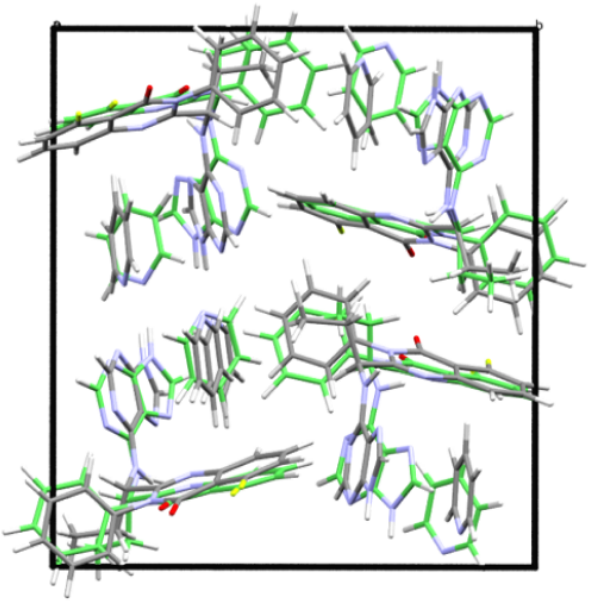


FIGURE 7.32: Overlay of predicted crystal of idelalisib:pyridine and the experimentally observed crystal structure before further re-optimisation. COMPACK 30/30 molecules within distance and angular tolerances of 20 % and 30° respectively. RMSD: 0.868 Å. **KEY:** Grey – carbon; white – hydrogen; red – oxygen; yellow – fluorine; green – carbons belonging to the experimental crystal.

Figure 7.32 illustrates the crystal structure prior to optimisation. The $RMSD_{30}$ value between the experimental and predicted crystals is 0.868 Å, indicating a reasonable resemblance between the two structures. However, the molecular conformation in the predicted structure requires further optimisation.

To address this, additional optimisation was performed by considering structures within 20 kJ $mol^{-1}$ of the global energy minimum. A total of 651 structures underwent DFTB+ - DMACRYS minimisation to refine the molecular geometries within each crystal structure.

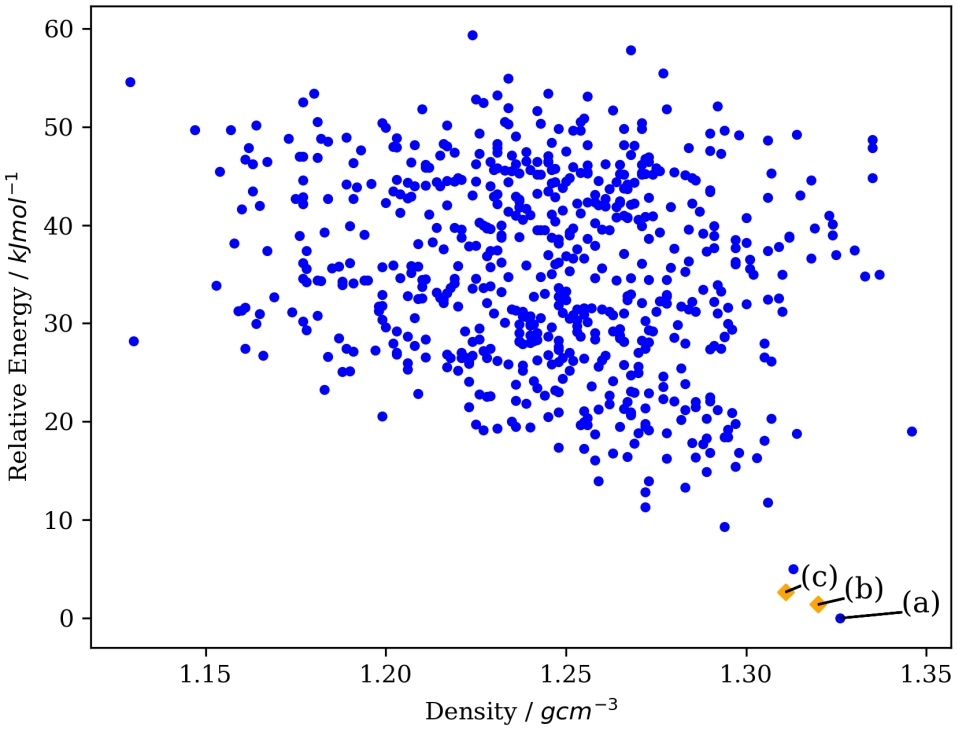
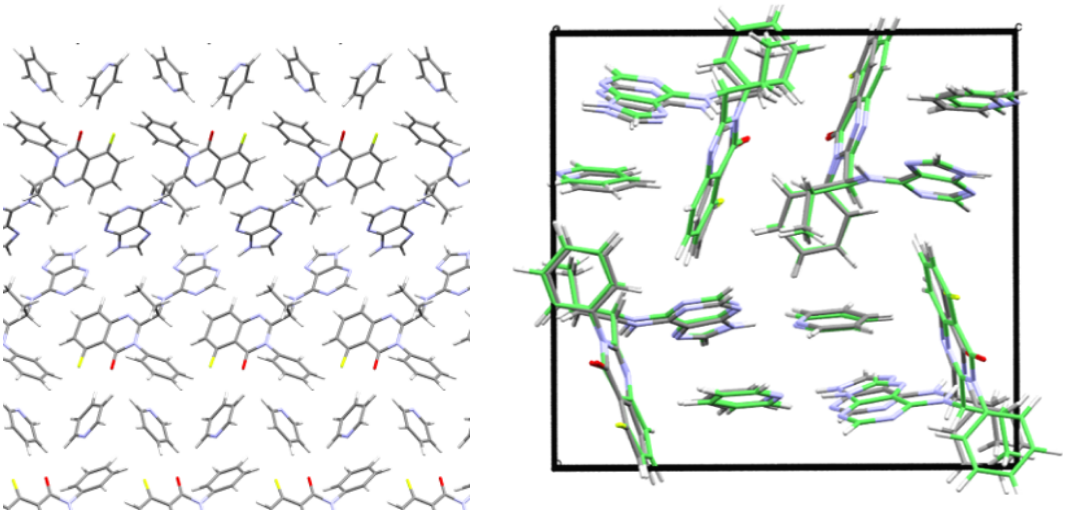
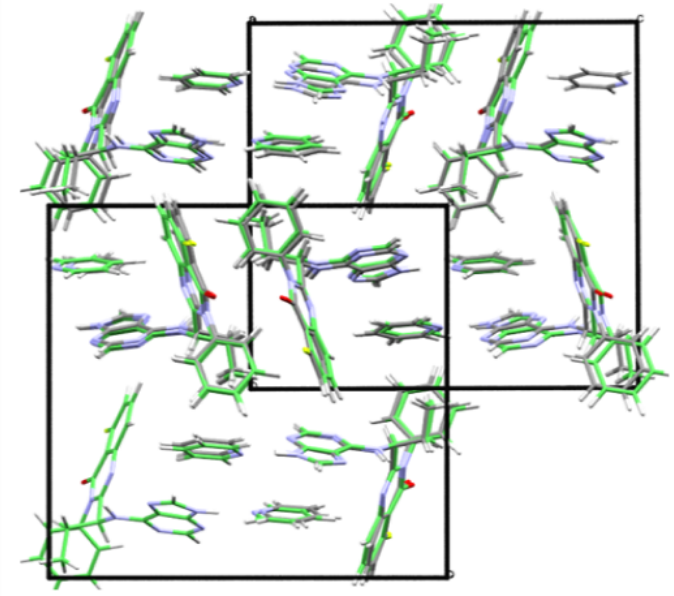



FIGURE 7.33: Crystal Landscape of idelalisib:pyridine 1:1 solvate after DFTB+ and DMACRYS optimisation. Structures shown as diamonds match to experimentally observed structure

After optimisation, two structures were identified as matching the experimental crystal structure of the idelalisib solvate. These structures are ranked second and third in terms of energy, indicating good agreement with the experimental data. Notably, the global minimum energy structure does not correspond to the experimental structure.



(A) 0.00 (B) 1.40



(C) 2.67

FIGURE 7.34: Overlay of predicted crystal to experimentally observed crystal structure after further re-optimisation with DFTB+ - DMACRYS. **KEY:** Grey – carbon; white – hydrogen; red – oxygen; blue - nitrogen; yellow – fluorine; green – carbons belonging to the experimental crystal. Shown are the relative total energies in kJ $mol^{-1}$.

The relative ranking of structures before and after re-optimisation can be examined to assess the extent of re-ranking. This information helps determine how large an energy window should be considered for subsequent analyses.

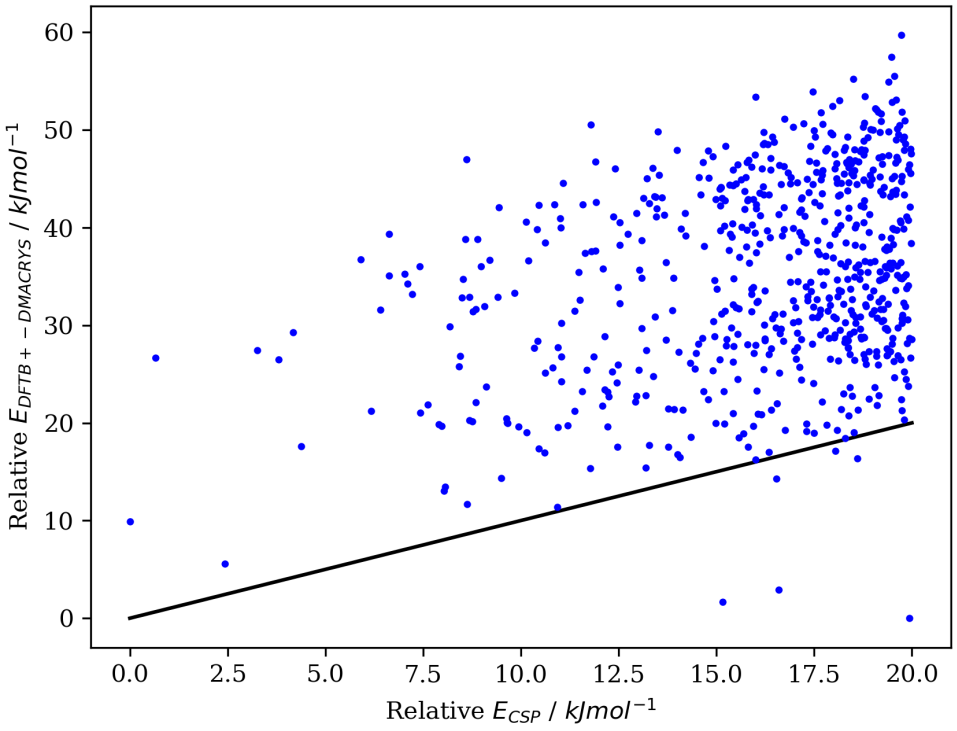



FIGURE 7.35: Energy re-ranking of idelalisib:pyridine solvate before and after DFTB-DMACRYS optimisation. The solid black line indicates y = x where the ranking of both structures is the same.

Fewer structures were minimised to low energy levels compared to the re-ranking of the neat idelalisib system. However, the new global minimum was obtained from a structure initially positioned approximately 20 kJ mol$^{-1}$ above the previous global minimum. To ensure that all potential structures capable of achieving a lower energy state are considered, it is recommended to expand the energy window to beyond 20 kJ mol$^{-1}$.

### 7.3.2 Idelalisib:Dimethylacetamide

CSP was conducted on an idelalisib–dimethylacetamide solvate using the sampling scheme outlined in Table 7.3.

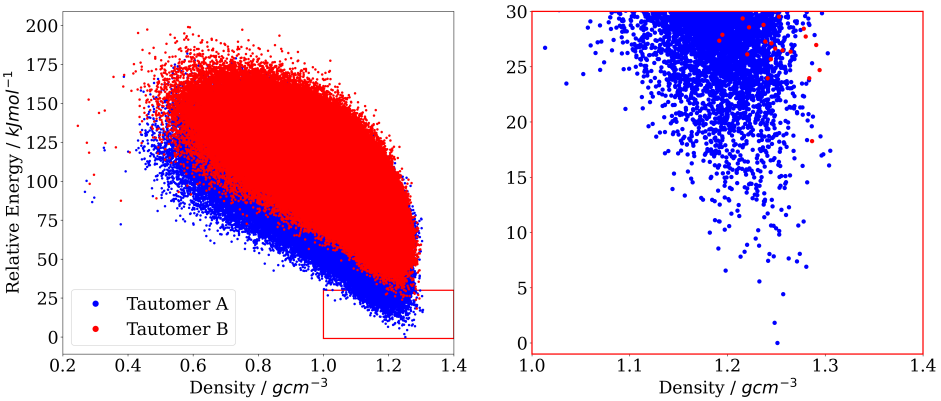



FIGURE 7.36: Crystal landscape of idelalisib:dimethylacetamide 1:1 crystal structure prediction. Shown in red is the magnified low energy region of the crystal landscape corresponding to 30 kJ mol$^{-1}$ above the global minimum.

Around 3 million crystal structures were found across all identified conformations of idelalisib. Again, the low energy region was dominated by tautomer A.

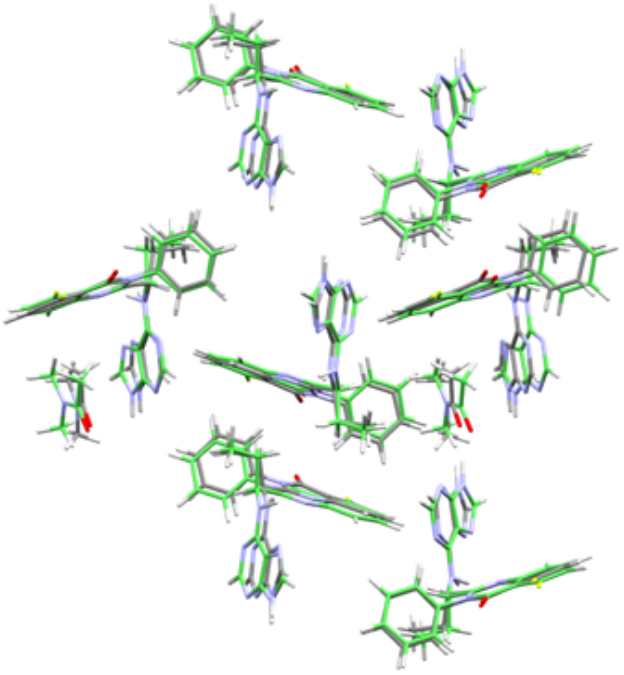


FIGURE 7.37: Overlay of predicted crystal for idelalisib:dimethylacetamide to experimentally observed crystal structure before further re-optimisation. COMPACK 30/30 molecules within distance and angular tolerances of 20 % and 30 Å respectively. RMSD$_{30}$: 0.483 Å. **KEY:** Grey – carbon; white – hydrogen; red – oxygen; blue - nitrogen; yellow – fluorine; green – carbons belonging to the experimental crystal.

Crystal structures were compared to the experimentally observed crystal structure. Figure 7.37 displays the closest match, indicating potential for further optimisation. A total of 1257 structures were identified within 25 kJ $mol^{-1}$ of the global minimum. These structures have been selected for further re-optimisation using DFTB+.

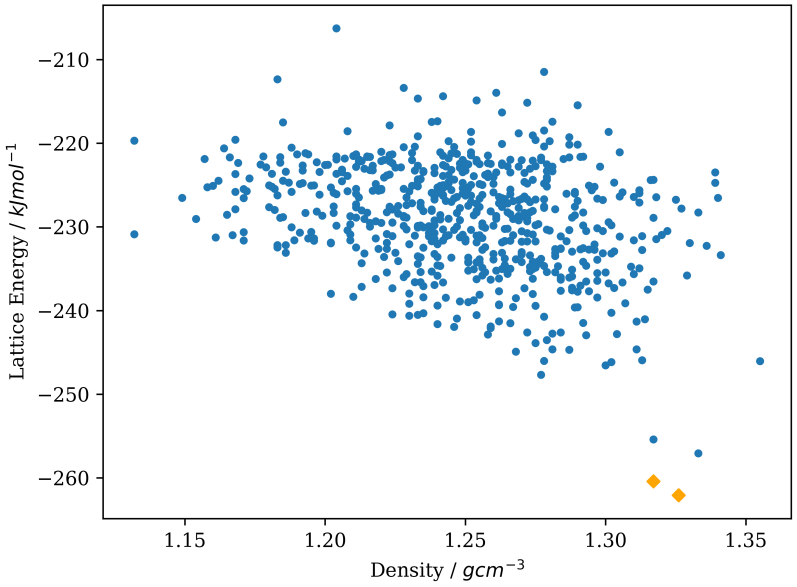



FIGURE 7.38: Idelalisib:dimethylacetamide landscape for 1:1 stoichiometry. Matches to experimental structure indicated with orange diamonds.

### 7.3.3 Idelalisib:Acetonitrile

CSP was performed on an idelalisib–acetonitrile solvate using the sampling approach detailed in Table 7.3. The resultant landscape is shown in Figure 7.39.

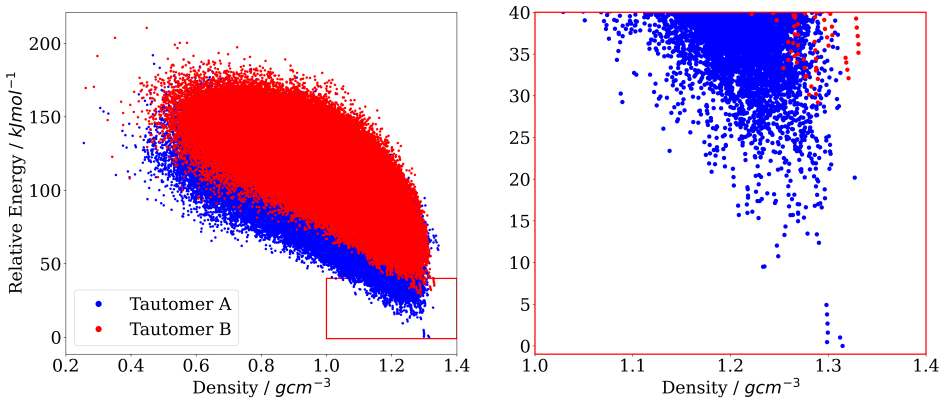



FIGURE 7.39: Crystal landscape of idelalisib:acetonitrile 1:1 search after CSP. Shown in red is the magnified low energy region of the crystal landscape corresponding to 40 kJ $mol^{-1}$ above the global minimum.

No close matches to structures were identified during this search. Further examination of the experimental structure revealed that the experimental form (Figure 7.40) adopts a 6:2 stoichiometry.

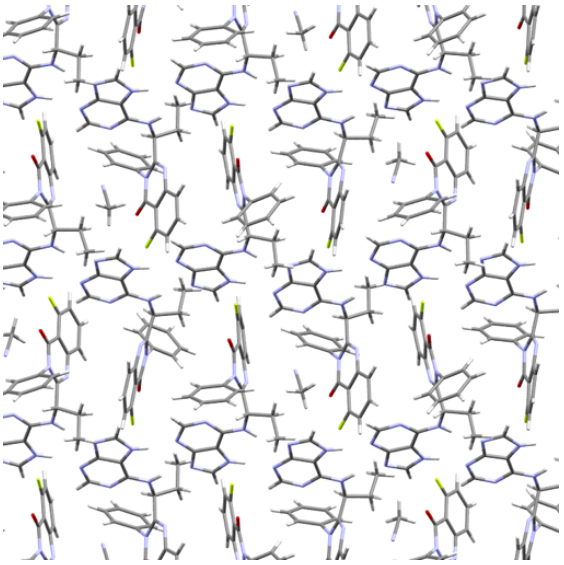


FIGURE 7.40: Experimental crystal structure of idelalisib:acetonitrile solvate. **KEY:** Grey – carbon; white – hydrogen; red – oxygen; blue - nitrogen; yellow – fluorine.

The low level of symmetry in the system may contribute to the difficulty in predicting the structure. As idelalisib molecules exhibit similar geometries, the crystal structure might be accessible by exploring a 3:1 stoichiometry. However, the computational cost of such an approach is substantial, and the CSP for this solvate under these conditions was not completed due to the excessive search space involved.

### 7.3.4 Solvate Stoichiometry Prediction

The energies of crystal structures cannot be directly compared across different stoichiometries because the number of molecules in each asymmetric unit varies. To address this issue, a convex hull was employed to evaluate the relative stabilities of the global minima for each stoichiometry [176].

These energies were subsequently compared to a constant stoichiometry comprising 2 moles of the API idelalisib and 2 moles of solvent, as presented in Equations 7.2.

$$\begin{aligned} E_{2:2}(1{:}1) &= E_{\text{total}}(1{:}1) + E_{\text{total}}(\text{Idel}) + E_{total}(\text{Solvent}) \\ E_{2:2}(2{:}1) &= E_{\text{total}}(2{:}1) + E_{\text{total}}(\text{Solvent}) \\ E_{2:2}(1{:}2) &= E_{\text{total}}(1{:}2) + E_{\text{total}}(\text{Idel}) \end{aligned} \tag{7.2}$$

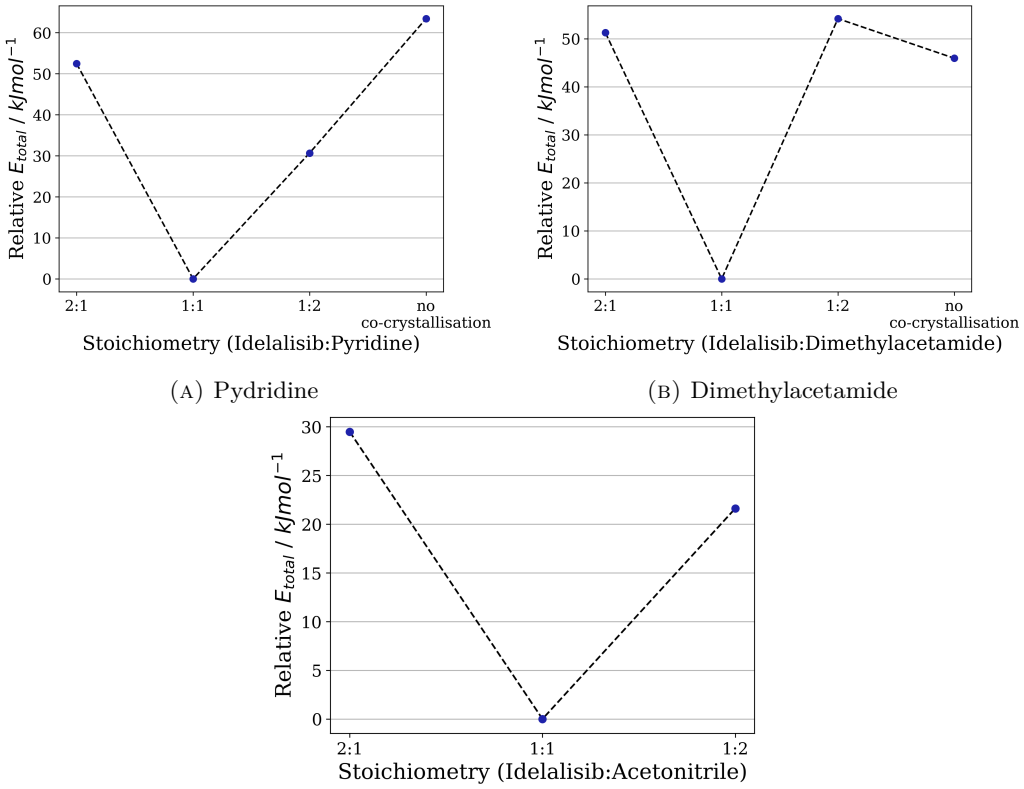
where $E_{2:2}$(1:1), $E_{2:2}$(2:1) and $E_{2:2}$(1:2) are the corrected energies compared to a constant 2:2 stoichiometry, $E_{\text{total}}$(Idel) and $E_{\text{total}}$(Solvent) are the total crystal energies of the experimentally observed crystal structures for idelalisib and the solvent respectively.

To determine the $E_{total}$(Solvent), crystals from the CSD for each solvent were used and optimised using DFTB+ followed by DMACRYS. For pyridine entries, PYRDNA01, PYRDNA02, PYRDNA03, PYRDNA04, PYRDNA05, PYRDNA06 were used from the CSD [177–180]. For acetonitrile, QQQCIV01 and QQQCIV08 were used [120, 181]. This ensured structures were energetically sensible. For dimethylacetamide, no experimental crystal structure could be found in the literature, so a Z' = 1 CSP search was performed, utilising the workflow described in this section. The lowest energy crystal structure was selected as a proxy for $E_{\text{total}}$(DMAC).

To identify if co-crystallisation would take place, a comparison was made to the sum of the total energies for 2 mols of neat idelalisib and 2 mols of solvent crystallising such that:

$$E_{2:2}(\text{Independent}) = 2E_{\text{total}}(\text{Idel}) + 2E_{\text{total}}(\text{Solvent}) \tag{7.3}$$

where $E_{2:2}$(Independent) is the total energy for neat idelalisib and neat solvent where no co-crystallisation occurs.




(A) Pydridine

(B) Dimethylacetamide

(C) Acetonitrile

FIGURE 7.41: Relative stabilities of idelalisib solvate stoichiometries. Lower relative total energy indicates greater stability of the stoichiometry. Each energy value is calculated relative to 2 mol of idelalisib and 2 mol of solvent.

The idelalisib–pyridine and idelalisib–dimethylacetamide solvates were both accurately predicted to exist in a 1:1 stoichiometry. The idelalisib–acetonitrile solvate was also predicted to adopt a 1:1 stoichiometry; however, stoichiometries with $Z' > 3$ were not examined. Future work should investigate the 3:1 and 6:2 stoichiometries and compare the resulting structures after re-optimisation using DFTB+ and DMACRYS.

## 7.4 Conclusions and Future Work

For neat idelalisib, a convincing match between the predicted and experimental crystal structures has not yet been identified. Although some of the hypothetical crystals appear similar, none provide a sufficiently strong correlation to the experimental diffraction data. This discrepancy may be due to insufficient sampling, particularly given that idelalisib is a flexible molecule featuring multiple rotatable bonds. Further additional sampling could therefore be crucial for improving the chances of locating the experimental structure within the predicted crystal energy landscape.

To see if more sampling is needed, assessing the completeness of the current landscape would be an informative next step. One way to evaluate this is to analyse the diversity of unique crystal structures generated towards the end of CSP runs. If the landscape is well-sampled and nearing completion, newly generated structures should tend to minimise into already identified energy minima rather than yielding distinct new crystal packings.

Further computational studies focusing on tautomer B of idelalisib are also recommended. Experimental evidence suggests that idelalisib may crystallise in the B tautomeric form, and additional CSP efforts targeting this conformer could potentially yield superior results. Moreover, the molecular conformation observed within the single known experimental crystal structure could be extracted and employed as the asymmetric unit in subsequent CSP calculations, potentially improving the chances of identifying a correct match.

Another avenue involves applying the MCSA algorithm for more structures. This could be utilised either for performing MC refinement on those predicted structures whose simulated PXRD patterns resemble the experimental pattern, or by executing full CSP runs starting from QR structures. However, it should be noted that the MCSA approach previously failed to produce satisfactory results for other complex systems such as ROY, suggesting that it may require further methodological enhancements before proving effective for idelalisib.

Where the MC refinement of crystal structures of idelalisib was carried out, no matches to the experimental crystal structure were identified. Although comparisons based on PXRD suggest reasonable agreement, the results remain unconvincing. Further work could involve increasing the number of structures taken forward from the crystal landscape for analysis. Additionally, ss-NMR studies could be performed on the idelalisib sample, and the resulting data incorporated into the cost function, potentially aiding in the search for the correct structure.

Given the significant flexibility of the idelalisib molecule and the diversity of its possible crystal packing arrangements, it is also advisable to maintain a relatively large energy window in the DFTB+–DMACRYS workflow. This ensures that potentially relevant

structures, even if initially predicted at higher energies, are not prematurely excluded from consideration. Moreover, the extent of sampling should remain high to maximise the likelihood of discovering the experimental structure.

It may also be beneficial to perform a dedicated DCM CSP to investigate whether the experimental crystal is, in fact, a solvate. Such an approach could help confirm the nature of the experimentally observed phase and explain discrepancies between the predicted and observed data.

The crystal structures of two solvate forms of idelalisib have been successfully predicted. However, the experimental structure of the acetonitrile solvate has not yet been determined, most likely due to its asymmetric nature. It is proposed that further CSP calculations could be undertaken using a stoichiometry of 6:2, which may enable matches to be found with the experimental structure. Nevertheless, the computational cost associated with performing CSP at this stoichiometry is considerable, and it may be prudent to restrict the search to the space group of the experimental structure. Consequently, a blind search for this crystal structure is likely not to be feasible.

An alternative approach for identifying the crystal structure of the idelalisib:acetonitrile solvate involves employing CSP with a 3:1 stoichiometry. This could be followed by perturbation techniques, which may assist in identifying the crystal structure and potentially refining it through the MC refinement process using a supercell.

# Chapter 8

# cspy-flex

*The following work was carried out in collaboration with Ramón Cuadrado, Joseph Glover, Christopher R. Taylor, and Graeme M. Day. As part of this research, conformer searches were performed and torsional ranges for different molecules were investigated by the author. Cuadrado completed flexible-CSP calculations for XBCN90 and developed a computational pipeline; Glover completed flexible-CSP calculations for FAHNOR; Taylor programmed a significant portion of the code; and Day provided expert advice.*

In Chapters 4 and 7, a rigid-conformer packing strategy was demonstrated, wherein gas-phase metastable conformations of the target molecule were generated, packed using the CLG, and subsequently relaxed with DMACRYS. The primary limitation of this approach lies in its reliance on the gas-phase conformer already closely resembling the crystal geometry.

Earlier studies addressed tackled flexible molecules in various ways. Fully flexible searches incorporate intramolecular torsions directly into the global optimisation, albeit at significant computational expense. Evolutionary algorithms perform parallel genetic searches using empirical force fields, simultaneously exploring both packing and conformation; although these methods have been successful for drug-like molecules, they remain constrained by the inherent inaccuracies of force fields [182]. More recent research has leveraged ML explicitly for flexible CSP: Butler et al. employed active learning to train neural potentials iteratively on density-functional data acquired during CSP, enabling fully flexible lattice relaxations at computational costs comparable to those of force fields [183]. This allows for easier exploration of configurational space.

To address the conformer-mismatch problem without incurring the full cost of ML training, a new approach, *cspy-flex*, is introduced. Drawing inspiration from previously employed methods, two potential workflows are proposed, incorporating both global and local conformer sampling. A global-sampling path executes a lightweight search to suggest diverse conformations, while a local-sampling path perturbs gas-phase conformers

within user-defined torsion windows. Both workflows pack conformations using the CLG and minimise with DMACRYS as before which preserves the proven accuracy of the energy model while expanding the scope of explored conformational space.

## 8.1 Global Sampling of Conformational Space

In principle, it is possible to sample conformational space by conducting a grid based search across flexible torsion angles of a molecule. The number of conformations in this instance would scale according to the degree of sampling and the number of flexible torsion angles a molecule possesses.

$$C = \prod_{t=0}^{T} \frac{2\pi}{s_t}\,, \tag{8.1}$$

where $C$ is the total number of conformations, $t$ is the torsion number, $T$ is the total number of torsions, $s_t$ is the angular step size of torsion $t$ in radians.

This method can generate a large number of conformations. Consequently, it is necessary either to limit the resolution of the grid or to restrict its application to molecules with few flexible torsions.

Focus has been concentrated on predicting the crystal structure of XBCN90 shown in Figure 8.1 [184].

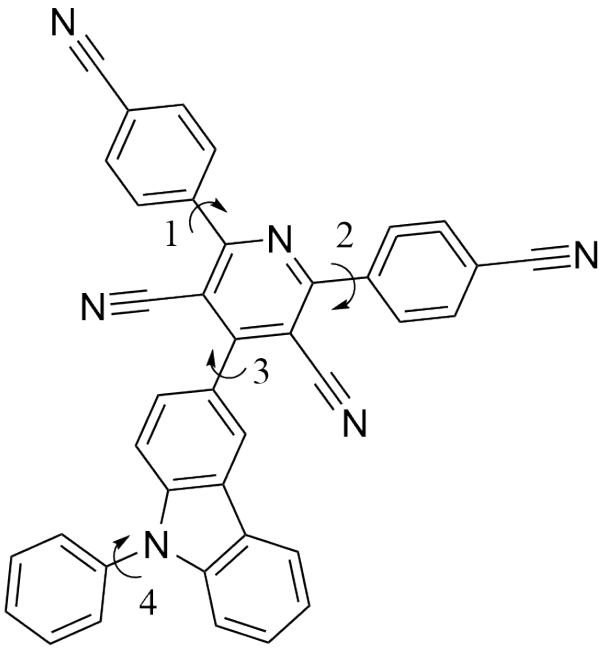



FIGURE 8.1: XBCN90. Flexible torsions (1-4) are shown with arrows.

If the global minimum conformation from a CREST search is distorted around all flexible torsion angles using an angular step size of $\frac{\pi}{4}$ radians, or 45°, a total of 4096 potential conformations are generated.

Conformational space can be more effectively sampled by adjusting the level of sampling around each flexible torsion, thereby placing greater emphasis on regions of particular interest. For example, sampling around XBCN90 could be performed using the following set:

| Torsion Number | Angular Range / ° | Angular Step Size / ° |
|---|---|---|
| 1 | 70 | 7 |
| 2 | 20 | 2 |
| 3 | 48 | 12 |
| 4 | 24 | 6 |

TABLE 8.1: Extent of sampling performed during global sampling in `cspy-flex` for XBCN90. Torsion numbers (1–4) correspond to flexible torsions labelled in Figure 8.1. The angular range represents the total region around a conformer sampled using the specified angular step size.

Some generated conformations resulted in atomic clashes and were consequently discarded. After removing all clashing geometries, a total of 3630 conformations remained. These conformations were subsequently used with the CLG to generate crystal structures in the P $2_1$ $2_1$ $2_1$ space group, in which XBCN90 has been experimentally observed to crystallise. The crystal energy landscape is shown in Figure 8.2.

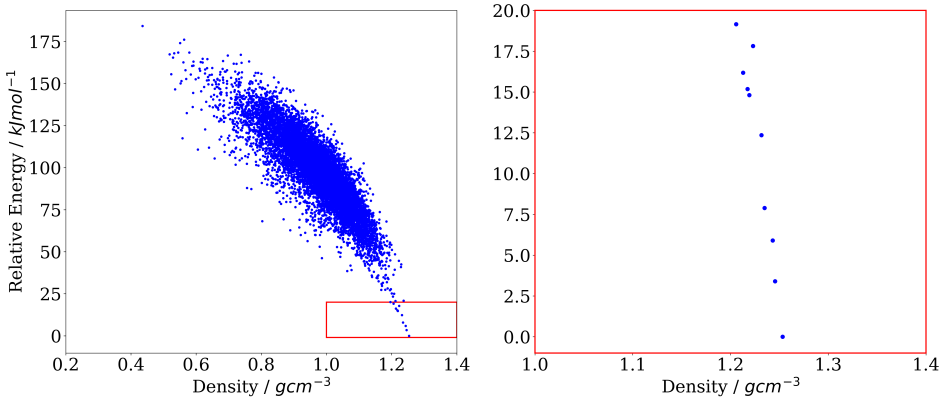



FIGURE 8.2: Crystal landscape of XBCN90 search after global flexible crystal structure prediction for space group P $2_1$ $2_1$ $2_1$ only. Shown in red is the magnified low energy region of the crystal landscape corresponding to 20 kJ mol$^{-1}$ above the global minimum.

The global energy minimum structure for this structure was then compared against the experimentally observed structure.

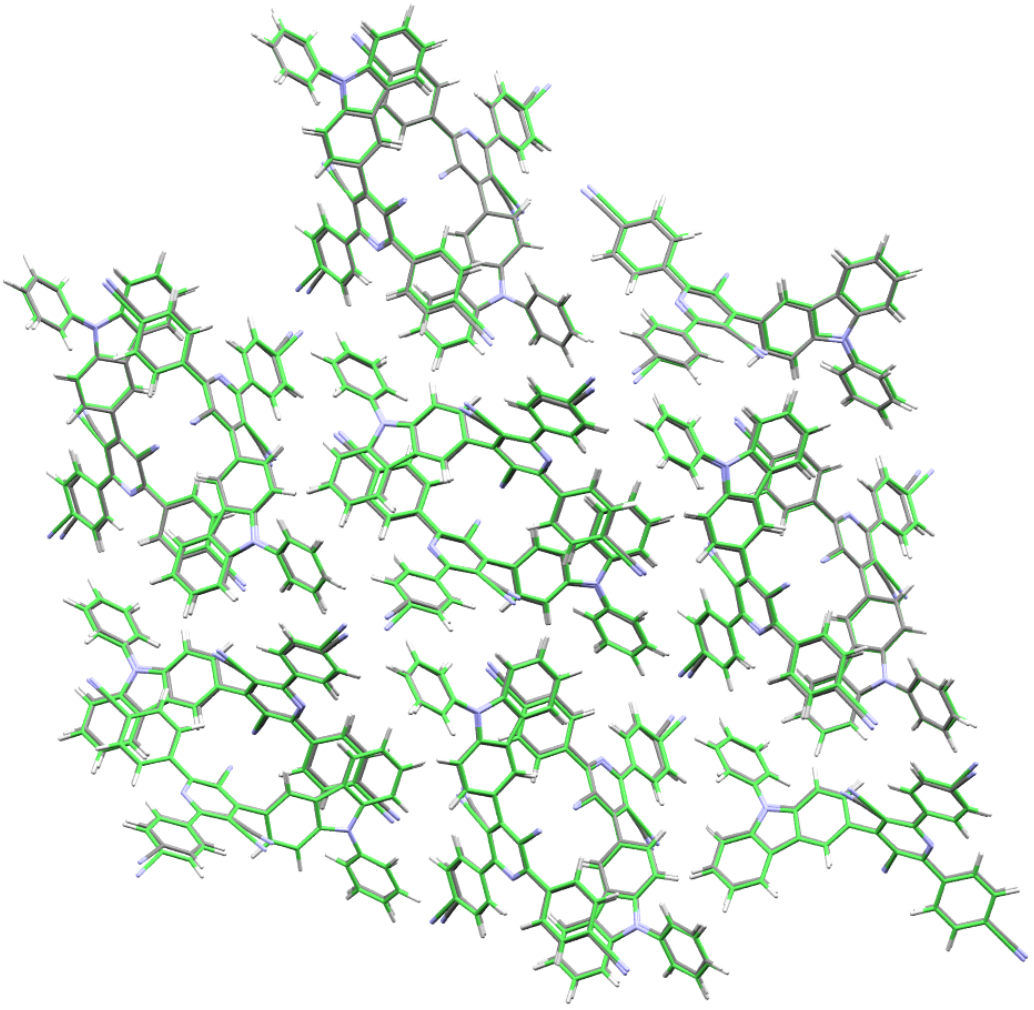


FIGURE 8.3: Overlay of predicted crystal of XBCN90 to experimentally observed crystal structure. COMPACK 30/30 molecules within distance and angular tolerances of 20 % and 30°. RMSD: 0.3300 Å **KEY:** Grey – carbon; white – hydrogen; blue - nitrogen; green – carbons belonging to the experimental crystal.

A comparison between the structures as seen in Figure 8.3, shows that there is an experimental match using COMPACK having an $RMSD_{30} = 0.33$ Å. Further optimisation using our post-CSP workflows was not needed.

## 8.2 Local Sampling of Conformational Space

An alternative approach to sampling global conformational space involves focusing on local sampling around each gas-phase conformer. Since molecules typically adopt conformations that are closely related to experimentally observed conformers, exploring local conformations can be achieved by introducing small distortions to flexible torsion angles [185, 186]. This technique allows for an effective exploration around each energy minimum as shown in Figure 8.4.

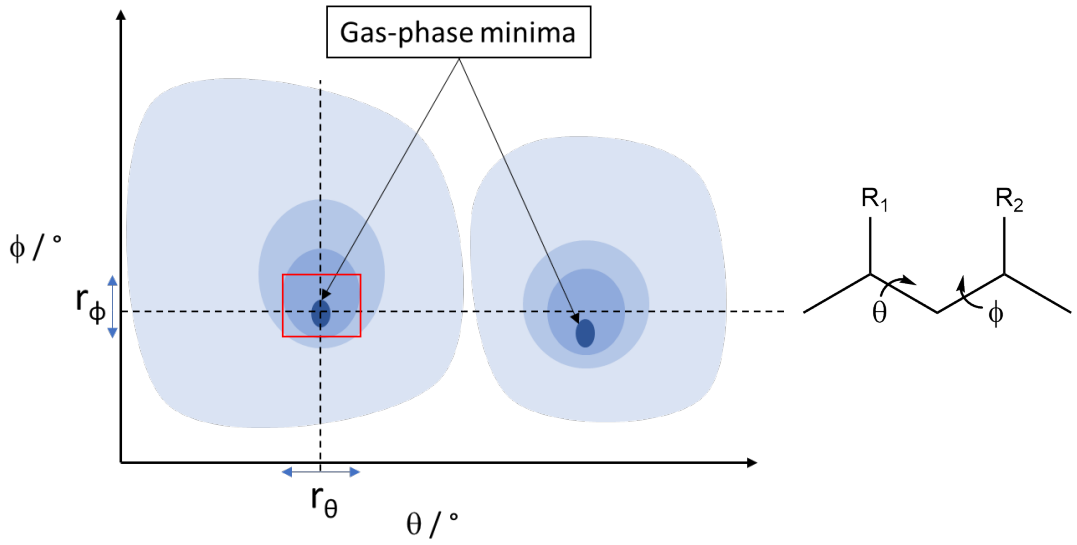



FIGURE 8.4: Illustration of local conformational sampling around each conformer. Conformers with low energy are indicated with darker blue. The region within the red square is area sampled by the local conformational sampling indicated by angular ranges $r_\theta$ and $r_\phi$.

By utilising local distortions the amount of conformations needed for the CSP calculations can be reduced.

$$C = n \prod_{t=0}^{T} \left(\frac{r_t}{s_t} + 1\right), \tag{8.2}$$

where $n$ is the number of gas phase conformers, $r_t$ is the angular range for torsion $t$ where $r_t < 2\pi$ and $s_t$ is the angular step size of torsion $t$. Again, some of these conformations may be discarded due to clashing atoms upon generation.

To identify the extent of sampling needed, experimental conformations found in a series of flexible pharmaceutical-like crystal structures were studied as shown in Figure 5.2. A conformational search was performed on each molecule using the program CREST. The extent to which each conformer within the ensemble would require rotation of its flexible torsions to align with the experimental conformation was calculated. The conformer

exhibiting the smallest maximum rotation among all its flexible torsions was selected from the ensemble. This angle value indicates the local sampling range necessary for the closest conformer to achieve the conformation observed in the crystal structure. The results are summarised in Figure 8.5.

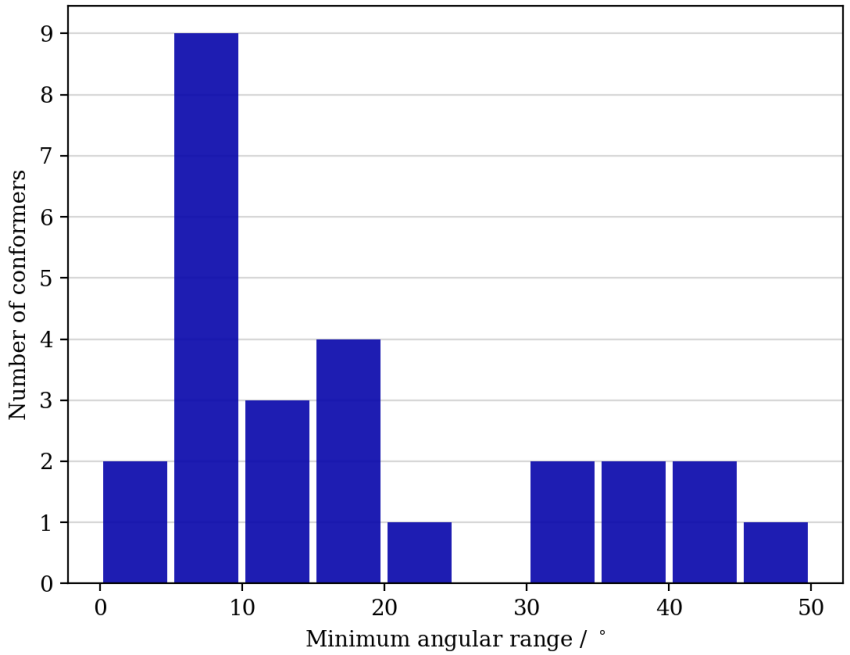



FIGURE 8.5: Minimum angular range in which a generated conformer can be distorted to reach conformations within experimental crystal structures for a series of pharaceutical-like molecules shown in Figure 5.2.

Around 70% of experimental conformations can be found by distorting gas phase conformers by 20°. However some angles require distortions up to 50° across all flexible torsions. This means that when using local conformer sampling, each torsion should be distorted by up to 50° in order to reach the experimentally observed conformation.

### 8.2.1 Generating Crystal Structures

Given the vast number of conformations, it is not computationally feasible to perform DMA on all of them. Consequently, lower accuracy point charges were used to model the electronic densities.

Conformations generated from each workflow were packed as described previously in section 2.4.3.2.

### 8.2.2 FAHNOR

A conformational search using mCREST on the FAHNOR molecule was performed shown in Figure 8.6.

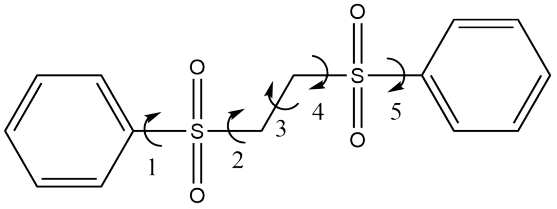



FIGURE 8.6: FAHNOR. Flexible torsions (1-5) are shown with arrows.

Local sampling was performed around each of the generated conformers.

| Torsion Number | Angular Range / ° | Angular Step Size / ° |
|---|---|---|
| 1 | 40 | 10 |
| 2 | 40 | 10 |
| 3 | 40 | 10 |
| 4 | 40 | 10 |
| 5 | 40 | 10 |

TABLE 8.2: Extent of sampling performed during local sampling in `cspy-flex` for FAHNOR. Torsion numbers (1–5) correspond to flexible torsions labelled in Figure 8.6. The angular range represents the total region around a conformer sampled using the specified angular step size.

35002 unique structures were generated using the sampling in Table 8.2. CSP was subsequently performed by randomly selecting conformers from the generated conformer database for packing. Crystal structures were generated in the usual method for space group P $2_1$ / c, resulting in the generation of 40,000 structures.

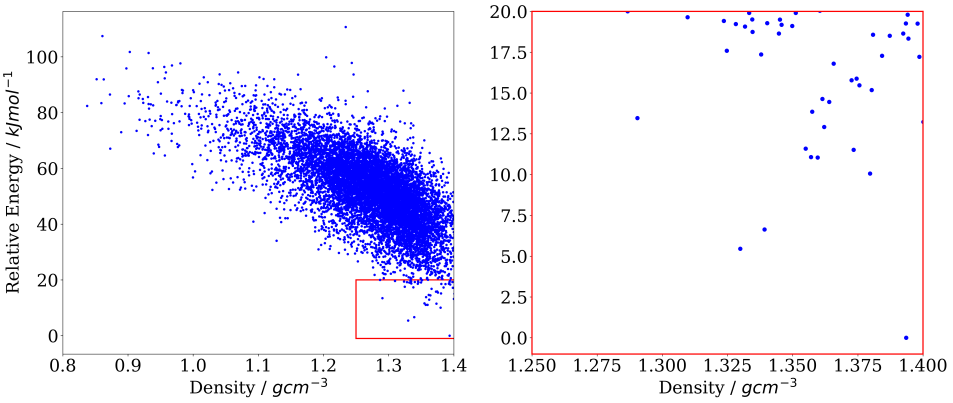



FIGURE 8.7: Crystal landscape of FAHNOR search after local flexible crystal structure prediction for space group P $2_1$ / c only. Shown in red is the magnified low energy region of the crystal landscape corresponding to 20 kJ mol$^{-1}$ above the global minimum.

One of the structures matched with the experimental structure with a COMPACK 30/30 search shown in Figure 8.8. This structure though had an energy of 13.1 kJ mol$^{-1}$ above the global minimum and ranked $42^{nd}$ in energy.

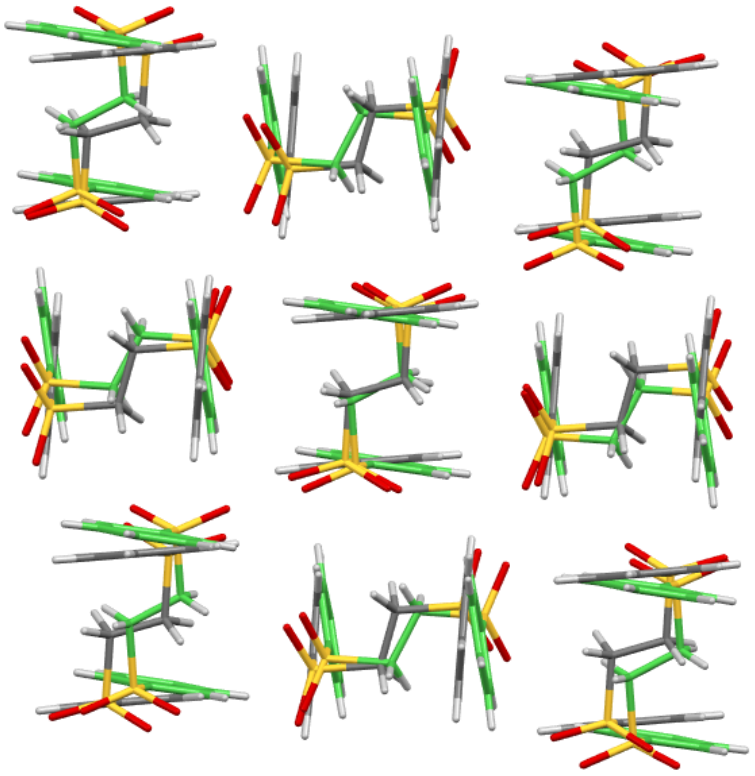


FIGURE 8.8: Overlay of predicted crystal of FAHNOR to experimentally observed crystal structure FAHNOR after initial crystal structure prediction. COMPACK 30/30 molecules within distance and angular tolerances of 20 % and 30°. RMSD: 0.64 Å **KEY:** Grey – carbon; white – hydrogen; yellow - sulphur; red - oxygen; green – carbons belonging to the experimental crystal.

For this last all the crystal structures were re-optimised up to 30 kJ mol$^{-1}$ above the global minimum using DFTB+. One of the structures matched with the experimental polymorph with an $RMSD_{30}$ of 0.23 Å and an energy of 6.78 kJ mol$^{-1}$ above the global minimum.

### 8.2.3 Idelalisib

It was not possible to produce a convincing match using our previous methodology employed for idelalisib. Therefore, an attempt was made using our *cspy-flex*.

A conformational search was conducted using mCREST on the idelalisib molecule, with sampling limited exclusively to idelalisib A, the lowest energy tautomer.

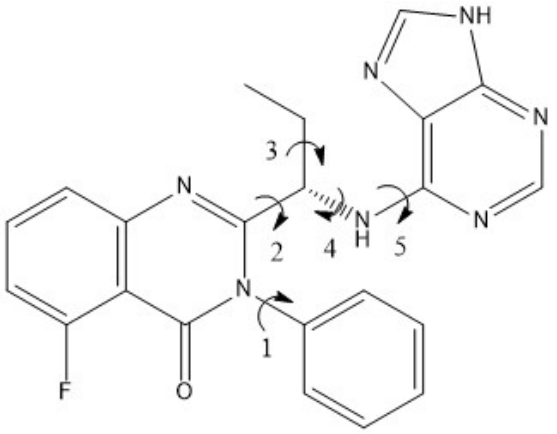



FIGURE 8.9: Idelalisib A. Flexible torsions (1-5) are shown with arrows.

Local sampling was performed around each of the conformers generated by our conformer search using the angular ranges and step sizes set out by Table 8.3.

| Torsion Number | Angular Range / ° | Angular Step Size / ° |
|---|---|---|
| 1 | 50 | 25 |
| 2 | 50 | 25 |
| 3 | 50 | 25 |
| 4 | 50 | 25 |
| 5 | 50 | 25 |

TABLE 8.3: Extent of sampling performed during local sampling in `cspy-flex` for idelalisib. Torsion numbers (1–5) correspond to flexible torsions labelled in Figure 8.9. The angular range represents the total region around a conformer sampled using the specified angular step size.

103,125 structures were generated using the sampling above. CSP was then performed by randomly selecting conformers for packing. The ten most common Sohncke space groups were sampled, generating 400,000 crystals for each space group. The resultant landscape is shown in Figure 8.10.

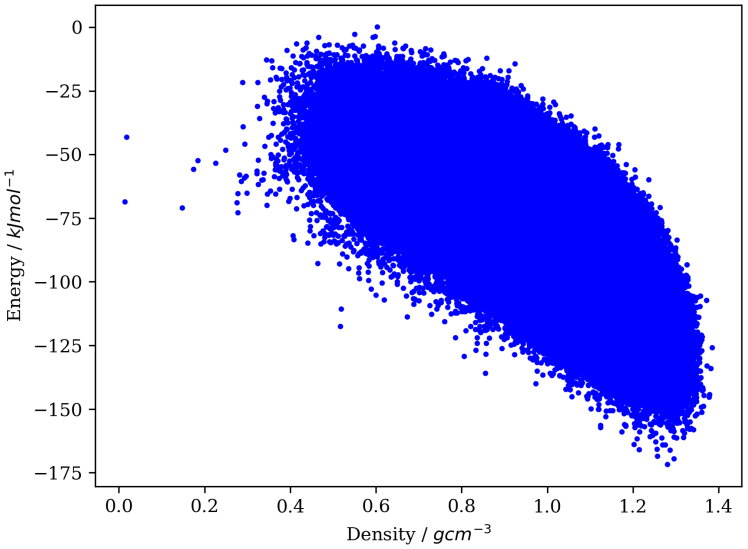



FIGURE 8.10: Crystal landscape of idelalisib search after local flexible crystal structure prediction across top 10 most common spacegroups.

Structures were optimised using the DFTB-DMACRYS workflow as described in section 7.2.1.2, taking structures 40 kJ $mol^{-1}$ above the global energy minimum. The resultant landscape is shown in Figure 8.11.

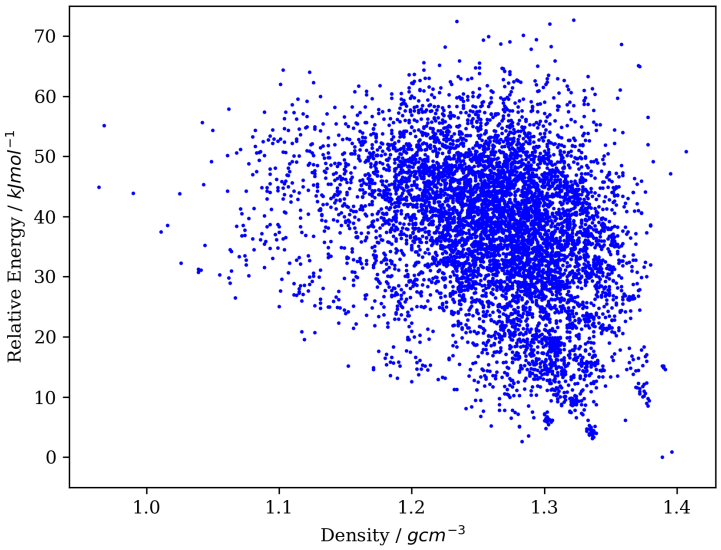
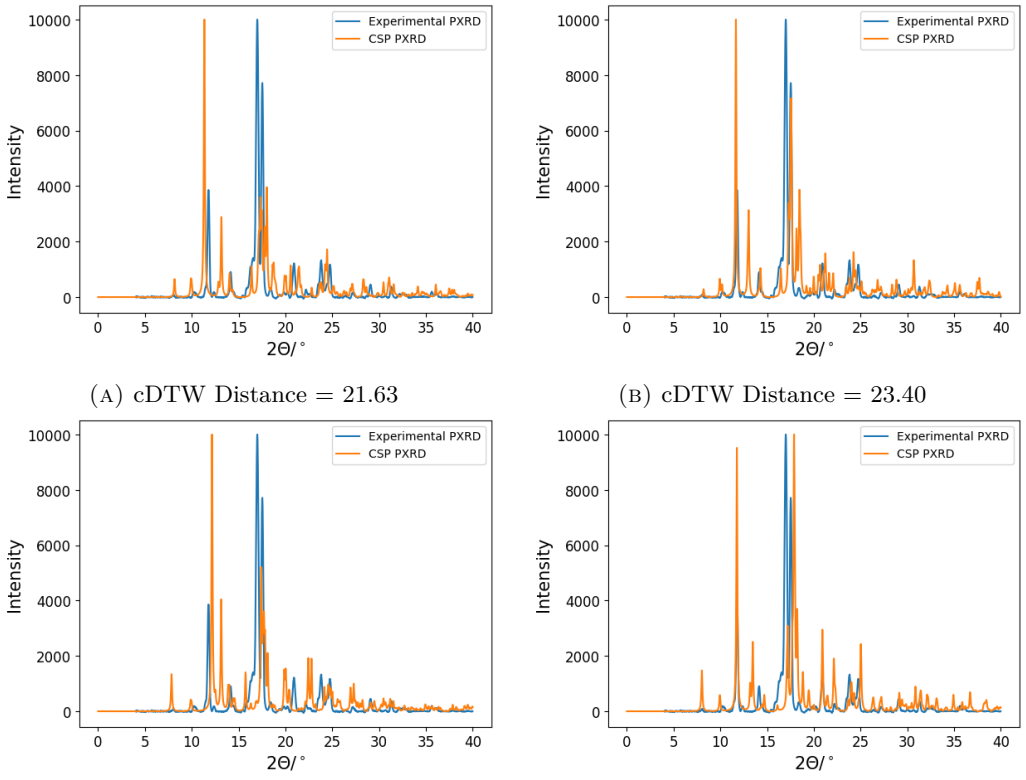



FIGURE 8.11: Crystal landscape of idelalisib search after local flexible crystal structure prediction across top 10 most common spacegroups after optimisation of structure up to 40 kJ $mol^{-1}$ above the global energy minimum. Only structure that have been optimised are shown.

Structures were compared with the experimental PXRD to identify any improvements made between the initial methodology and *cspy-flex* using local distortions.




(A) cDTW Distance = 21.63

(B) cDTW Distance = 23.40

(C) cDTW Distance = 24.85

(D) cDTW Distance = 25.05

FIGURE 8.12: Powder X-ray diffraction (PXRD) overlays between experimental and generated crystal structures for idelalisib using the cspy-flex workflow. Patterns shown are those with the lowest constrained dynamic time warping distance between patterns.

Some structures exhibit peaks that closely align with one another. However, the cDTW distance remains high, and peak intensities are also poorly matched. As previously observed, no convincing matches to the experimental PXRD pattern were identified.

## 8.3 Conclusions and Future Work

The use of *cspy-flex* eliminates the need for extensive post-CSP optimisation of structures, as molecular conformations are already packed in a perturbed state. This approach allows for the adoption of new geometries that might otherwise be overlooked when packing only standard molecular conformers.

Future calculations may benefit by performing post-CSP re-optimisation using DFTB+. As in the case of FAHNOR, it was seen to improve the geometric match between the predicted crystal and experimental structure.

However, the computational cost of this method is significant. While the full effectiveness of the method has not yet been established, future work is necessary to test this methodology on a range of flexible molecules. The challenge of performing multiple GDMA calculations is compounded by this computational expense. Approximations, such as assuming that the multipoles of a perturbed conformer are the same as the unperturbed conformer, could help mitigate this issue. Nevertheless, further testing is required to confirm the viability of this approach.

It is possible to use the relative energies of conformations to bias this selection, as described by Equation 8.3. Here, instead of selecting conformers uniformly with probability $\frac{1}{C}$, each conformer $i$ is assigned a weight $w_i$ that depends on its relative intramolecular energy $E_i$. The probability $p_i$ of selecting conformer $i$ is then given by:

$$p_i = \frac{w_i}{\sum_{j=1}^{C} w_j} \tag{8.3}$$

where $w_i$ is a weight function defined to decrease with increasing conformer energy, for example:

$$w_i = \frac{1}{1 + E_i} \tag{8.4}$$

In this way, lower-energy conformers are preferentially selected, while still allowing the possibility of sampling higher-energy conformations.

# Chapter 9

# Conclusions

This thesis has detailed the development of several CSP methodologies. These approaches were designed to address the challenges of predicting the crystal structures of flexible, drug-like small molecules, which typically possess multiple rotatable bonds, can adopt different tautomeric states, crystallise in a range of solvate stoichiometries, and exhibit complex packing motifs. The main objective of the work was therefore to refine CSP protocols so they can reliably handle such systems and to introduce a set of novel techniques that overcome the limitations of existing approaches.

Key features of dealing with such molecules has been identified which must be considered when performing CSP on pharmaceuticals: conformational and crystal sampling, post CSP optimisation and integration of experimental data into workflows.

### Conformational Sampling

A initial challenge was achieving comprehensive sampling of the PES. A single conformational search rarely explores the full PES; instead, running multiple, independent searches markedly increases the diversity of conformers obtained. The resulting ensemble is best clustered through the use of molecular torsion angles rather than by RMSD, as the former affords finer control over which conformers are retained. The mCREST protocol produced broader coverage of the PES and consistently out-performed both LMCS and a single CREST run.

Tautomerism must also be considered. High-energy gas-phase tautomers can become thermodynamically competitive in the solid state if their geometry stabilises otherwise inaccessible hydrogen-bonding patterns. Therefore it is recommended that performing conformer searches for every plausible tautomer.

In addition to packing pre-optimised conformers into the CLG, performing local sampling around each conformer and packing them, may yield lower-energy structures that

require minimal further optimisation with DFTB. Conversely, a global, grid-based exploration of the conformational space may be employed to ensure that crystal conformations differing substantially from the gas-phase ensemble are not overlooked. Both approaches, however, face challenges associated with the rapid expansion of candidate structures and the limitations of point-charge electrostatics in reliably ranking them. To achieve quantitative accuracy, it will be necessary to incorporate multipole electrostatics or the use of ML forcefields, or to re-rank candidate structures using DFT.

### Sampling Crystal Landscape

To assess whether sampling is complete, one approach is to inspect if late-stage CSP runs continue to yield new minima or predominantly relax into previously identified energy wells. This evaluation helps determine whether additional sampling such as incorporating more conformers or generating more structures is warranted for a challenging flexible target. The strategy can also be tailored by space group, enabling more efficient allocation of computational resources.

### Post Crystal Structure Prediction Optimisation

Taking a subset of the low-energy crystals from the crystal landscape through post-CSP optimisation was essential to obtain close matches with experimentally observed structures. In several systems, substantial re-ranking of structures was observed; therefore, generous energy cut-offs on the order of tens of kJ $mol^{-1}$, should be employed when selecting structures for higher levels of theory. Prematurely discarding higher-energy candidates risks omitting the experimental form.

Two methodologies were identified as particularly promising for this process. A composite approach combining DFTB with DMACRYS was found to generate structures that were geometrically similar to experimental structures. However, the energy rankings of these structures relative to the broader landscape were only moderately reliable. In contrast, the use of a higher-level method such as periodic DFT provided both strong geometric agreement and improved energy ranking. Nevertheless, this approach incurred a significantly higher computational cost.

### Utilising Experimental Data

If experimental data is available, it can be used to validate predicted structures. One such approach involves using a DTW algorithm to measure the dissimilarity between two PXRDs. When PXRD patterns are generated from hypothetical crystals, this comparison can be performed with relative ease. However, patterns that do not visually

match can still yield low DTW distances. Comparing structures across a range of different Sakoe–Chiba bands offers some improvement but often requires manual intervention and has limited effectiveness.

The newly developed MCSA method demonstrates significant promise as a tool for determining molecular crystal structures by integrating experimental and computational techniques. It is robust enough to reliably resolve the crystal structures of rigid molecules and distinguish polymorphs formed under different conditions, such as varying pressure. For the method to be truly effective, the PXRD data must be free from significant preferred orientation. Therefore, for an automated synthesis and PXRD workflow to be feasible, preferred orientation should be minimised. Despite this limitation, the method successfully predicted the $\alpha$ polymorph of benzimidazole from 6 out of 8 samples, which is encouraging. Nonetheless, the method currently struggles to determine the structures of flexible molecules with the same reliability. Of the two flexible molecules investigated, only one could be determined consistently. Further development is required, including testing a wider range of molecules with different numbers of flexible torsions, and optimising parameters such as trajectory types, step sizes, and temperatures. A significant parameter space remains unexplored, and the method's limitations are not yet fully understood. The approach could allow for the incorporation of multiple experimental data sources simultaneously such as PXRD and NMR with the potential to include additional data types. Provided that a suitable structural similarity metric and a rapid simulation method exist for the data type in question, implementation within our codebase is straightforward.

A slightly different approach utilises MC refinement which facilitates the matching of CSP datasets to experimental data. It has proven successful for both rigid and flexible molecules in the systems tested thus far and could be used towards the end a CSP to determine whether a match has been made where there is ambiguity. The refinement process provides insight into the reliability of structural matches.

## Closing Remarks

The contents of this thesis present promising developments in the advancement of new methods for CSP applications, though further refinement is necessary to realise their full potential.

The integration of conformational sampling strategies, improved post-CSP optimisation techniques, and the incorporation of experimental data into predictive workflows represents a significant step forward in addressing the complexities of flexible pharmaceutical molecules. While the protocols presented here have demonstrated success across a range

of systems, several limitations persist. Among these are the challenges in handling extreme conformational flexibility, reliably predicting rare or metastable polymorphs, and confidently matching predicted structures to experimental data.

Future work should focus on the continued refinement of these techniques, particularly through the adoption of machine learning models to efficiently estimate the energies of crystal structures and to develop methods for ranking structures whose total energy comprises contributions from multiple levels of theory. Parallel efforts should also aim to broaden the integration of experimental data by incorporating additional data types into the CSP process, thereby further enhancing predictive accuracy.

Ultimately, reliable CSP for flexible, drug-like molecules remains a significant challenge. However, the advances made in this thesis represent progress toward establishing CSP as a routine and dependable tool in pharmaceutical solid-form development. As computational capabilities expand and algorithms become more sophisticated, the prospect of a fully predictive, experimentally guided CSP for flexible molecules becomes increasingly attainable.

# Appendix A

# Conformer Search Settings

| **Job Settings** | |
|---|---|
| Sort Z-matrix | F |
| **CRE Settings** | |
| Energy window (kcal) | 6.0000 |
| RMSD threshold (Å) | 0.1250 |
| Energy threshold (kcal) | 0.0500 |
| Rot. const. threshold | 0.01 |
| T (K) (for boltz. weight) | 298.15 |
| **General MD/MTD Settings** | |
| Time step (fs) | 5.0 |
| Shake mode | 2 |
| MTD temperature (K) | 300.00 |
| Trj dump step (fs) | 100 |
| MTD $V_{\text{bias}}$ dump (ps) | 1.0 |
| **XTB Settings** | |
| GFN method | GFN2 |
| Final optimisation level | very tight |

Table A.1: iMTD-GC conformational search settings used throughout.

| **Job Settings** | |
|---|---|
| Sort Z-matrix | F |
| **CRE Settings** | |
| Energy window (kcal) | 6.0000 |
| RMSD threshold (Å) | 0.1250 |
| Energy threshold (kcal) | 0.0500 |
| Rot. const. threshold | 0.01 |
| T (K) (for boltz. weight) | 298.15 |
| **General MD/MTD Settings** | |
| Time step (fs) | 5.0 |
| Shake mode | 2 |
| MTD temperature (k) | 300.00 |
| Trj dump step (fs) | 100 |
| MTD Vbias dump (ps) | 1.0 |
| **XTB Settings** | |
| GFN method | GFN2 |
| Final optimisation level | very tight |

TABLE A.2: iMTD-sMTD conformational search settings used throughout.

| **Parameter** | **Setting** |
|---|---|
| Optimisation Algorithm | Berny Algorithm |
| Convergence Criteria | |
| Maximum Force | $4.5 \times 10^{-4}$ Hartree/Bohr |
| RMS Force | $3.0 \times 10^{-4}$ Hartree/Bohr |
| Maximum Displacement | $1.8 \times 10^{-3}$ Bohr |
| RMS Displacement | $1.2 \times 10^{-3}$ Bohr |
| Step Size | 0.01 Bohr |
| Hessian Update | Broyden–Fletcher–Goldfarb–Shanno (BFGS) update |
| Optimisation | |
| Maximum Steps | 29 |
| Maximum cycles | 500 |
| SCF | |
| Maximum cycles | 65 |
| Level | tight |
| Initial Hessian | Model Hessian (calculated analytically for the first point) |
| Symmetry | Retained during optimisation unless 'NoSymm' keyword is used |
| Charge | 0 |
| Spin Multiplicity | Singlet |

Table A.3: Optimisation Settings used in Gaussian09 for conformer search comparisons

# Appendix B

# Idelalisib Landscapes

## B.1 Idelalisib Experimental Details

| | |
|---|---|
| Empirical formula | $C_{22}H_{18}FN_7O$ |
| Formula weight | 415.43 |
| Temperature / K | 100.00(10) |
| Crystal system | monoclinic |
| Space group | P21 |
| a/Å | 20.8926(17) |
| b/Å | 11.1874(5) |
| c/Å | 21.2003(18) |
| $\alpha$/deg | 90 |
| $\beta$/deg | 116.919(10) |
| $\gamma$/deg | 90 |
| Volume/Å$^3$ | 4418.3(7) |
| Z | 8 |
| $\rho$calcg/cm$^3$ | 1.249 |
| $\mu$/mm 1 | 0.725 |
| F(000) | 1728 |
| Crystal size/mm3 | 0.17 × 0.1 × 0.04 |
| Radiation | Cu K$\alpha$ ($\lambda$ = 1.54184) |
| 2$\theta$ range for data collection/deg | 4.744 to 138.85 |
| Index ranges | -23 ≤ h ≤ 25, -13 ≤ k ≤ 13, -25 ≤ l ≤ 14 |
| Reflections collected | 32992 |
| Independent reflections | 13364 [$R_{int}$ = 0.0530, $R_\sigma$ = 0.0597] |
| Data/restraints/parameters | 13364/1/1122 |
| Goodness-of-fit on F2 | 1.072 |
| Final R indexes [I≥2$\sigma$ (I)] | R1 = 0.0908, wR2 = 0.2483 |
| Final R indexes [all data] | R1 = 0.1115, wR2 = 0.2754 |
| Largest diff. peak/hole / e Å$^{-3}$ | 0.36/-0.37 |
| Flack parameter | -0.08(13) |

TABLE B.1: Crystal data and structure refinement for idelalisib ($C_{22}H_{18}FN_7O$).

## B.2 MC Refinement Parameters

| **Parameter** | **Value** |
|---|---|
| Max. accepted steps | 1000 |
| Starting move scale | 1.0 |
| Move Type | Adaptive |
| Energy evaluation | DFTB |
| $\alpha$ | 1 |
| $\lambda$ / kJ mol$^{-1}$ | 10 |
| Temperature profile | Linear |
| Starting temperature | 0 |
| Final temperature | 0 |
| Reject count limit | 120 |
| Bandwarping limit | 0.5 |
| $2\theta$ range / $^\circ$ | 0-40 |

Table B.2: Monte Carlo refinement parameters for neat idelalisib

## B.3 Idelalisib Solvates

### B.3.1 Pyridine

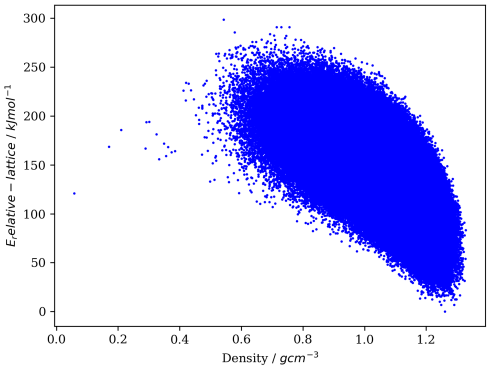



FIGURE B.1: Idelalisib-Pyridine 2:1 Crystal Landscape sampling the 10 most common space groups for Z' > 1.

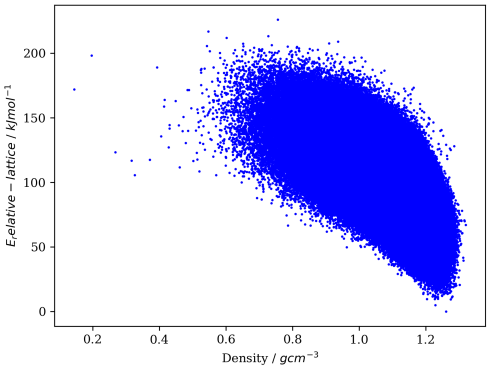



FIGURE B.2: Idelalisib-Pyridine 1:2 Crystal Landscape sampling the 10 most common space groups for Z' > 1.

### B.3.2 Dimethylacetamide

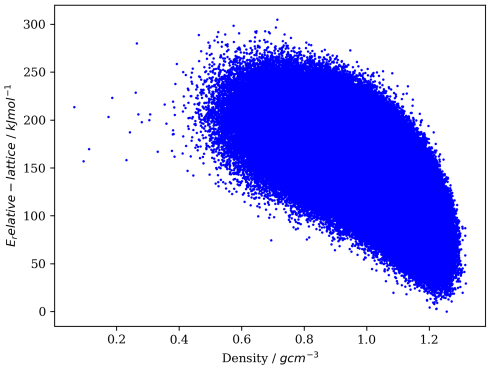



FIGURE B.3: Idelalisib-Dimethylacetamide 2:1 Crystal Landscape sampling the 10 most common space groups for Z' > 1.

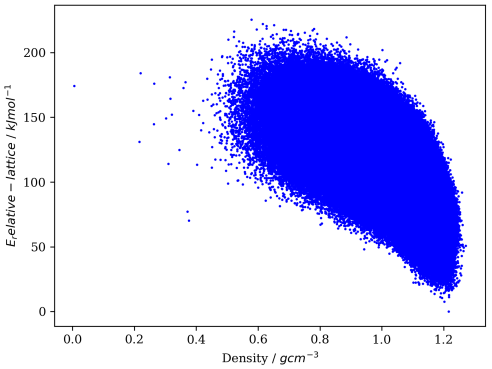



FIGURE B.4: Idelalisib-Dimethylacetamide 1:2 Crystal Landscape sampling the 10 most common space groups for Z' > 1.

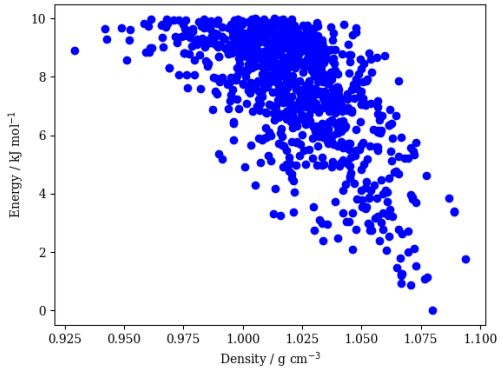



FIGURE B.5: Post crystal structure prediction dimethylacetamide low energy landscape for Z' = 1. Each structure has been optimised using DFTB+ and DMACRYS followed by VASP.

### B.3.3 Acetonitrile

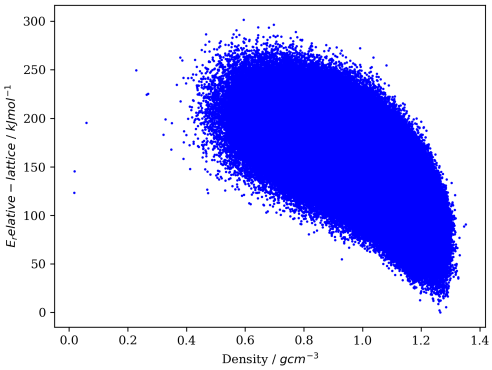



FIGURE B.6: Idelalisib-Acetonitrile 2:1 Crystal Landscape sampling the 10 most common space groups for Z' > 1.

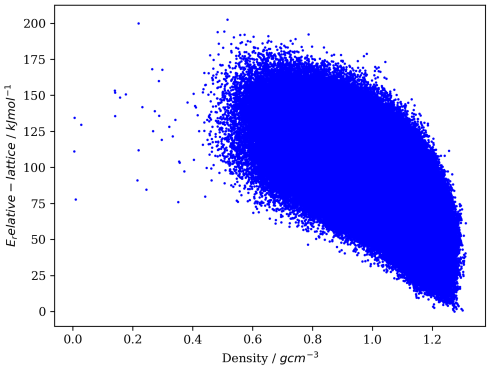



FIGURE B.7: Idelalisib-Acetonitrile 1:2 Crystal Landscape sampling the 10 most common space groups for Z' > 1.

# Bibliography


[1] *mol-CSPy*. URL: https://gitlab.com/mol-cspy/mol-cspy G. M. Day, J. Bramley, P. W. V. Butler, P. J. Bygrave, D. H. Case, C. Y. Cheng, R. Cuadrado, J. Dickman, J. Dorrell, J. Glover, R. Hafizi, J. Johal, D. P. McMahon, J. Nyman, P. Spackman, C. R. Taylor, J. Yang, and S. Yang. *mol-CSPy*.

[2] T. Hahn. *International Tables for Crystallography, Volume A: Space-Group Symmetry*. 5th revised. Dordrecht, Boston, London: Kluwer Academic Publishers, 2002.

[3] R. Hilfiker. *Polymorphism: in the Pharmaceutical Industry*. John Wiley $ Sons, Ltd, 2006.

[4] J. Bernstein. *Polymorphism in Molecular Crystals*. Oxford University Press, 2020. ISBN: 9780199655441.

[5] N. Blagden and et al. "Polymorphism in Pharmaceuticals: Challenges and Opportunities". In: *Advanced Drug Delivery Reviews* 56.3 (2004), pp. 241–274.

[6] J. F. B. Black, P. T. Cardew, A. J. Cruz-Cabeza, R. J. Davey, S. E. Gilks, and R. A. Sullivan. "Crystal nucleation and growth in a polymorphic system: Ostwald's rule, p-aminobenzoic acid and nucleation transition states". In: *CrystEngComm* 20 (6 2018), pp. 768–776.

[7] 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". In: *Acta Crystallographica Section B* 61.5 (2005), pp. 511–527.

[8] P. W. V. Butler and G. M. Day. "Reducing overprediction of molecular crystal structures via threshold clustering". In: *Proceedings of the National Academy of Sciences* 120.23 (2023), e2300516120.

[9] J. Pillardy, Y. A. Arnautova, C. Czaplewski, K. D. Gibson, and H. A. Scheraga. "Conformation-family Monte Carlo: A new method for crystal structure prediction". In: *Proceedings of the National Academy of Sciences* 98.22 (2001), pp. 12351–12356. ISSN: 0027-8424.

[10] Q. Zhu, A. R. Oganov, and A. O. Lyakhov. "Evolutionary metadynamics: a novel method to predict crystal structures". In: *CrystEngComm* 14.10 (2012), p. 3596. ISSN: 1466-8033.

[11] E. C. Dybeck, N. S. Abraham, N. P. Schieber, and M. R. Shirts. "Capturing Entropic Contributions to Temperature-Mediated Polymorphic Transformations Through Molecular Modeling". In: *Crystal Growth & Design* 17.4 (2017), pp. 1775–1787. ISSN: 1528-7483.

[12] A. Gavezzotti. "A solid-state chemist's view of the crystal polymorphism of organic compounds". In: *Journal of Pharmaceutical Sciences* 96.9 (2007), pp. 2232–2241.

[13] S. Grimme, A. Hansen, J. G. Brandenburg, and C. Bannwarth. "Dispersion-Corrected Mean-Field Electronic Structure Methods". In: *Chemical Reviews* 116.9 (2016). PMID: 27077966, pp. 5105–5154.

[14] V. V. Gobre and A. Tkatchenko. "Scaling laws for van der Waals interactions in nanostructured materials". In: *Nature Communications* 4 (2013), p. 2341. ISSN: 2041-1723.

[15] J. Behler and M. Parrinello. "Generalized Neural-Network Representation of High-Dimensional Potential-Energy Surfaces". In: *Physical Review Letters* 98.14 (2007), p. 146401.

[16] P. W. V. Butler, R. Hafizi, and G. M. Day. "Machine-Learned Potentials by Active Learning from Organic Crystal Structure Prediction Landscapes". In: *The Journal of Physical Chemistry A* 128.5 (2024). PMID: 38277275, pp. 945–957.

[17] J. Lommerse, W. Motherwell, H. Ammon, J. Dunitz, A. Gavezzotti, D. Hofmann, F. Leusen, W. Mooij, S. Price, B. Schweizer, M. Schmidt, B. van Eijck, P. Verwer, and D. Williams. "A test of crystal structure prediction of small organic molecules". English. In: *ACTA CRYSTALLOGRAPHICA SECTION B-STRUCTURAL SCIENCE CRYSTAL ENGINEERING AND MATERIALS* 56.4 (2000), pp. 697–714.

[18] W. Motherwell, H. Ammon, J. Dunitz, A. Dzyabchenko, P. Erk, A. Gavezzotti, D. Hofmann, F. Leusen, J. Lommerse, W. Mooij, S. Price, H. Scheraga, B. Schweizer, M. Schmidt, B. van Eijck, P. Verwer, and D. Williams. "Crystal structure prediction of small organic molecules: a second blind test". English. In: *ACTA CRYSTALLOGRAPHICA SECTION B-STRUCTURAL SCIENCE CRYSTAL ENGINEERING AND MATERIALS* 58.4 (2002), pp. 647–661. ISSN: 2052-5206.

[19] A. Gavezzotti and G. Filippini. "Polymorphic forms of organic crystals at room conditions: thermodynamic and structural implications". In: *Journal of the American Chemical Society* 117.49 (1995), pp. 12299–12305.

[20] D. W. Hofmann and J. Apostolakis. "Crystal structure prediction by data mining". In: *Journal of Molecular Structure* 647.1-3 (2003), pp. 17–39.

[21] 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. Wolf, and B. Schweizer. "Significant progress in predicting the crystal structures of small organic molecules - a report on the fourth blind test". English. In: *ACTA CRYSTALLOGRAPHICA SECTION B-STRUCTURAL SCIENCE CRYSTAL ENGINEERING AND MATERIALS* 65.2 (2009), pp. 107–125.

[22] 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". English. In: *ACTA CRYSTALLOGRAPHICA SECTION B-STRUCTURAL SCIENCE CRYSTAL ENGINEERING AND MATERIALS* 67.6 (2011), pp. 535–551.

[23] 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". In: *Acta Crystallographica Section B* 72.4 (2016), pp. 439–459.

[24] H. P. G. Thompson and G. M. Day. "Which conformations make stable crystal structures? Mapping crystalline molecular geometries to the conformational energy landscape". In: *Chem. Sci.* 5 (8 2014), pp. 3173–3182. ISSN: 2041-6520.

[25] D. H. Case, J. E. Campbell, P. J. Bygrave, and G. M. Day. "Convergence Properties of Crystal Structure Prediction by Quasi-Random Sampling". In: *Journal of Chemical Theory and Computation* 12 (2 2016), pp. 910–924.

[26] I. Sobol. "Distribution of Points in a Cube and Approximate Evaluation of Integrals". In: *USSR Computational Mathematics and Mathematical Physics* 7 (4 1967), pp. 86–112.

[27] K. Shoemake. "Uniform Random Rotations". In: *Graphics Gems III*. Ed. by D. Kirk. San Diego, CA, USA: Academic Press Professional, Inc., 1992, pp. 124–132.

[28] C. R. Groom, I. J. Bruno, M. P. Lightfoot, and S. C. Ward. "The Cambridge Structural Database". In: *Acta Crystallographica Section B* 72.2 (2016), pp. 171–179.

[29] S. L. Price. "Why Don't We Find More Polymorphs?" In: *Acta Crystallographica Section B* 69.4 (2013), pp. 313–328.

[30] C. C. Pantelides. "Crystal Structure Prediction Using Genetic Algorithms". In: *Computational Materials Science* 4.3 (1995), pp. 121–125.

[31] S. Yang and G. M. Day. "Global analysis of the energy landscapes of molecular crystal structures by applying the threshold algorithm". In: *Communications Chemistry* 5 (2022), p. 86.

[32] A. Erba. "Hybrid Quantum-Classical Methods for Accurate Crystal Structure Prediction". In: *The Journal of Chemical Physics* 150.2 (2019), p. 024101.

[33] 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". In: *Acta Crystallographica Section B Structural Science, Crystal Engineering and Materials* 72.4 (2016), pp. 439–459. ISSN: 2052-5206.

[34] A. Szabo and N. S. Ostlund. *Modern Quantum Chemistry: Introduction to Advanced Electronic Structure Theory*. Dover Publications, 1989.

[35] C. J. Cramer. *Essentials of Computational Chemistry: Theories and Models*. John Wiley & Sons, 2004.

[36] R. McWeeny. *Methods of Molecular Quantum Mechanics*. Academic Press, 1992.

[37] R. G. Parr and W. Yang. *Density-Functional Theory of Atoms and Molecules*. Oxford University Press, 1989.

[38] A. Szabo and N. S. Ostlund. *Modern Quantum Chemistry: Introduction to Advanced Electronic Structure Theory*. Dover Publications, 1996.

[39] J. C. Slater. "Note on Hartree's Method". In: *Physical Review* 35.2 (1930), p. 210.

[40] S. F. Boys. "Electronic Wave Functions. I. A General Method of Calculation for the Stationary States of Any Molecular System". In: *Proceedings of the Royal Society of London. Series A. Mathematical and Physical Sciences* 200.1063 (1950), pp. 542–554.

[41] W. J. Hehre, R. F. Stewart, and J. A. Pople. "Self-consistent Molecular-orbital Methods. I. Use of Gaussian Expansions of Slater-type Atomic Orbitals". In: *The Journal of Chemical Physics* 51.6 (1969), pp. 2657–2664.

[42] J. A. Pople and W. J. Hehre. *Approximate Molecular Orbital Theory*. John Wiley & Sons, 1971.

[43] W. J. Hehre, R. F. Stewart, and J. A. Pople. "Self-consistent molecular-orbital methods. I. Use of Gaussian expansions of Slater-type atomic orbitals". In: *Journal of Chemical Physics* 51 (1969), p. 2657.

[44] R. Ditchfield, W. J. Hehre, and J. A. Pople. "Self-consistent molecular-orbital methods. IX. An extended Gaussian-type basis for molecular-orbital studies of organic molecules". In: *The Journal of Chemical Physics* 54 (1971), p. 724.

[45] T. Clark, J. Chandrasekhar, G. W. Spitznagel, and P. v. R. Schleyer. "Efficient diffuse function-augmented basis sets for anion calculations. III. The 3-21+G basis set for first-row elements, Li–F". In: *Journal of Computational Chemistry* 4 (1983), pp. 294–301.

[46] M. M. Francl, W. J. Pietro, W. J. Hehre, J. S. Binkley, M. S. Gordon, D. J. DeFrees, and J. A. Pople. "Self-consistent molecular orbital methods. XXIII. A polarization-type basis set for second-row elements". In: *The Journal of Chemical Physics* 77 (1982), p. 3654.

[47] W. J. Hehre, L. Radom, P. v. R. Schleyer, and J. A. Pople. *Ab initio Molecular Orbital Theory*. Wiley, 1976.

[48] J. B. Foresman and A. Frisch. *Exploring Chemistry With Electronic Structure Methods*. 2nd ed. Gaussian, 1996. ISBN: 3825208028.

[49] H. B. Schlegel. "Optimization of equilibrium geometries and transition structures". In: *Journal of Computational Chemistry* 3 (2 1982), pp. 214–218.

[50] B. Hourahine, B. Aradi, V. Blum, F. Bonafé, A. Buccheri, C. Camacho, C. Cevallos, M. Y. Deshaye, T. Dumitrică, A. Dominguez, S. Ehlert, M. Elstner, T. van der Heide, J. Hermann, S. Irle, J. J. Kranz, C. Köhler, T. Kowalczyk, T. Kubař, I. S. Lee, V. Lutsker, R. J. Maurer, S. K. Min, I. Mitchell, C. Negre, T. A. Niehaus, A. M. N. Niklasson, A. J. Page, A. Pecchia, G. Penazzi, M. P. Persson, J. Řezáč, C. G. Sánchez, M. Sternberg, M. Stöhr, F. Stuckenberg, A. Tkatchenko, V. W.-z. Yu, and T. Frauenheim. "DFTB+, a software package for efficient approximate density functional theory based atomistic simulations". In: *The Journal of Chemical Physics* 152 (12 2020), p. 124101. ISSN: 0021-9606.

[51] S. Grimme, C. Bannwarth, and P. Shushkov. "A Robust and Accurate Tight-Binding Quantum Chemical Method for Structures, Vibrational Frequencies, and Noncovalent Interactions of Large Molecular Systems Parametrized for All spd-Block Elements (Z = 1-86)". In: *Journal of Chemical Theory and Computation* 13 (5 2017), pp. 1989–2009. ISSN: 15499626.

[52] C. Bannwarth, S. Ehlert, and S. Grimme. "GFN2-xTB - An Accurate and Broadly Parametrized Self-Consistent Tight-Binding Quantum Chemical Method with Multipole Electrostatics and Density-Dependent Dispersion Contributions". In: *Journal of Chemical Theory and Computation* 15 (3 2019), pp. 1652–1671. ISSN: 15499626.

[53] G. Kresse and J. Furthmüller. "Efficiency of ab-initio total energy calculations for metals and semiconductors using a plane-wave basis set". In: *Computational Materials Science* 6 (1 1996), pp. 15–50. ISSN: 0927-0256.

[54] G. Kresse and J. Hafner. "Ab initio molecular dynamics for liquid metals". In: *Phys. Rev. B* 47 (1 1993), pp. 558–561.

[55] G. Kresse and J. Hafner. "Ab initio molecular-dynamics simulation of the liquid-metal–amorphous-semiconductor transition in germanium". In: *Phys. Rev. B* 49 (20 1994), pp. 14251–14269.

[56] G. Kresse and J. Furthmüller. "Efficient iterative schemes for ab initio total-energy calculations using a plane-wave basis set". In: *Phys. Rev. B* 54 (16 1996), pp. 11169–11186.

[57] 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". In: *Physical Chemistry Chemical Physics* 12 (30 2010), p. 8478. ISSN: 1463-9076.

[58] A. J. Stone and M. Alderton. "Distributed multipole analysis: Methods and applications". In: *Molecular Physics* 100 (1 2002), pp. 221–233. ISSN: 1362-3028.

[59] P. J. Winn, G. G. Ferenczy, and C. A. Reynolds. "Toward Improved Force Fields. 1. Multipole-Derived Atomic Charges". In: *The Journal of Physical Chemistry A* 101 (30 1997), pp. 5437–5445. ISSN: 1089-5639.

[60] G. G. Ferenczy, P. J. Winn, and C. A. Reynolds. "Toward Improved Force Fields. 2. Effective Distributed Multipoles". In: *The Journal of Physical Chemistry A* 101 (30 1997), pp. 5446–5455. ISSN: 1089-5639.

[61] K. Pearson. "On lines and planes of closest fit to systems of points in space". In: *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science* 2.11 (1901), pp. 559–572.

[62] H. Hotelling. "Analysis of a complex of statistical variables into principal components". In: *Journal of Educational Psychology* 24.6 (1933), pp. 417–441.

[63] N. Halko, P. G. Martinsson, and J. A. Tropp. "Finding structure with randomness: Probabilistic algorithms for constructing approximate matrix decompositions". In: *SIAM Review* 53.2 (2011), pp. 217–288.

[64] S. Hayward, A. Kitao, and N. Go. "Harmonicity and anharmonicity in protein dynamics: A normal mode analysis and principal component analysis". In: *Protein Science* 23 (1995), pp. 177–186.

[65] E. Reich, P. Torsky, and M. Warme. "Principal component and clustering analysis on molecular dynamics data of the ribosomal L11 · 23S subdomain". In: *Journal of Molecular Modeling* 21.7 (2015), p. 188.

[66] E. Forgy. "Cluster analysis of multivariate data: efficiency versus interpretability of classifications". In: *Biometrics* 21 (3 1965), pp. 768–769.

[67] S. Lloyd. "Least squares quantization in PCM". In: *IEEE Transactions on Information Theory* 28 (2 1982), pp. 129–137. ISSN: 0018-9448.

[68] J. M. Blaney and J. S. Dixon. "Distance Geometry in Molecular Modeling". In: *Reviews in Computational Chemistry*. John Wiley & Sons, Ltd, 1994, pp. 299–335. ISBN: 9780470125823.

[69] I. Borg and P. J. F. Groenen. "Matrix Algebra for MDS". In: *Modern Multidimensional Scaling: Theory and Applications*. New York, NY: Springer New York, 2005, pp. 137–168. ISBN: 978-0-387-28981-6.

[70] T. A. Halgren. “Merck molecular force field. II. MMFF94 van der Waals and electrostatic parameters for intermolecular interactions”. In: *Journal of Computational Chemistry* 17.5-6 (1996), pp. 520–552.

[71] T. A. Halgren. “Merck molecular force field. III. Molecular geometries and vibrational frequencies for MMFF94”. In: *Journal of Computational Chemistry* 17.5-6 (1996), pp. 553–586.

[72] T. A. Halgren. “Merck molecular force field. I. Basis, form, scope, parameterization, and performance of MMFF94”. In: *Journal of Computational Chemistry* 17.5-6 (1996), pp. 490–519.

[73] T. A. Halgren. “MMFF VI. MMFF94s option for energy minimization studies”. In: *Journal of Computational Chemistry* 20.7 (1999), pp. 720–729.

[74] T. A. Halgren and R. B. Nachbar. “Merck molecular force field. IV. conformational energies and geometries for MMFF94”. In: *Journal of Computational Chemistry* 17.5-6 (1996), pp. 587–615.

[75] A. Laio and M. Parrinello. “Escaping free-energy minima”. In: *Proceedings of the National Academy of Sciences* 99.20 (2002), pp. 12562–12566.

[76] A. Barducci, G. Bussi, and M. Parrinello. “Metadynamics”. In: *Wiley Interdisciplinary Reviews: Computational Molecular Science* 1.5 (2011), pp. 826–843.

[77] G. Fiorin, M. L. Klein, and J. Hénin. “Using collective variables to drive molecular dynamics simulations”. In: *Molecular Physics* 111.22-23 (2013), pp. 3345–3362.

[78] A. Laio and F. L. Gervasio. “Predicting rare events in molecular dynamics”. In: *Reports on Progress in Physics* 71.12 (2008), p. 126601.

[79] A. Barducci, G. Bussi, and M. Parrinello. “Well-Tempered Metadynamics: A Smoothly Converging and Tunable Free-Energy Method”. In: *Phys. Rev. Lett.* 100 (2 2008), p. 020603.

[80] O. Valsson, P. Tiwary, and M. Parrinello. “Enhancing sampling in molecular dynamics simulations with metadynamics, replica-exchange, and variational approaches”. In: *Annual Review of Physical Chemistry* 67 (2016), pp. 159–184.

[81] P. Pracht, F. Bohle, and S. Grimme. “Automated exploration of the low-energy chemical space with fast quantum chemical methods”. In: *Physical Chemistry Chemical Physics* 22 (14 2020), pp. 7169–7192. ISSN: 14639076.

[82] J. Simons, P. Joergensen, H. Taylor, and J. Ozment. “The Journal of Physical Chemistry”. In: *The Journal of Physical Chemistry* 87.15 (1983), pp. 2745–2753.

[83] I. Kolossváry and W. C. Guida. “Low Mode Search. An Efficient, Automated Computational Method for Conformational Analysis: Application to Cyclic and Acyclic Alkanes and Cyclic Peptides”. In: *Journal of American Chemical Society* 118 (21 1996), pp. 5011–5019.

[84] I. Kolossváry and W. C. Guida. "Low-mode Conformational Search Elucidated. Application to C39H80 and Flexible Docking of 9-Deazaguanine Inhibitors to PNP". In: *Journal of Computational Chemistry* 20.13 (1999), pp. 1671–1681.

[85] J. Doye and D. Wales. "Surveying a potential energy surface by eigenvector-following". In: *Zeitschrift für Physik D Atoms, Molecules and Clusters* 40 (1-4 1997), pp. 194–197. ISSN: 0178-7683.

[86] A. Shrake and J. A. Rupley. "Environment and exposure to solvent of protein atoms. Lysozyme and insulin". In: *Journal of Molecular Biology* 79 (2 1973), pp. 351–371. ISSN: 00222836.

[87] G. J. Piermarini, A. D. Mighell, C. E. Weir, and S. Block. "Crystal Structure of Benzene II at 25 Kilobars". In: *Science* 165.3899 (1969), pp. 1250–1255.

[88] H. Sakoe and S. Chiba. "Dynamic programming algorithm optimization for spoken word recognition". In: *IEEE Transactions on Acoustics, Speech, and Signal Processing* 26.1 (1978), pp. 43–49.

[89] J. A. Chisholm and W. D. S. Motherwell. "COMPACK: A program for identifying crystal structure similarity using distances". In: *Journal of Applied Crystallography* 38 (2005), pp. 228–231.

[90] C. P. Brock. "Crystallographic comparison tools in structure validation". In: *Acta Crystallographica Section B: Structural Science* 65.6 (2009), pp. 479–486.

[91] F. H. Allen. "The Cambridge Structural Database: a quarter of a million crystal structures and rising". In: *Acta Crystallographica Section B: Structural Science* 58.3 (2002), pp. 380–388.

[92] S. L. Price. "Predicting crystal structures of organic compounds". In: *Chemical Society Reviews* 38.7 (2009), pp. 2143–2153.

[93] W. D. S. Motherwell et al. "Crystal structure prediction of small organic molecules: techniques and applications". In: *Current Opinion in Solid State and Materials Science* 6.2 (2002), pp. 105–114.

[94] M. Cordova, E. A. Engel, A. Stefaniuk, F. Paruzzo, A. Hofstetter, M. Ceriotti, and L. Emsley. "A Machine Learning Model of Chemical Shifts for Chemically and Structurally Diverse Molecular Solids". In: *The Journal of Physical Chemistry C* 126.39 (2022), pp. 16710–16720.

[95] T. Ottersen, E. Rosenqvist, C. E. Turner, and F. S. El-Feraly. "The Crystal and Molecular Structure of Cannabinol". In: *Acta Chemica Scandinavica, Series B* 31 (1977), pp. 781–787.

[96] C. Parish, R. Lombardi, K. Sinclair, E. Smith, A. Goldberg, M. Rappleye, and M. Dure. "A comparison of the Low Mode and Monte Carlo conformational search methods". In: *Journal of Molecular Graphics and Modelling* 21.2 (2002), pp. 129–150. ISSN: 1093-3263.

[97] J. Goodman and W. Still. “Conformational space sampling of small molecules: Comparison of deterministic and stochastic methods”. In: *Journal of Chemical Information and Computer Sciences* 36.1 (1996), pp. 112–120.

[98] X. Daura and O. Conchillo-Solé. “On Quality Thresholds for the Clustering of Molecular Structures”. In: *Journal of Chemical Information and Modeling* 62.22 (2022), pp. 5738–5745.

[99] *SQLite*. URL: https://www.sqlite.org/index.html R. D. Hipp. *SQLite*. Version 3.31.1. 2020.

[100] F. Mohamadi, N. G. J. Richards, W. C. Guida, R. Liskamp, M. Lipton, C. Caufield, G. Chang, T. Hendrickson, and W. C. Still. “Macromodel—an integrated software system for modeling organic and bioorganic molecules using molecular mechanics”. In: *Journal of Computational Chemistry* 11.4 (1990), pp. 440–467.

[101] K. Sargsyan, J. Wright, and C. Lim. “GeoPCA: a new tool for multivariate analysis of dihedral angles based on principal component geodesics”. In: *Nucleic Acids Research* 40.3 (2012), e25. ISSN: 1362-4962.

[102] K. D. M. Harris, R. L. Johnston, and B. M. Kariuki. “The Genetic Algorithm: Foundations and Apllications in Structure Solution from Powder Diffraction Data”. In: *Acta Crystallographica Section A* 54.5 (1998), pp. 632–645.

[103] O. J. Lanning, S. Habershon, K. D. Harris, R. L. Johnston, B. M. Kariuki, E. Tedesco, and G. W. Turner. “Definition of a ‘guiding function’ in global optimization: a hybrid approach combining energy and R-factor in structure solution from powder diffraction data”. In: *Chemical Physics Letters* 317.3 (2000), pp. 296–303. ISSN: 0009-2614.

[104] A. A. Coelho. “*TOPAS* and *TOPAS-Academic*: an optimization program integrating computer algebra and crystallographic objects written in C++”. In: *Journal of Applied Crystallography* 51.1 (2018), pp. 210–218.

[105] S. Racioppi, A. Otero-de-la-Roza, S. Hajinazar, and E. Zurek. “Powder X-ray diffraction assisted evolutionary algorithm for crystal structure prediction”. In: *Digital Discovery* 4 (2025), pp. 73–83.

[106] G. M. Day, J. van de Streek, A. Bonnet, J. C. Burley, W. Jones, and W. D. S. Motherwell. “Polymorphism of Scyllo-Inositol: Joining Crystal Structure Prediction with Experiment to Elucidate the Structures of Two Polymorphs”. In: *Crystal Growth & Design* 6.10 (2006), pp. 2301–2307.

[107] P. Cui, D. P. McMahon, P. R. Spackman, B. M. Alston, M. A. Little, G. M. Day, and A. I. Cooper. “Mining predicted crystal structure landscapes with high throughput crystallisation: old molecules, new insights”. In: *Chem. Sci.* 10 (43 2019), pp. 9988–9997.

[108] R. A. Mayo, K. M. Marczenko, and E. R. Johnson. "Quantitative matching of crystal structures to experimental powder diffractograms". In: *Chem. Sci.* 14 (18 2023), pp. 4777–4785.

[109] M. Balodis, M. Cordova, A. Hofstetter, G. M. Day, and L. Emsley. "De Novo Crystal Structure Determination from Machine Learned Chemical Shifts". In: *Journal of the American Chemical Society* 144.16 (2022). PMID: 35416661, pp. 7215–7223.

[110] M. D. Eddleston, K. E. Hejczyk, E. G. Bithell, G. M. Day, and W. Jones. "Polymorph Identification and Crystal Structure Determination by a Combined Crystal Structure Prediction and Transmission Electron Microscopy Approach". In: *Chemistry – A European Journal* 19.24 (2013), pp. 7874–7882.

[111] M. D. Eddleston, K. E. Hejczyk, E. G. Bithell, G. M. Day, and W. Jones. "Determination of the Crystal Structure of a New Polymorph of Theophylline". In: *Chemistry – A European Journal* 19.24 (2013), pp. 7883–7888.

[112] M. Baias, C. M. Widdifield, J.-N. Dumez, H. P. G. Thompson, T. G. Cooper, E. Salager, S. Bassil, R. S. Stein, A. Lesage, G. M. Day, and L. Emsley. "Powder crystallography of pharmaceutical materials by combined crystal structure prediction and solid-state 1H NMR spectroscopy". In: *Phys. Chem. Chem. Phys.* 15 (21 2013), pp. 8069–8080.

[113] A. Hofstetter, M. Balodis, F. M. Paruzzo, C. M. Widdifield, G. Stevanato, A. C. Pinon, P. J. Bygrave, G. M. Day, and L. Emsley. "Rapid Structure Determination of Molecular Solids Using Chemical Shifts Directed by Unambiguous Prior Constraints". In: *Journal of the American Chemical Society* 141.42 (2019). PMID: 31117663, pp. 16624–16634.

[114] M. K. Dudek, P. Paluch, and E. Pindelska. "Crystal structures of two furazidin polymorphs revealed by a joint effort of crystal structure prediction and NMR crystallography". In: *Acta Crystallographica Section B* 76.3 (2020), pp. 322–335.

[115] M. Baias, J.-N. Dumez, P. H. Svensson, S. Schantz, G. M. Day, and L. Emsley. "De Novo Determination of the Crystal Structure of a Large Drug Molecule by Crystal Structure Prediction-Based Powder NMR Crystallography". In: *Journal of the American Chemical Society* 135.46 (2013). PMID: 24168679, pp. 17501–17507.

[116] A. M. Lunt, H. Fakhruldeen, G. Pizzuto, L. Longley, A. White, N. Rankin, R. Clowes, B. Alston, L. Gigli, G. M. Day, A. I. Cooper, and S. Y. Chong. "Modular, multi-robot integration of laboratories: an autonomous workflow for solid-state chemistry". In: *Chem. Sci.* 15 (7 2024), pp. 2456–2463.

[117] R. T. Stibrany, J. A. Potenza, and H. J. Schugar. CSD Communication. 2001.

[118] S. Krawczyk and M. Gdaniec. "Polymorph $\beta$ of 1*H*-benzimidazole". In: *Acta Crystallographica Section E* 61.12 (2005), pp. 4116–4118.

[119] W. Zieliński and A. Katrusiak. "Hydrogen Bonds NH · · · N in Compressed Benzimidazole Polymorphs". In: *Crystal Growth & Design* 13.2 (2013), pp. 696–700.

[120] 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". In: *Journal of the American Chemical Society* 122.4 (2000), pp. 585–591.

[121] J. C. Moore, A. Yeadon, and R. A. Palmer. "Crystal and molecular structure of an amber polymorph of 4-methyl-2-nitroacetanilide (MNA)". In: *Journal of Crystallographic and Spectroscopic Research* 14.3 (1984), pp. 283–291. ISSN: 1572-8854.

[122] X. Wang, A. Mueen, H. Ding, and E. Keogh. "A Novel Shapelet-Based Approach for Time Series Classification". In: *Neurocomputing* 112 (2013), pp. 1–11.

[123] A. Mueen, E. Keogh, and J. Lin. "Shapelet-Based Representation and Classification of Time Series". In: *Knowledge and Information Systems* 26.2 (2010), pp. 153–172.

[124] E. Keogh, J. Lin, L. Wei, and C.-K. Lee. "Time Series Shapelet Transform for Efficient and Effective Classification". In: *Knowledge and Information Systems* 28.2 (2011), pp. 181–201.

[125] A. Dukkipati, M. Murty, and S. Bhatnagar. "Cauchy Annealing Schedule: An Annealing Schedule for Boltzmann Selection Scheme in Evolutionary Algorithms". In: *CoRR* cs.AI/0408055 (2004).

[126] R. H. J. M. Otten and L. P. P. P. van Ginneken. *The Annealing Algorithm*. Boston, MA, USA: Kluwer Academic Publishers, 1989.

[127] C. Yuan, J. Wang, Q. Yang, Y. Xu, S. Feng, X. Zhu, and H. Li. "High-pressure polymorphism in amoxicillin". In: *Chemical Physics Letters* 828 (2023), p. 140743. ISSN: 0009-2614.

[128] G. J. O. Beran, I. J. Sugden, C. Greenwell, D. H. Bowskill, C. C. Pantelides, and C. S. Adjiman. "How many more polymorphs of ROY remain undiscovered". In: *Chem. Sci.* 13 (5 2022), pp. 1288–1297.

[129] R. R. Furman, J. P. Sharman, S. E. Coutre, B. D. Cheson, J. M. Pagel, P. Hillmen, J. C. Barrientos, A. D. Zelenetz, T. J. Kipps, I. Flinn, P. Ghia, H. Eradat, T. Ervin, N. Lamanna, B. Coiffier, A. R. Pettitt, S. Ma, S. Stilgenbauer, P. Cramer, M. Aiello, D. M. Johnson, L. L. Miller, D. Li, T. M. Jahn, R. D. Dansey, M. Hallek, and S. M. O'Brien. "Idelalisib and Rituximab in Relapsed Chronic Lymphocytic Leukemia". In: *New England Journal of Medicine* 370.11 (2014), pp. 997–1007.

[130] *ZYDELIG*. URL: https://www.zydelig.com/ (visited on 09/05/2024) *ZYDELIG*.

[131] C. Fonseca Guerra, F. M. Bickelhaupt, S. Saha, and F. Wang. "Adenine Tautomers: Relative Stabilities, Ionization Energies, and Mismatch with Cytosine". In: *The Journal of Physical Chemistry A* 110.11 (2006), pp. 4012–4020. ISSN: 1089-5639.

[132] R. S. Paton and J. M. Goodman. "Hydrogen bonding and pi-stacking: how reliable are force fields? A critical evaluation of force field descriptions of nonbonded interactions". In: *J. Chem. Inf. Model.* 49 (2009), pp. 944–55.

[133] C. D. Kruif and C. V. Ginkel. "Torsion-weighing effusion vapour-pressure measurements on organic compounds". In: *The Journal of Chemical Thermodynamics* 9 (1977), pp. 725–730.

[134] A. H. Jones. "Sublimation Pressure Data for Organic Compounds". In: *Journal of Chemical $ Engineering Data* 5 (1960), pp. 196–200.

[135] L. A. Torres-Gomez, G. Barreiro-Rodriguez, and A. Galarza-Mondragon. "A new method for the measurement of enthalpies of sublimation using differential scanning calorimetry". In: *Thermochimica Acta* 124 (1988), pp. 229–233.

[136] F. Wania, W.-Y. Shiu, and D. Mackay. "Measurement of the Vapor Pressure of Several Low-Volatility Organochlorine Chemicals at Low Temperatures with a Gas Saturation Method". In: *Journal of Chemical $ Engineering Data* 39 (1994), pp. 572–577.

[137] W. J. Sonnefeld, W. H. Zoller, and W. E. May. "Dynamic coupled-column liquid-chromatographic determination of ambient-temperature vapor pressures of polynuclear aromatic hydrocarbons". In: *Analytical Chemistry* 55 (1983), pp. 275–280.

[138] M. Colomina, P. Jimenez, and C. Turrion. "Vapour pressures and enthalpies of sublimation of naphthalene and benzoic acid". In: *The Journal of Chemical Thermodynamics* 14 (8 1982), pp. 779–784.

[139] C. de Kruif, T. Kuipers, J. van Miltenburg, R. Schaake, and G. Stevens. "The vapour pressure of solid and liquid naphthalene". In: *The Journal of Chemical Thermodynamics* 13 (11 1981), pp. 1081–1086.

[140] C. D. Kruif. "Enthalpies of sublimation and vapour pressures of 11 polycyclic hydrocarbons". In: *The Journal of Chemical Thermodynamics* 12 (3 1980), pp. 243–248.

[141] D. Ambrose, I. Lawrenson, and C. Sprake. "The vapour pressure of naphthalene". In: *The Journal of Chemical Thermodynamics* 7 (12 1975), pp. 1173–1176.

[142] J. S. Chickos. "A simple equilibrium method for determining heats of sublimation". In: *Journal of Chemical Education* 52 (1975), p. 134.

[143] D. McEachern, O. Sandoval, and J. C. Iniguez. "The vapor pressures, derived enthalpies of sublimation, enthalpies of fusion, and resonance energies of acridine and phenazine". In: *The Journal of Chemical Thermodynamics* 7 (1975), pp. 299–306.

[144] D. McEachern and O. Sandoval. “A molecular flow evaporation apparatus for measuring vapour pressures and heats of sublimation of organic compounds”. In: *Journal of Physics E: Scientific Instruments* 6 (1973), pp. 155–161.

[145] G. A. Miller. “Vapor Pressure of Naphthalene. Thermodynamic Consistency with Proposed Frequency Assignments”. In: *Journal of Chemical $ Engineering Data* 8 (1963), pp. 69–72.

[146] A. Aihara. “Estimation of the Energy of Hydrogen Bonds Formed in Crystals. I. Sublimation Pressures of Some Organic Molecular Crystals and the Additivity of Lattice Energy”. In: *Bulletin of the Chemical Society of Japan* 32 (1959), pp. 1242–1248.

[147] R. S. Bradley and T. G. Cleasby. “The vapour pressure and lattice energy of some aromatic ring compounds”. In: *Journal of the Chemical Society (Resumed)* (1953), p. 1690.

[148] P. C. Hansen and C. A. Eckert. “An Improved Transpiration Method for the Measurement of Very Low Vapor Pressure.” In: *Journal of Chemical & Engineering Data* 31 (1986), pp. 1–3.

[149] B. F. Rordorf. “Thermal properties of dioxins, furans and related compounds”. In: *Chemosphere* 15 (1986), pp. 1325–1332.

[150] R. Bender, V. Bieling, and G. Maurer. “The vapour pressures of solids: anthracene, hydroquinone, and resorcinol”. In: *The Journal of Chemical Thermodynamics* 15 (1983), pp. 585–594.

[151] J. W. Taylor and R. J. Crookes. “Vapour pressure and enthalpy of sublimation of 1,3,5,7-tetranitro-1,3,5,7-tetra-azacyclo-octane (HMX)”. In: *”J. Chem. Soc., Faraday Trans. 1”* 72 (1976), p. 723.

[152] L. Malaspina, R. Gigli, and G. Bardi. “Microcalorimetric determination of the enthalpy of sublimation of benzoic acid and anthracene”. In: *The Journal of Chemical Physics* 59.1 (1973), pp. 387–394. ISSN: 0021-9606.

[153] G. Beech and R. M. Lintonbon. “The measurement of sublimation enthalpies by differential scanning calorimetry”. In: *Thermochimica Acta* 2 (1971), pp. 86–88.

[154] J. D. Kelley and F. O. Rice. “The Vapor Pressures of Some Polynuclear Aromatic Hydrocarbons”. In: *The Journal of Physical Chemistry* 68 (1964), pp. 3794–3796.

[155] G. W. Sears and E. R. Hopke. “Vapor Pressures of Naphthalene, Anthracene and Hexachlorobenzene in a Low Pressure Region”. In: *Journal of the American Chemical Society* 71 (1949), pp. 1632–1634.

[156] M. V. Roux, M. Temprado, J. S. Chickos, and Y. Nagano. “Critically Evaluated Thermochemical Properties of Polycyclic Aromatic Hydrocarbons”. In: *Journal of Physical and Chemical Reference Data* 37 (2008), p. 1855.

[157] K. Nass, D. Lenoir, and A. Kettrup. "Calculation of the Thermodynamic Properties of Polycyclic Aromatic Hydrocarbons by an Incremental Procedure". In: *Angewandte Chemie International Edition in English* 34 (1995), pp. 1735–1736.

[158] H. Inokuchi, S. Shiba, T. Handa, and H. Akamatu. "Heats of Sublimation of Condensed Polynuclear Aromatic Hydrocarbons". In: *Bulletin of the Chemical Society of Japan* 25 (1952), pp. 299–302.

[159] N. Wakayama and H. Inokuchi. "Heats of Sublimation of Polycyclic Aromatic Hydrocarbons and Their Molecular Packings". In: *Bulletin of the Chemical Society of Japan* 40 (1967), pp. 2267–2271.

[160] C. Lenchitz and R. W. Velicky. "Vapor pressure and heat of sublimation of three nitrotoluenes". In: *Journal of Chemical $ Engineering Data* 15 (1970), pp. 401–403.

[161] R. Jochems, H. Dekker, C. Mosselman, and G. Somsen. "The use of the LKB 8721-3 Vaporization calorimeter to measure enthalpies of sublimation The enthalpies of sublimation of bicyclo[2.2.1]hept-2-ene (norbornene), bicyclo[2.2.1]heptane (norbornane), and tricyclo[3.3.1.13,7]decane (adamantane)". In: *The Journal of Chemical Thermodynamics* 14 (1982), pp. 395–398.

[162] T. Clark, T. M. Knox, M. A. McKervey, H. Mackle, and J. J. Rooney. "Thermochemistry of bridged-ring substances. Enthalpies of formation of some diamondoid hydrocarbons and of perhydroquinacene. Comparisons with data from empirical force field calculations". In: *Journal of the American Chemical Society* 101 (1979), pp. 2404–2410.

[163] T. Clark, T. Knox, H. Mackle, M. A. McKervey, and J. J. Rooney. "Heats of sublimation of some cage hydrocarbons by a temperature scanning technique". In: *Journal of the Chemical Society, Faraday Transactions 1: Physical Chemistry in Condensed Phases* 71 (1975), p. 2107.

[164] R. H. Boyd, S. N. Sanwal, S. Shary-Tehrany, and D. McNally. "The Thermochemistry, Thermodynamic Functions, and Molecular Structures of Some Cyclic Hydrocarbons". In: *The Journal of Physical Chemistry* 75 (1971), pp. 1264–1271.

[165] R. Butler, A. Carson, P. Laye, and W. Steele. "The enthalpy of formation of adamantane". In: *The Journal of Chemical Thermodynamics* 3 (1971), pp. 277–280.

[166] P.-J. Wu, L. Hsu, and D. A. Dows. "Spectroscopic Study of the Phase Transition in Crystalline Adamantane". In: *The Journal of Chemical Physics* 54 (1971), p. 2714.

[167] W. K. Bratton, I. Szilard, and C. A. Cupas. "Enthalpy of sublimation of adamantane". In: *The Journal of Organic Chemistry* 32 (1967), pp. 2019–2021.

[168] P. M. Burkinshaw and C. T. Mortimer. "Enthalpies of sublimation of transition metal complexes". In: *Journal of the Chemical Society, Dalton Transactions* (1984), p. 75.

[169] A. S. Christensen, T. Kubař, Q. Cui, and M. Elstner. "Semiempirical Quantum Mechanical Methods for Noncovalent Interactions for Chemical and Biochemical Applications". In: *Chemical Reviews* 116.9 (2016). PMID: 27074247, pp. 5301–5337.

[170] 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". In: *Journal of Chemical Theory and Computation* 4.10 (2008). PMID: 26620182, pp. 1795–1805.

[171] *Idelalisib*. URL: https://www.selleckchem.com/products/cal-101.html selleckchem.com. *Idelalisib*. Last accessed 10 February 2023. 2023.

[172] C. Hamilton. Private Communication. 2022.

[173] G. M. Sheldrick. "*SHELXT* – Integrated space-group and crystal-structure determination". In: *Acta Crystallographica Section A* 71.1 (2015), pp. 3–8.

[174] G. M. Sheldrick. "Crystal structure refinement with *SHELXL*". In: *Acta Crystallographica Section C* 71.1 (2015), pp. 3–8.

[175] O. V. Dolomanov, L. J. Bourhis, R. J. Gildea, J. A. K. Howard, and H. Puschmann. "*OLEX2*: a complete structure solution, refinement and analysis program". In: *Journal of Applied Crystallography* 42.2 (2009), pp. 339–341.

[176] A. J. Cruz-Cabeza, S. Karki, L. Fábián, T. Friščić, G. M. Day, and W. Jones. "Predicting stoichiometry and structure of solvates". In: *Chem. Commun.* 46 (13 2010), pp. 2224–2226.

[177] D. Mootz and H.-G. Wussow. "Crystal structures of pyridine and pyridine trihydrate". In: *The Journal of Chemical Physics* 75.3 (1981), pp. 1517–1522. ISSN: 0021-9606.

[178] S. Crawford, M. T. Kirchner, D. Bläser, R. Boese, W. I. F. David, A. Dawson, A. Gehrke, R. M. Ibberson, W. G. Marshall, S. Parsons, and O. Yamamuro. "Isotopic Polymorphism in Pyridine". In: *Angewandte Chemie International Edition* 48.4 (2009), pp. 755–757.

[179] M. Podsiadło, K. Jakóbek, and A. Katrusiak. "Density, freezing and molecular aggregation in pyridazine, pyridine and benzene". In: *CrystEngComm* 12 (9 2010), pp. 2561–2567.

[180] N. Giordano, C. M. Beavers, B. J. Campbell, V. Eigner, E. Gregoryanz, W. G. Marshall, M. Peña-Álvarez, S. J. Teat, C. E. Vennari, and S. Parsons. "High-pressure polymorphism in pyridine". In: *IUCrJ* 7.1 (2020), pp. 58–70.

[181] A. Olejniczak and A. Katrusiak. "Supramolecular Reaction between Pressure-Frozen Acetonitrile Phases $\alpha$ and $\beta$". In: *The Journal of Physical Chemistry B* 112.24 (2008). PMID: 18491934, pp. 7183–7190.

[182] S. Kim, A. M. Orendt, M. B. Ferraro, and J. C. Facelli. "Crystal structure prediction of flexible molecules using parallel genetic algorithms with a standard force field". In: *Journal of Computational Chemistry* 30.13 (2009), pp. 1973–1985.

[183] P. W. V. Butler, R. Hafizi, and G. M. Day. "Machine-Learned Potentials by Active Learning from Organic Crystal Structure Prediction Landscapes". In: *J. Phys. Chem. A* 128 (2024), pp. 945–957.

[184] X. Li. Private Communication. 2021.

[185] N. J. Harris and K. Lammertsma. "Ab Initio Density Functional Computations of Conformations and Bond Dissociation Energies for Hexahydro-1,3,5-trinitro-1,3,5-triazine". In: *Journal of the American Chemical Society* 119 (28 1997), pp. 6583–6589.

[186] A. I. Kitaigorodsky. *Molecular crystals and molecules*. Vol. 29. Academic Press, 1973, p. 553. ISBN: 0124105505.