ASSESSMENT AND IMPROVEMENT OF STRUCTURE GENERATION IN CRYSTAL STRUCTURE PREDICTION PROTOCOLS

by

Sarah M. Clarke

Submitted in partial fulfillment of the requirements
for the degree of Master of Science

at

Dalhousie University
Halifax, Nova Scotia
March 2023

CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBREVIATIONS AND SYMBOLS USED

Abbreviation	Description
AM1-BCC	Austin Model 1 with Bond Charge Corrections
AMBER	Assisted Model Building with Energy Refinement
B3LYP	Becke's 3-parameter exchange and Lee-Yang-Parr correlation functional
B86bPBE	Becke's B86b exchange and Perdew-Burke-Ernzerhof correlation functional
BT	Blind Test for crystal structure prediction
CCDC	Cambridge Crystallographic Data Centre
CHARMM	Chemistry at Harvard Macromolecular Mechanics
CSP	Crystal Structure Prediction
DFA	Density-Functional Approximation
DFT	Density-Functional Theory
DMA	Distributed Multipole Analysis
FF	Force Field
FFAT	Force Field Atom Type
FHI-aims	Fritz Haber Institute *ab initio* materials simulations
GAFF	Generalized AMBER Force Field
GGA	Generalized Gradient Approximation
GULP	General Utility Lattice Program
HF	Hartree-Fock Theory
LDA	Local Density Approximation
LJ	Lennard-Jones
MAE	Mean Absolute Error
MD	Molecular Dynamics
MM	Molecular Mechanics
MM2	Allinger's Second Molecular Mechanics FF
MM3	Allinger's Third Molecular Mechanics FF
MMFF94	1994 Merck Molecular Force Field
MP2	Møller Plesset Second-Order Perturbation Theory
MP4	Møller Plesset Fourth-Order Perturbation Theory

Abbreviation	Description
OPLS	Optimized Potentials for Liquid Simulations
OPLS-AA	All Atom version of OPLS
PAW	Projector Augmented-Wave
POWDIFF	Powder Pattern Difference
PV17	Polymorph Vibration benchmark containing 17 compounds
PXRD	Powder X-Ray Diffraction
QE	Quantum ESPRESSO
QM	Quantum Mechanical
RESP	Restrained Electrostatic Potential
RMSD	Root-Mean-Square Deviation
UFF	Universal Force Field
USPEX	Universal Structure Predictor: Evolutionary Xtallography
VC	Variable Cell
VC-PWDF	Variable Cell Powder Difference
vdW	van der Waals
XC	Exchange-Correlation
XDM	Exchange-Hole Dipole Moment Dispersion Model

Symbol	Description
A, B	Molecule labels
A, B, C	Empirical parameters
a, b	Empirical parameters
$\mathbf{a}$	Lattice vector
$c_{fg}(\delta)$	Cross-correlation function
C_6	Dispersion coefficient
C_n	Fourier expansion coefficients
d	Fingerprint distance
D_e	Dissociation energy
E	Energy
f	Scaling factor
f	Fitness
$f(R)$	Fingerprint function
$f(x)$	Continuous function

Symbol	Description
$g(x)$	Continuous function
G	Empirical parameter
$\hat{H}$	Hamiltonian operator
k	Harmonic spring constant
i, j	Atom indices
J	Coulomb repulsion energy
K	Force constant
L	Length
m, n, N	Integer numbers
Q_i^t,Q_j^u	Multipole moments
q_i, q_j	Atomic charges
r, r_{ij}, R	Interatomic distances
$\mathbf{r}$	Electron position
$\hat{T}$	Kinetic energy operator
$T_\circ$	Kinetic energy
T_{ij}^{tu}	Multipole interaction tensor
t, u	Multipole orders
$\hat{U}$	Electron-electron potential energy operator
V	Potential
$\hat{V}$	Electron-nuclear potential energy operator
W	Gaussian height
x	Position variable
Z	Atomic number
Z	Number of molecules per unit cell
α	Atomic polarizability
β, θ, ϕ	Angles
δ	Buffering constant
δ	Offset shift
$\delta(R)$	Delta function
ϵ	Well depth
ε	Dielectric constant
μ	Reduced mass
ρ	Electron density
σ	Collision diameter

Symbol	Description
σ	Gaussian width
φ	Dihedral angle
ψ	Wavefunction
ω	Vibrational frequency

ABSTRACT

The prediction of 3D crystal structures of a compound solely from its single-molecule structure is one of the major challenges of physical chemistry. Crystal structure prediction, or CSP, refers to the ability to predict these structures computationally, without additional input from experiments. This is an important field for industries involving the discovery and screening of solids such as pharmaceuticals, organic electronics, and dyes. The forefront method for performing CSP involves a hierarchical approach, where structures are ranked, pruned, and re-ranked with increasing levels of theory. When structures are first generated, the initial energy ranking method must be computationally efficient enough to relax the geometries for thousands of structures in a reasonable time frame. Force field (FF) methods, therefore, have been frequently used during this stage of CSP.

Advances in CSP are tracked by the Cambridge Crystallographic Data Centre's series of blind test competitions, and a full CSP analysis of methyl anthranilate was completed and submitted as part of the seventh blind test. With our methods, the crystal packing of the predicted structure shared many similarities with the experimental one, although the structures were not identical. Based on the result for the blind test, a benchmark of accessible FFs was completed to test their performance in the structure generation stage of CSP. Here, the FFs were coupled with an evolutionary algorithm structure generator and assessed on their ability to identify the experimental polymorphs of compounds comprising the PV17 benchmark set, plus 5-fluorouracil. The performance was determined not only by how many polymorph matches were found, but by the relative energies of these matches as well. It was concluded that the generalized AMBER force field (GAFF) is the optimal choice for CSP at this time, based on the high rate of polymorph matches with low relative energies. The results from this benchmark identify key FF features that would be necessary for a successful CSP protocol.

ACKNOWLEDGEMENTS

This work would not have been possible without the support from my supervisor, Dr. Erin Johnson, and the blind test team with whom I continued to work closely with. Having had little experience in the computational world prior to starting, I've learned so much working with Erin. I'd also like to specially acknowledge Dr. Alberto Otero-de-la-Roza, our collaborator, without whom I would have lost my mind script-writing ages ago.

A big thank you as well to the Johnson group members past and present. The variety of knowledge and perspective you each bring to the group is exceptional. I particularly want to thank Alex, who was the first person to show me the ropes, and was instrumental to this project for making sense of the thousands and thousands of structures I sent his way. I'm also incredibly grateful to have had such amazing office mates (and friends) this past year. There's no one I'd rather sit at a computer for hours in a concrete room with!

I want to extend my sincere thanks to my family and my friends, although at this point – same difference. I wouldn't be anywhere without their constant love and support, and I'm so lucky to have such an amazing support system surrounding me. Special thanks to Joel for helping me make pretty figures even though he doesn't know what's going on in them, and to Noah upon whom I've become reliant for having a nightly debrief. Finally, 친한친구 진심으로 감사합니다.

CHAPTER 1

INTRODUCTION

For many industries the potential for compounds to exhibit polymorphism, or crystallize into multiple different structures, can be both a difficulty and an opportunity. Different polymorphs of the same material can have vastly different chemical and physical properties. Molecular crystals are of particular interest for their applications in pharmaceuticals, organic electronics, dyes, and other industries.[1] These are crystals where the repeating asymmetric unit is an organic molecule, and are the type of solid considered throughout this work. For pharmaceuticals in particular, it is estimated that up to 80% of marketed pharmaceuticals have polymorphs that can be experimentally produced.[2] Pharmaceutical function is highly dependent on the adopted crystal structure, so polymorph conversion during the life cycle of a drug can prove disastrous for consumer and company alike. In 1998, production of the antiviral compound ritonavir (marketed as Norvir) was halted when many samples of the capsule form failed dissolution tests. This was caused by conversion of the crystal into a new polymorph, now known as form II, which, due to decreased solubility, had a much lower bioavailability. The discovery of this form meant the drug could no longer be stored or administered in the same way that was intended for form I.[3]

The discrepancy between forms I and II of ritonavir, or any given pair of crystalline polymorphs, is due to the interplay of thermodynamics and kinetics. The polymorph with the global minimum free energy, or thermodynamic minimum, is in theory the most likely polymorph to form. Kinetic factors, however, can lead to the growth of a local minimum polymorph which can eventually transform to the thermodynamically favourable form.[4] Molecular crystals can have numerous potential polymorph forms, so using traditional lab methods for studying them all can quickly become an arduous, time-consuming, and environmentally unfriendly task. Herein lies the problem that faces manufacturers and materials scientists – there must be an effective method for polymorph screening that does not solely rely on experiment. To do this, computational methods can be employed to assess the thermodynamic properties or relative energies of crystalline solids.

In response to this issue, the field of crystal structure prediction (CSP) has emerged. Referring to the ability to predict the 3D structure of a compound based solely on its 2D molecular structure, CSP is a constant work-in-progress that can have huge impacts on any industry where molecular crystal polymorphism is a concern. While CSP is not yet advanced enough to replace experimental screening entirely, it can be a powerful tool for many aspects of the manufacturing process. For simplicity, pharmaceutical manufacturing will be discussed. As a complement to experimental screening, CSP enhances the drug discovery process by identifying novel crystal forms, or predicting whether an unwanted polymorph is likely to arise. A full CSP analysis of a molecule yields a crystal energy landscape, where the relative energies and densities of potential structures are plotted against each other. Based on the landscape, structures with low energies and high densities are most likely to be experimentally isolable and the thermodynamic minimum can be identified. In lieu of preparing and analyzing each potential structure experimentally, the results from a CSP search can both point researchers to polymorphs of interest, and rationalize existing experimental knowledge.[5]

Performing full CSP is no trivial endeavour, and there are three main stages in a typical CSP protocol with their own unique challenges. Beginning from the 2D molecular structure, the first stage is to assess the possible molecular conformations. This is no problem if the molecule is rigid. Useful manufacturing targets, however, are becoming larger and more complex, so the molecules subjected to CSP can have a multitude of different conformations requiring consideration. Consequently, this affects the second stage of CSP – structure generation. In the structure generation stage, care must be taken to ensure all potentially relevant crystal phases are "found" by the program used. The complexity of the high-dimensional crystal landscape is due to degrees of freedom associated with the unit cell, molecule position, and flexibility.[6] The high-dimensional landscape is essentially a potential energy surface (PES) where there are $6 + 3\,(N - 1)$ dimensions for N number of atoms, attributed to the coordinates of each atom and the six lattice parameters.[7] The analytic form of the PES is unknown; therefore, adequate sampling is necessary to find the global minimum. There are many existing programs for this stage, using methods including random sampling, genetic algorithms, or machine learning.[7–9] Basin-hopping, simulated annealing, and metadynamics are other methods that exist and are used, but become less useful as molecule size increases.[10] CSP structure generation suffers from the "curse of dimensionality", so true random sampling is insufficient even for the simplest of systems. Thus, genetic algorithms have become popular for their effective sampling while still minimizing computational cost.[11]

Throughout the generation stage using the methods employed in this work, an evolutionary algorithm is paired with an external program to assess the energy of each structure produced. This is related to the final challenge in CSP, the energy ranking, where structures are ranked, pruned, and

re-ranked at increasingly accurate levels of theory. Throughout the protocol there must be a careful balance maintained between assessing the energies with sufficient accuracy to ensure that true low-energy structures are carried forward, however, without wasting resources performing high-level density-functional theory (DFT) calculations on unstable structures that will be eventually discarded. To do this, a classical force field (FF) method is used to perform geometry optimizations during structure generation. Because of the high number of structures generated, a lower level of theory is necessary at this stage. In this work alone, over one million structures were generated and ranked using force field methods. While the relative free energies determine the thermodynamically favourable structure, the electronic lattice energies are typically used in lieu of free energies to forgo computing costly vibrational energy contributions.[12] These contributions are typically small, < 2 kJ/mol, but can still result in energetic reordering of polymorphs. Irrespective of the vibrational contribution, energy differences between isolable polymorphs are quite small – less than 10 kJ/mol, with 80% of polymorphs separated by 4.2 kJ/mol or less.[1] Thus, the methods considered must provide relative energies that are sufficiently accurate to meet this threshold. Dispersion-corrected density-functional theory is one such method that is ubiquitous in computational chemistry, and has been proven to be extremely effective at assessing the lattice energies of molecular crystals.[4,13]

Given the increasing complexity of CSP targets and the wide variety of methods that can be employed, progress in the field has been tracked by a series of blind tests (BTs) organized by the Cambridge Crystallographic Data Centre (CCDC).[6,14–18] Target molecules with unpublished, but known, crystal structures are provided to researchers, who then have a set amount of time to submit their best predictions for review. There have been seven tests since 1999, which have highlighted the many improvements CSP methods have undergone over the last 25 years, and exposed problems that still require work. In the sixth blind test, all experimental structures were found by at least one of the teams, except for one particularly disordered polymorph. While the success rate of the sixth blind test was comparable to the fifth (36/70 and 24/68 successful attempts, respectively), the targets in the sixth test were more complex.[6] As mentioned, there has been great success in the accurate assessment of polymorph lattice energies, albeit at a high computational cost. The most significant challenge seen in the sixth blind test was dealing with conformational flexibility. With each blind test, the targets have grown from simple rigid and partially flexible molecules to salts, co-crystals, metal-containing, and generally larger systems.[6] The increasing complexity of target systems is the driving force behind the difficulty of CSP, as more robust methods are necessary to effectively model the greater diversity of atoms and interactions.[12] The targets from the seventh, and most recent blind test (BT7) are shown in Figure 1.1.

Detailed in this thesis are two projects – a full CSP analysis of compound XXIX for submission

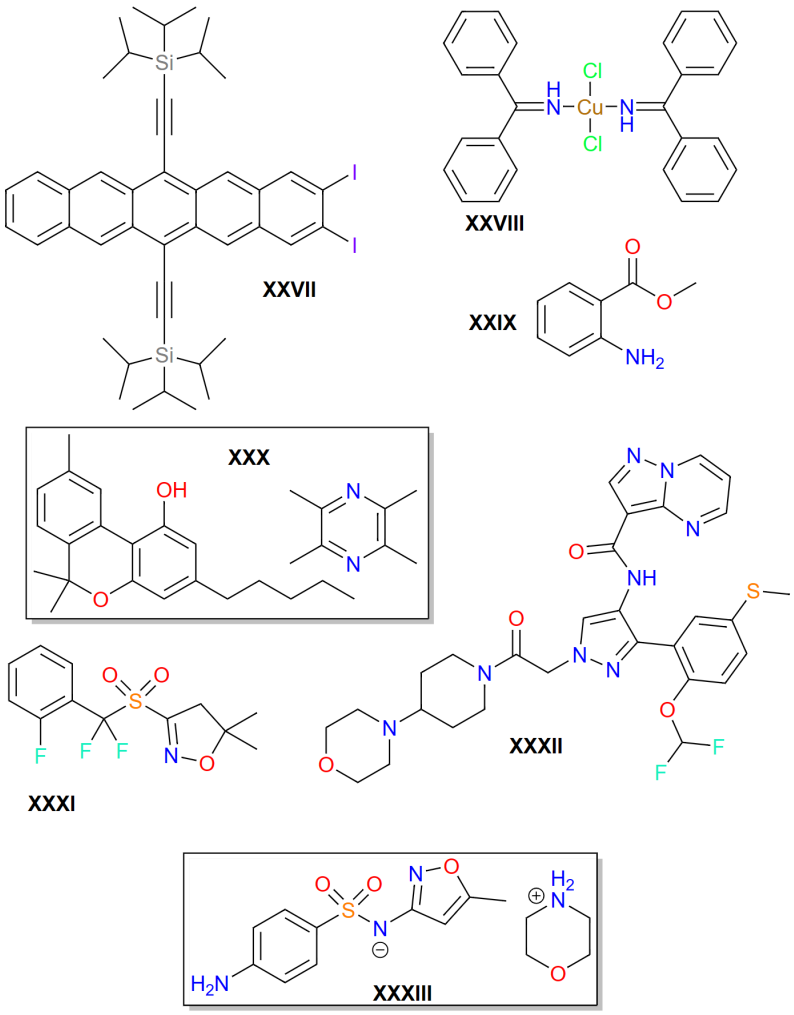


Figure 1.1: Target compounds in the 7th CCDC blind test.

to the 7th blind test, and a benchmark of classical force field methods to improve the structure generation stage. The BT7 submission marked the first time that CSP analysis was performed from start to finish within the Johnson group, and it applied a hierarchical approach combining force fields and dispersion-corrected DFT. We saw that, while our energy-ranking methods were sound, the search space and efficacy needed to be improved. This provided the basis and motivation for the second project, the structure generation benchmark. Here, a collection of force fields were combined with USPEX,[19–22] a popular program that employs an evolutionary algorithm to generate crystal structures for CSP. Each of these force fields were assessed on their ability to find each experimental polymorph pair for a set of 18 small organic molecules (consisting of the PV17 phonon benchmark set[23] and 5-fluorouracil[1]). Force fields are designed to reproduce experimental or quantum-mechanical (QM) results using classical physics approximations, and use parameter-fitting to adjust the description of atoms in different chemical environments. Thus, the

quality of data used to parameterize the force field or the specificity of how “chemical environment" is defined contributes directly to the accuracy of the method. For example, having a different parameter description of an oxygen atom for each functional group it participates in should be more accurate than a generic description for all sp^2-hybridized oxygen atoms. Each of the force fields used will be described in Chapter 2, so the differences in construction can be understood before the results of the force field benchmark are discussed in Chapter 3. Of the seven force fields investigated, those with the most robust description of electrostatic interactions or sophisticated parameterization performed best.

CHAPTER 2

THEORY AND COMPUTATIONAL METHODS

2.1 EMPIRICAL FORCE FIELDS

At their core, force field (FF) methods approximate forces between atoms or molecules using classical physics. These methods use a combination of a functional form and fitted parameters to estimate the potential energy. While many force fields may use similar functional forms, the combination of the functional and parameters are what differentiate FFs that are tailored for different uses. The following general description of FFs has been written with reference to *Molecular Modelling: Principles and Applications* by Andrew R. Leach.[24]

For all functional forms, the bonded and non-bonded energy terms are added together to give the total force-field energy:

$$E_{\mathrm{FF}} = E_{\mathrm{bonded}} + E_{\mathrm{non\text{-}bonded}} \,. \tag{2.1}$$

The bonded terms include bond stretching, angle bending, and torsion,

$$E_{\mathrm{bonded}} = E_{\mathrm{stretching}} + E_{\mathrm{bending}} + E_{\mathrm{torsion}} \,, \tag{2.2}$$

while the non-bonded terms include the electrostatic and van der Waals interactions,

$$E_{\mathrm{non\text{-}bonded}} = E_{\mathrm{elec}} + E_{\mathrm{vdW}} \,. \tag{2.3}$$

While specific implementations of the functional form differ between force fields, there are certain approximations that unite them, particularly concerning the bonded contribution. It is important to note that, while many FF methods may use similar functional forms for the various energy terms, they are inherently empirical and there is no "correct" form for any particular term.

Because of these different functional forms, the total energy values cannot be compared between different force fields, only the relative energies.

One approximation for potential energy, V, is written as a function of distance between atoms, r. Given the shape of the potential energy curve for bond dissociation, the Morse potential,

$$V(r) = D_e \left(1 - e^{-a(r-r_0)}\right)^2 , \tag{2.4}$$

can be used. Here, D_e is the dissociation energy that indicates the difference in energy between the minimum energy at equilibrium bond distance, r_0, and the energy in the dissociation limit. The spring constant, k, is incorporated in the potential through the width of the potential well a,

$$a = \sqrt{\frac{k}{2D_e}} , \tag{2.5}$$

and the harmonic vibrational frequency for the bond is $\omega = \sqrt{\frac{k}{\mu}}$, where μ is the reduced mass.

While the Morse potential is effective at modelling bonds from equilibrium to dissociation, this is not the optimal approximation implemented in FFs. In molecular mechanics, bonds are typically oscillating close to their equilibrium distance, so the simpler harmonic oscillator approximation can be applied without sacrificing accuracy. Given Hooke's law, the equation

$$E_{\text{stretching}} = \frac{k}{2} (r - r_0)^2 \tag{2.6}$$

can be used to simply model bond vibrations in ground-state molecules near equilibrium. Higher-order terms can be included for more effective approximations as the bond separation increases. As seen in Figure 2.1, this can fit the Morse potential more closely, however a cubic expansion alone can be problematic due to the drop in energy past the maximum.

Bond angle bending is typically described in a similar way using Hooke's law,

$$E_{\text{bending}} = \frac{k}{2} (\theta - \theta_0)^2 , \tag{2.7}$$

where k is once again a force constant and θ_0 is the equilibrium value for the angle. Similarly, higher-order terms can be included to improve accuracy or consider cases such as extremely strained molecules.

Next is the torsional term which models how energy changes with bond rotation. This term is closely related to the non-bonded term, and is considered "soft" compared to the "hard" bond stretching and bending terms, where it takes a significant deformation for the structure to deviate

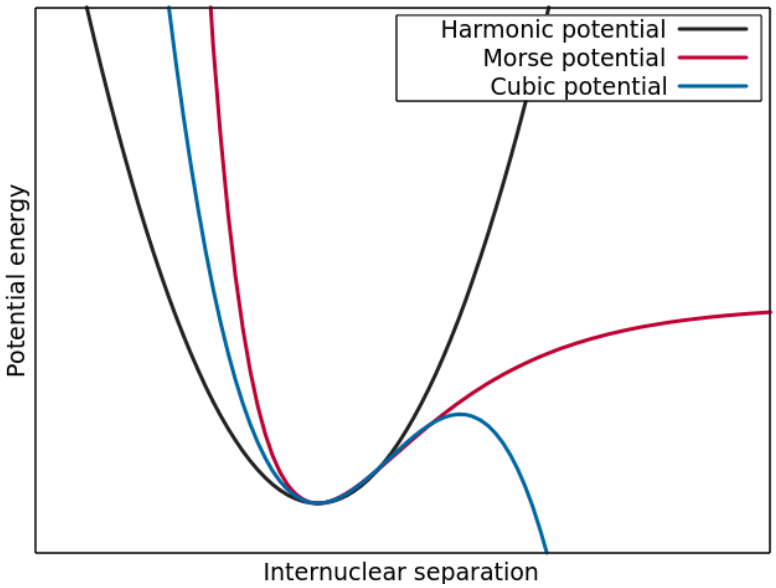


Figure 2.1: Comparison of harmonic, Morse, and cubic potential approximations.

from the reference values used. The torsion energy is represented by a cosine series expansion,

$$E_{\text{torsion}} = \sum_{n=0}^{N} \frac{V_n}{2} \left[1 + \cos\left(n\varphi - \varphi_0\right)\right], \tag{2.8}$$

where φ is the dihedral angle, V_n is the barrier height, n is the number of minimum energy points as the bond rotates from 0°to 360°, and φ_0 denotes where the torsion potential has a minimum value. As with the previous two contributions, the torsional component can be expanded to include higher-order terms. Including these terms can significantly improve the accuracy over a single-termed torsional energy, however, it requires many more parameters which may be tedious to compute.

While the torsional term is used for four atoms bonded in succession, improper torsional terms can be defined for a group of four atoms bonded in any order. For example, this can be three atoms around a trigonal center and improper torsions are typically used for controlling the geometry or stereochemistry around atoms. In the sample case of cyclobutanone in the geometry it is known to adopt experimentally, an improper torsion term is required for the oxygen atom to lie in plane with the ring. There are various ways to model these contributions, but one such improper torsion term can have the form

$$E_{\text{improper}} = k(1 - \cos 2\varphi), \tag{2.9}$$

where φ is once again the dihedral, and k is a stiffness parameter. Out-of-plane terms based on the harmonic potential discussed previously can also be used to control geometries.

Non-bonded interactions are the next major portion of the force field energy, which are broken

down into the electrostatic and van der Waals terms. Typically, these interactions are only calculated between atoms that are separated by at least three chemical bonds, and can be applied to atoms both within and outside of the same molecule. The electrostatic term can be most simply represented by summing interactions between point charges using Coulomb's Law,

$$E_{\text{elec}} = \sum_{i=1}^{N_A} \sum_{j \neq i}^{N_B} \frac{q_i q_j}{r_{ij}} \,, \tag{2.10}$$

where N_A and N_B are the number of point charges (i.e. atoms) in the two molecules A and B being considered, q_i and q_j are their respective charges, and r_{ij} is the distance between them. This method can be applied when there are partial atomic charges assigned to atoms in the molecule. This is specifically important when considering polar or charged species, such as salts.[25] For example, GAFF requires partial charges assigned from the restrained electrostatic potential fit (RESP) model.[26] The electrostatic term can also be adapted to describe the charge description with multipole moments, such as dipoles or quadrupoles. This is given by the summation over molecules for all multipoles with the form

$$E_{\text{elec}}^{\text{multipoles}} = \frac{1}{2} \sum_{A} \sum_{B \neq A} Q_i^t Q_j^u T_{ij}^{tu} \,, \tag{2.11}$$

where Q_i^t is the t-order multipole moment at atomic site i in molecule A, and vice versa in molecule B for Q_j^u. The interaction function T_{ij}^{tu} relates the multipoles to the fixed molecular axis system, and depends only on the orientation and distance between multipoles.[27,28]

The final non-bonded interaction to be considered is the van der Waals term. Comprised of attractive (London dispersion) and repulsive forces, there are numerous ways to model these interactions. One particularly common choice is the Lennard-Jones 6-12 potential, which has the form

$$E_{\text{LJ}} = 4\epsilon \left[\left(\frac{\sigma}{r} \right)^{12} - \left(\frac{\sigma}{r} \right)^{6} \right] \tag{2.12}$$

with collision diameter σ, well depth ϵ, and atomic separation r. This potential is illustrated in Figure 2.2, showing the energy as atoms separate, with the optimal atom separation at r_m. These attractive (r^{-6}) and repulsive (r^{-12}) components have proven useful for modeling of noble gases, but the repulsive part of the potential is too steep for more complex systems such as molecules. Regardless, this potential is ubiquitous, and serves as the basis for many of the vdW energy terms whose descriptions will follow.

While the energy contributions described in this section are employed in all force fields, the specific functional form of any given component is unique to the implementation. Some force

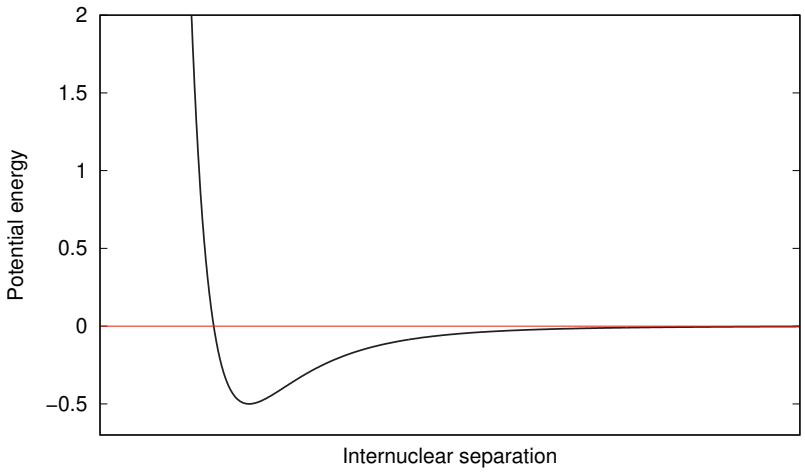


Figure 2.2: Graph of the Lennard-Jones potential.

fields may include cross terms which couple various internal variables such as the bond lengths and angles. An illustration of these terms can be seen in Figure 2.3. The addition of cross terms has led to a classification system for force fields, in which class I includes no cross terms or higher-order expansions of harmonic terms. A class II force field includes cross terms and higher-order expansions, and class III includes more specialized chemical properties such as polarization. These terms will be described as necessary in the following section.

The final pieces in every force field are the parameters. It is integral that these values are as accurate as possible, specifically for the non-bonded or torsional terms, since the performance of the force field is highly reliant upon these. The parameterization process can be described simply as one of trial and error, where parameters are adjusted until the result of the force field calculation matches the data set used to parameterize it. This data set can come from experimental information, although it is typically derived from more advanced quantum mechanical (QM) methods such as DFT or correlated wavefunction theories. The use of a force field can be as specific or general as the developer desires, so they are often specialized for specific classes of compounds, calculating different properties, or have more a general purpose. Many force fields, for example, are designed to model proteins or nucleic acids. These force fields would not be used when one is investigating molecular crystals, and force fields designed for molecules would not be used when trying to model large biomolecules. In this work, we will focus on force fields that are generally applicable for modeling small organic molecules.

2.1.1 TINKER

Developed by the Jay Ponder Lab, Tinker[29] is a molecular mechanics and dynamics toolbox designed to easily apply numerous different force fields to chemical systems. The Tinker package

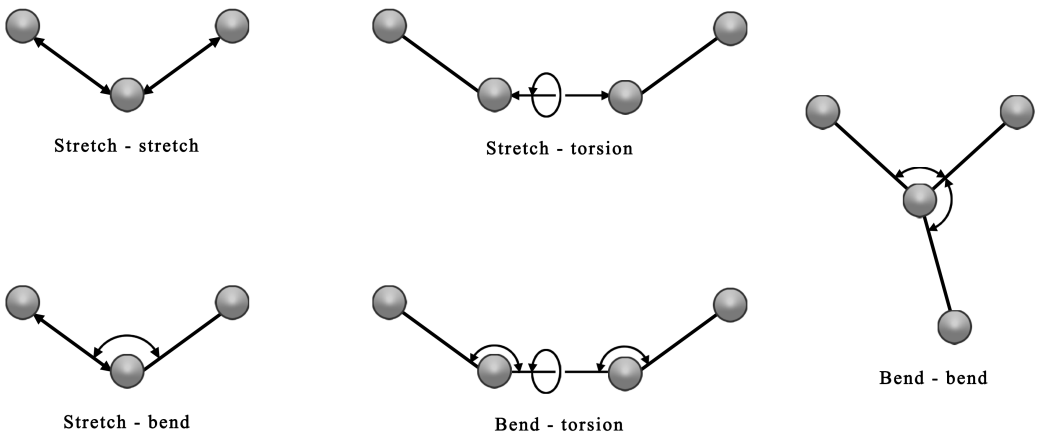


Figure 2.3: Illustration of coupled interactions important in evaluating the force field energy.

contains a collection of functions, subroutines, and parameter sets than can be tailored by the user to meet their needs. Of the eight force fields covered in this work, four are part of the Tinker package. In this section, the theoretical details will be summarized for those four force fields.

2.1.1.1 TINY

"TINY" is a rudimentary force field included in the Tinker package.[29] This FF is not technically published, as it is not meant for any special purposes other than simple geometry optimizations. TINY uses a very generic set of parameters, with LJ-type van der Waals (vdW) interactions and near omission of electrostatics.

Parameters are given for each atom type in the FF; therefore, calculating interactions between two atoms of the same type is as trivial as using those parameters in the relevant equations. Systems with more than one atom type would theoretically need $N(N-1)/2$ sets of parameters to describe interactions between each type. Atom types in TINY are determined by element and the number of bonded groups. There are, for example, four possible atom types for carbon – a mono, di, tri, or tetravalent atom type. As mentioned, parameterization is an arduous and time-consuming process and it is unrealistic to fully parameterize every possible interaction between atoms. Thus, it is commonly assumed that these parameters can be obtained by combining existing parameters via mixing rules. In this case, the popular Lorentz-Berthelot mixing rules for the LJ parameters are applied, where σ_{ij} and ϵ_{ij} between different atom types i and j are obtained from combination of

their homoatomic values:

$$\sigma_{ij} = \frac{1}{2}(\sigma_{ii} + \sigma_{jj})\,, \tag{2.13}$$

$$\epsilon_{ij} = \sqrt{\epsilon_{ii}\epsilon_{jj}}\,. \tag{2.14}$$

2.1.1.2 MM3

The MM3 force field, developed by Allinger and coworkers in the late 1980s, is one entry in a series of force fields designed for organic molecules.[30] This version addressed problems systematic to the MM2 force field, its predecessor, such as the underestimation of C-C rotational barriers, improper bond lengths, and improving the bending function. There are seven main contributors to the MM3 energy: bond stretching, angle bending, torsion, stretch-bend, torsion-bend, bend-bend, and van der Waals interactions.

The bond stretching, angle bending, and torsion terms are similar to those described by Equations 2.7, 2.6, and 2.8, adjusted to include higher-order terms. The stretching energy,

$$E^{\mathrm{MM3}}_{\mathrm{stretching}} = 71.94 k_s (r - r_0)^2 \left[r - 2.55(r - r_0) + \left(\frac{7}{12}\right) 2.55 (r - r_0)^2 \right]\,, \tag{2.15}$$

is expanded to include a quartic term, so the FF can be more tolerant to poor starting geometries. The angle bending term is also expanded up to a sextic term,

$$E^{\mathrm{MM3}}_{\mathrm{bending}} = 0.021914 k_\theta \left(\theta - \theta_0\right)^2 \sum_{i=0}^{4} a_i \left(\theta - \theta_0\right)^i\,, \tag{2.16}$$

for $a_0 = 1$, $a_1 = -0.014$, $a_2 = 5.6 \times 10^{-5}$, $a_3 = -7.0 \times 10^{-7}$, and $a_4 = 9.0 \times 10^{-10}$. The resulting function is monotonic up to 180° to prevent artificial minima from occurring. Lastly, the torsional energy is represented by a three-term expansion of Equation 2.8, given by

$$E^{\mathrm{MM3}}_{\mathrm{torsion}} = \frac{V_1}{2}(1 + \cos\varphi) + \frac{V_2}{2}(1 - \cos 2\varphi) + \frac{V_3}{2}(1 + \cos 3\varphi)\,, \tag{2.17}$$

where each term corresponds physically to different energy contributions. The first term is attributed to interactions between bond dipoles, the second accounts for conjugation and hyperconjugation effects, and the third coresponds to steric interactions between the atoms in the 1 and 4 position for torsion of the 2–3 bond.

The first cross term included in MM3 is the stretch-bend interaction. This allows bonds to stretch slightly when the angle between them decreases, and vice versa. This energy is given by

$$E^{\mathrm{MM3}}_{r\theta} = 2.51118 K_{s\theta}[(r - r_0) + (r' - r'_0)](\theta - \theta_0)\,, \tag{2.18}$$

where $K_{r\theta}$ is a force constant, r is the length of the first bond, r' the length of the second, and θ the angle between them. This term is only applied to cases where the bonds being stretched and angle being bent are centered around the same atom.

The torsion-stretch interaction is the next cross term to be included. This is to improve the description of bond lengths when the molecular conformation is anything other than staggered. This energy term is given by

$$E_{\varphi r}^{\mathrm{MM3}} = 11.995 \left(\frac{K_{\varphi s}}{2} \right) (r - r_0)(1 + \cos 3\varphi) \,, \tag{2.19}$$

where $K_{\varphi s}$ is a force constant, r is the bond length, and φ is the dihedral angle. The torsion-bend term had been used in the previous iteration of the force field (MM2), but this interaction alone has little effect on most systems, and can be accounted for within the torsion-stretch term. The bend-bend interaction is the final cross term to consider here, given by

$$E_{\theta\theta'}^{\mathrm{MM3}} = -0.021914 K_{\theta\theta'}(\theta - \theta_0)(\theta' - \theta_0') \,, \tag{2.20}$$

where $K_{\theta\theta'}$ is again a force constant, and θ and θ_0 are the respective bond angles. The bend coupling term is particularly important for calculating vibrational frequencies, and is only used for angles involving two carbon (or heavier) atoms, a carbon and a hydrogen, or two hydrogens.

Finally, the non-bonded terms are the last to be considered. In this implementation, the van der Waals interaction is based on the two-parameter Hill potential,[31] with the form

$$E_{\mathrm{vdW}}^{\mathrm{MM3}} = \epsilon \left[-2.25 \left(\frac{r_{\mathrm{vdW}}}{r} \right)^6 + \left(1.84 \times 10^5 \right) e^{-12.00(r/r_{\mathrm{vdW}})} \right] \,, \tag{2.21}$$

where the adjustable parameters are the van der Waals radii, r_{vdW}, and an energy parameter, ϵ. The parameters were only adjusted slightly from MM2, due to the predecessor's success. The electrostatic contribution is modeled by a collection of bond dipoles as opposed to point charges, as there are negligible differences between the two methods if parameterized correctly.[32,33] MM3 also includes an explicit hydrogen-bonding potential. For a X$-$H$\cdots$Y hydrogen bond, this is given by

$$E_{\mathrm{HB}}^{\mathrm{MM3}} = \frac{\epsilon_{\mathrm{HB}}}{\varepsilon} \left[1.84 \times 10^5 e^{-12(R_{\mathrm{YH}}/r)} - 2.25 \cos\beta \left(\frac{R_{\mathrm{X-H}}}{R_{\mathrm{X-H}}^0} \right) \left(\frac{r}{R_{\mathrm{YH}}} \right)^6 \right] \,, \tag{2.22}$$

where ϵ_{HB} is the H-bonding energy parameter, ε is the dielectric constant, r is the equilibrium H-bonding distance, R_{YH} is the H$\cdots$Y H-bonding distance, $\cos\beta$ is the cosine of $\angle$H$-$X$\cdots$Y, and $R_{\mathrm{X-H}}$ and $R_{\mathrm{X-H}}^0$ are the X$-$H bond length and equilibrium length, respectively.

Parameterization of MM3 was completed with reference to spectroscopic force constants, and

the geometries of known structures were used to obtain reference values for r_0, θ_0, φ_0, etc. Finally, once adequate geometries were determined, vibrational spectra and heats of formation for the molecules were examined to further tune the structural parameters.

2.1.1.3 MMFF94

The MMFF94 force field, developed in the mid-1990s at Merck Research Laboratories, is closely related to the MM3 force field with some key changes.[34] It was designed to be applicable to both organic molecules and proteins, used *ab initio* results throughout the development process, and was validated with experimental data.

The functional form of MMFF94 is extremely similar to that of MM3, so this description will focus only on the differences between them – primarily the non-bonded terms. Like MM3, the bond stretching term uses a quartic expansion to avoid the so-called "cubic stretch" error, and uses a cubic expansion for the angle bending. This is given by

$$E_{\text{bending}}^{\text{MMFF94}} = 0.043844 \frac{k_\theta}{2} (\theta - \theta_0)^2 \left[1 - 0.4(\theta - \theta_0)\right] , \tag{2.23}$$

where k_θ is the force constant. The torsional and stretch-bend terms are identical to that of MM3, shown in Equations 2.17 and 2.18, respectively. Improper torsion (out-of-plane bending) around trigonal centers is included, using the form

$$E_{\text{improper}}^{\text{MMFF94}} = 0.043844 \frac{k_{\text{imp}}}{2} \phi^2 , \tag{2.24}$$

where k_{imp} is a force constant and ϕ is the angle between the bond formed by the central atom and one bonded atom, and the plane formed by the other two bonded atoms.

The van der Waals interactions in MMFF94 utilize Halgren's buffered 14-7 potential, developed to be an alternative function to improve upon the well-known Lennard-Jones 6-12 potential while still being computationally simple. This has the form

$$E_{\text{vdW}}^{\text{MMFF94}} = \epsilon_{ij} \left(\frac{1.07 r_{ij}^*}{r_{ij} + 0.07 r_{ij}^*} \right)^7 \left(\frac{1.12 r_{ij}^{*7}}{r_{ij}^7 + 0.12 r_{ij}^{*7}} - 2 \right) \tag{2.25}$$

with well depth ϵ_{ij}, minimum-energy separation dependent on atomic polarizability r_{ij}^*, and separation r_{ij}. Modified values for ϵ_{ij} and r_{ij}^* are used to incorporate hydrogen-bonding interactions, since there is no specific term for H-bonds in MMFF94. This term, like all non-bonding terms, is only applied to atoms that are separated by at least three chemical bonds. The buffered 14-7 potential follows more elaborate combination rules. For the minimum separation parameter, this is

given by

$$r_{ij}^{*} = \frac{r_{ii}^{*3} + r_{jj}^{*3}}{r_{ii}^{*2} + r_{jj}^{*2}}\,, \tag{2.26}$$

and the well depth mixing by

$$\epsilon_{ij} = \frac{181.16 G_i G_j \alpha_i \alpha_j}{4C_{6,ii}/3\alpha_i + 4C_{6,jj}/3\alpha_j} \frac{1}{r_{ij}^{*6}}\,, \tag{2.27}$$

where the G parameters are constants to reproduce the well depths of like-atom pairs, α is the atomic polarizability, and C_6 is the dispersion coefficient.[35]

For electrostatics, a buffered coulombic form using partial charges is employed. This has the form

$$E_{\text{elec}}^{\text{MMFF94}} = 332.0716 \left(\frac{q_i q_j}{\varepsilon (r_{ij} + \delta)^n} \right)\,, \tag{2.28}$$

where q_i and q_j are partial atomic charges, ε is the dielectric constant, δ is the buffering constant, and r_{ij} is internuclear separation. The exponent, n, is usually 1, although it can also be 2 in the case of a distance-dependent dielectric constant. Distance buffering ($\delta > 0$) prevents the infinite attractive electrostatic interaction from overwhelming the finite repulsive interaction as oppositely charged atoms closely approach each other.

To paramaterize MMFF94, MP2/6-31G* optimized geometries were obtained for a wide variety of molecules. This set included the vast majority of important organic functional groups such as alcohols, amines, carbonyl derivatives, and many more. Over 20 chemical families were included in this initial parameterization, and an additional set of structures from the CCDC were later included. The process aimed to be mutually consistent, where all parameters are determined simultaneously so the accuracy of the force field could be confidently attributed to the functional form itself, and not the parameterization method. This is extremely computationally expensive, so parameters that depend closely on each other were grouped together and optimized. For example, the reference bond lengths and angles (r_0 and θ_0) only depend very weakly on the V_1, V_2, and V_3 torsional parameters, so they could be adjusted independently from each other.

2.1.1.4 OPLS-AA

The last FF used with Tinker was the Optimized Potentials for Liquid Simulations all-atom (OPLS-AA) force field developed by Jorgensen and coworkers at Yale University.[36] This FF derived many of its energy terms from the AMBER force field,[26] although the torsional and non-bonding terms are unique to OPLS-AA. In this description, each atom has its own type and parameter set, while the united-atom version of OPLS implicitly includes hydrogen atoms in the various carbon atom types. OPLS was originally designed to conduct molecular dynamics (MD) simulations on liquids,

where the united-atom approach saves considerable time due to the reduced number of interactions that need to be calculated. The all-atom approach, however, gives a better description for the torsional and non-bonding interactions, which have the greatest effect on the overall accuracy of a force field for molecular crystals.

The terms for bond stretching and angle bending are both based on the Hooke's Law approximation, seen in Equations 2.6 and 2.7, and are the same as those used in AMBER. All parameters for these terms are also the same as those in AMBER, except for the parameters for alkanes, which were adapted instead from the CHARMM[37] force field and led to much better results for the geometries and energetics of alkanes. The torsional energy term also takes on a form described previously, given by Equation 2.17.

Non-bonded interactions are modelled by a combined coulombic and LJ term, with the form

$$E_{\mathrm{non-bonded}}^{\mathrm{OPLS-AA}} = \sum_i \sum_j \left[\frac{q_i q_j}{r_{ij}} + 4\epsilon_{ij} \left(\left(\frac{\sigma_{ij}}{r_{ij}} \right)^{12} - \left(\frac{\sigma_{ij}}{r_{ij}} \right)^{6} \right) \right] f_{ij} \tag{2.29}$$

and using the standard combination rules described by Equations 2.13 and 2.14. The inclusion of the scaling factor, f_{ij}, accounts for including both inter- and intramolecular non-bonded terms, so long as the atoms involved are separated by at least three bonds. This scaling factor is simply one (1) for intermolecular interactions, and 0.5 for intramolecular interactions, so the same parameters can be used in all cases. Charges in OPLS are determined empirically by fitting to reproduce organic liquid properties. These are associated with different atom types as parameters, separated according to the functional group. This allows the parameters to be more transferable, and improves the speed of calculations for large molecules.

The parameterization of OPLS-AA is what sets it apart from AMBER, given their extremely similar functional forms. The parameters were developed and validated against experimental data, such as gas-phase calculations of molecular geometries and torsional energy profiles at the HF/6-31G* level, and the thermodynamic properties of liquids through MD simulations. Torsional parameters were derived from the gas-phase HF calculations by comparing the energies at every dihedral angle to single out the torsion contribution. These energies were fit against the functional form, unnecessary near-zero terms were removed, and the data was refit for each dihedral until the parameters were optimized to reproduce the torsional energy profiles with the fewest number of terms. Additionally, the torsion parameters needed to be refit any time the partial-charge parameters were adjusted, usually due to unsatisfactory agreement with the liquid properties. Radial distribution functions, heats of vaporization, and densities of liquids were calculated via MD simulations and used to validate and further optimize the torsion and non-bonded parameters. A total of 34 organic

liquids were included in the parameterization process, including alkanes, alkenes, alcohols, thiols, and amides, the geometries of which were allowed to be fully flexible during the MD simulations.

2.1.2 General Utility Lattice Program

The General Utility Lattice Program (GULP), developed at the Curtin University of Technology by Julian Gale and coworkers, is designed to implement force-field methods for a variety of calculation types centered on the solid state.[38,39] Described here are the three force fields used in this work that are implemented within the GULP software package.

2.1.2.1 DREIDING

Developed by Mayo and coworkers in the late 1980s, the DREIDING force field was inspired by the success of the MM2, AMBER, and CHARMM force fields, but aimed to be parameterized based on hybridization of atoms, not by functional group.[40] The bond stretching term employs the harmonic description from Equation 2.6. Parameters for the Morse potential approximation are also included, but in order to be more tolerant to poor starting geometries, the harmonic description is the default. Equilibrium bond radii are determined by $r_0 = r_i + r_j - 0.01$ Å, where r_i and r_j are atomic radii based on standard experimental reference data. The angle bending term also implements a harmonic form, this time involving the cosine of the angles involved. This has the form

$$E_{\text{bending}}^{\text{DREIDING}} = \frac{C_{ij}}{2} \left(\cos\theta - \cos\theta_0\right)^2 , \tag{2.30}$$

where θ is the angle under scrutiny, θ_0 is the reference angle, and C_{ij} is a constant derived from the force constant, k_{ij}, with the form $C_{ij} = k_{ij}/(\sin\theta_0)^2$. This is preferred over the harmonic form described in Equation 2.7 because the simpler harmonic potential does not lead to a zero slope as the angle approaches 180°. For molecules with linear geometries, an alternate form of Equation 2.30 is used,

$$E_{\text{linear}}^{\text{DREIDING}} = k_{ij}(1 + \cos\theta) , \tag{2.31}$$

where the only parameter is the force constant, k_{ij}. The torsional term employs an altered form of Equation 2.8 to ensures that torsional energy is always zero at equilibrium, instead of a negative value. This is given by

$$E_{\text{torsion}}^{\text{DREIDING}} = \frac{V}{2} \left(1 - \cos\left[n(\varphi - \varphi_0)\right]\right) \tag{2.32}$$

with dihedral angle φ, periodicity n, rotation barrier V, and equilibrium dihedral φ_0. Due to the nature of the function, the torsional parameters are limited by symmetry. There must be a slope of zero at 0°and 180°, and the equilibrium dihedral must be some multiple of $180°/n$. The torsional parameters are based solely on hybridization and defined independently of the specific

atom involved, i.e., the parameters for a torsion around an sp^3 carbon and sp^2 carbon bond is the same as those for an sp^3 carbon and sp^2 nitrogen bond. Improper torsions in DREIDING have a functional form identical to the angle bending function in Equation 2.30. Here, θ and θ_0 are replaced by ϕ and ϕ_0, the angle and equilibrium angle between the plane formed by two of the bonds attached to the central atom, and the third bond formed. The constant C_{ij} is also replaced with $C = k/(\sin\phi_0)$, with force constant k.

Van der Waals interactions in DREIDING are modelled by both the 6-12 LJ potential from Equation 2.12 and the exponential-6 (X6) form,

$$E_{\mathrm{X6}}^{\mathrm{DREIDING}} = Ae^{-Cr} - Br^{-6} \tag{2.33}$$

with internuclear separation r and parameters A, B, and C. The A and B parameters follow the geometric mean combination rule from Equation 2.14, and the C parameter follows the arithmetic mean combination rule from Equation 2.13. Parameters for the van der Waals interactions are adapted from those published by Williams and coworkers.[41] Electrostatic interactions are either ignored, or calculated based on Gasteiger charge estimates with Coulomb's law.[42]

Finally, a hydrogen-bonding term is included in the non-bonded interactions. This term is invoked if the atom type for a hydrogen capable of H-bonding is included. The hydrogen-bonding potential is adapted from the CHARMM force field, or the LJ 10-12 potential, with the form

$$E_{\mathrm{HB}}^{\mathrm{DREIDING}} = D_e \left[5 \left(\frac{r_{\mathrm{HB}}}{r_{\mathrm{DA}}} \right)^{12} - 6 \left(\frac{r_{\mathrm{HB}}}{r_{\mathrm{DA}}} \right)^{10} \right] (\cos\theta)^4 , \tag{2.34}$$

where D_e is the well depth, r_{HB} is the distance between the H-bond donor atom and the hydrogen, r_{DA} is the distance between the donor and acceptor atoms, and θ is the angle between the donor, hydrogen, and acceptor. There are no cross terms included in DREIDING. Parameterization is minimized, as only generic force constants are used and strictly applied within the formulae described. This FF aims to simplify atom typing, and does away with determining parameters that depend on the combination of the stretching, bending, and torsional terms.

2.1.2.2 UFF

The next GULP-implemented force field is the Universal Force Field, or UFF. Developed by Rappé and coworkers in the early 1990s, UFF is intended to approximate parameters for all elements in the periodic table.[43] Thus, atom types are only determined by the element and hybridization or geometry. The bond stretching term once again usedsthe harmonic potential from Equation 2.6,

and the angle bending term uses the cosine Fourier expansion,

$$E_{\text{bending}}^{\text{UFF}} = k \sum_{n=0}^{m} C_n \cos n\theta \,, \tag{2.35}$$

where the C_n coefficients are selected to ensure the function has a minimum when angle θ is at the equilibrium angle θ_0. This cosine expansion was chosen instead of the harmonic approximation because large-amplitude movement is better described. The torsional term is identical to that from the DREIDING FF, seen in Equation 2.32. Like DREIDING, the parameters in this equation are determined by hybridization only.

The van der Waals term is once again given by the LJ 6-12 potential from Equation 2.12. While an X6 form has also been included, UFF prefers the LJ approach due to the tendency for X6 energies to be excessively high for small interatomic separations. The Lorentz-Berthelot mixing rules are applied to derive heteronuclear pair parameters, given by Equations 2.13 and 2.14. Electrostatics are described via Coulomb's law, with partial charges assigned based on a charge equilibration approach.[44] Parameters in UFF were primarily obtained by fitting experimental data.

2.1.2.3 GAFF

The general AMBER force field (GAFF) is a parameter set designed to model organic molecules and be compatible with the original AMBER force field for biomolecules. Developed by Wang and coworkers, it employs a simple functional form and general parameters derived from experimental or *ab initio* data to reproduce properties of organic pharmaceutical molecules both inside and outside of the parameterization set.[45] The GAFF functional forms for bond stretching, angle bending, and torsion energy have already been described. The stretching and bending term use the simple harmonic potentials described in Equations 2.6 and 2.7, and the torsional potential has the same form as that described in Equation 2.8. The van der Waals term is given by the 6-12 LJ potential, Equation 2.12. While the original AMBER force field includes an optional hydrogen bonding term using the LJ 10-12 potential seen in Equation 2.34, this is not implemented for the GAFF parameters and there are no specific hydrogen-bond-capable atom types defined.[26]

Despite GAFF's more simplistic functional forms, its distinguishing feature from the previous FFs detailed thus far is its description of electrostatics. GAFF requires the user to provide restrained electrostatic potential (RESP) charges at the HF/6-31G* level for any systems of interest.[46] These can be calculated with the Gaussian software.[47] Since calculation of the RESP charges is relatively expensive if considering a large number of different molecules, the cheaper AM1-BCC method can also be used to assign partial charges instead.[48,49] These partial charges are then used in the Coulomb potential to determine the electrostatic energy contribution.

GAFF is similar to DREIDING in that atom types are generic, and are defined primarily according to element, hybridization, and aromaticity. The LJ parameters were the same as those in the original AMBER force field, and the equilibrium bond lengths and angles were determined from a combination of MP2/6-31G* calculations, experimental crystal data, and previous AMBER parameters. Over 3000 molecule optimizations were performed to determine these structural parameters. Torsion parameters were determined by scanning the angles at the MP4/6-311G(d,p) level, and fitting the V_1, V_2, and V_3 parameters to replicate the calculated profiles.

Conversely, force constants are determined via empirical formulae that are more complex than the combination rules seen previously. The heteronuclear bond stretching constant is determined by

$$k_{\text{stretching}}^{\text{GAFF}} = \frac{k_{ii}|r_{0,ij} - r_{0,jj}| + k_{jj}|r_{0,ij} - r_{0,ii}|}{|r_{0,ij} - r_{0,jj}| + |r_{0,ij} - r_{0,ii}|} \left(\frac{1}{r_{ij}}\right)^m , \tag{2.36}$$

where m is a power parameter, r_{ij} is the actual bond length, the various r_0 reference values are those determined from the structural parameters discussed previously, and k_{ii} and k_{jj} are the homoatomic force constants for atoms i and j, respectively. Angle bending force constants are determined by the formula

$$k_{\text{bending}}^{\text{GAFF}} = \frac{143.9 a_i b_j a_k}{(r_{0,ij} + r_{0,jk})\theta_0^2} \exp\left[-2\frac{(r_{0,ij} - r_{0,jk})^2}{(r_{0,ij} + r_{0,jk})^2}\,,\right] \tag{2.37}$$

with parameters a and b given for atoms i, j, and k, θ_0 is the equilibrium bond angle, and the various r_0 parameters are the equilibrium bond distances between the noted atoms.

2.1.3 DMACRYS

The crystal energy program DMACRYS, developed by Price and coworkers at University College London, combines a QM calculation of conformer energies with an anisotropic potential model to determine the lattice energy of structures.[27] Like the previous force field methods, the total energy of the system is given by a summation of the contributors. In this case, the QM conformer energy replaces the bonded energy terms, and the non-bonded energy is the sum of electrostatic and vdW potentials, like the FFs previously described. The electrostatic contribution is given by Equation 2.11, following the distributed multipole analysis (DMA) performed on the molecule. The van der Waals interaction is then modelled by the Buckingham potential, with the form

$$E_{\text{vdW}}^{\text{DMACRYS}} = A_{ij} e^{-B_{ij} r_{ij}} - \frac{C_{ij}}{r_{ij}^6}\,, \tag{2.38}$$

where A_{ij}, B_{ij}, and C_{ij} are parameters for atoms i and j, and r_{ij} is the internuclear distance. The parameters for heteronuclear pairs are determined again via the Lorentz-Berthelot mixing rules,

with A_{ij} and C_{ij} determined by Equation 2.14 and B_{ij} determined by Equation 2.13. DMACRYS employs either the transferable FIT[25] or Williams[41] parameters for the vdW term.

Following the QM calculation of the conformer energy and DMA, the molecules in the crystal are kept rigid during the lattice energy determination. This is because the DMA is dependent on the conformation of the molecule, so any change in structure would require the DMA to be recalculated. The choice of electronic-structure method to determine conformer energy is up to the user, with a DFT calculation used in this work.

2.2 DENSITY-FUNCTIONAL THEORY

Conversely to the force fields described previously, density-functional theory (DFT) computes properties of many-electron systems using the electron density, ρ. The energy of a many-electron system with wavefunction $\psi(\mathbf{r}_1, ..., \mathbf{r}_N)$ for N electrons can be determined using the time-independent Schrödinger equation,

$$\hat{H}\psi = \left[\hat{T} + \hat{V} + \hat{U}\right]\psi = E\psi\,, \tag{2.39}$$

where $\hat{H}$ is the Hamiltonian, and $\hat{T}$, $\hat{V}$, and $\hat{U}$ are the operators for kinetic energy, electron-nuclear, and electron-electron interaction energy, respectively. Instead of the wavefunction, the central quantity in DFT is the electron density, which is a function of position $\mathbf{r}$, and can be evaluated from a sum over occupied orbitals ψ_i:

$$\rho = \sum_{i=1}^{N} |\psi_i|^2. \tag{2.40}$$

The total Kohn-Sham DFT energy can be written as

$$E(\rho) = T_\circ + V_{\mathrm{nuc}}(\rho) + J(\rho) + E_{\mathrm{XC}}(\rho)\,. \tag{2.41}$$

The energy contributions, in atomic units, include the kinetic energy $T_\circ$,

$$T_\circ = -\frac{1}{2}\sum_i \int \psi_i^*(\mathbf{r})\nabla^2\psi_i(\mathbf{r})\,d\mathbf{r}\,, \tag{2.42}$$

electron-nuclear energy V_{nuc},

$$V_{\mathrm{nuc}} = \int V_{\mathrm{ext}}\rho(\mathbf{r})\,d\mathbf{r}\,, \tag{2.43}$$

and Coulomb repulsion energy,

$$J = \frac{1}{2}\iint \frac{\rho(\mathbf{r_1})\rho(\mathbf{r_2})}{r_{12}}\,d\mathbf{r_1}\,d\mathbf{r_2}\,. \tag{2.44}$$

The final contribution to the DFT energy is E_{XC}, which describes the exchange-correlation (XC) energy of the electrons. While E_{XC} represents a relatively small contribution to total energy compared to the first three terms, it is still an essential component. Unfortunately, the exact form of the XC functional is unknown and there is no systematic method to obtain it. There are a number of density-functional approximations (DFAs) proposed, however, to estimate E_{XC}. The simplest of these is the local density approximation (LDA), which depends only on ρ, but this results in large errors for molecular systems. More sophisticated XC functionals depend on the density gradient as well. These generalized gradient approximations (GGAs), while still imperfect, provide a more realistic description of the XC energy. The most popular DFAs are hybrid functionals, which include a mixture of density-functional and Hartree-Fock exchange. The B86bPBE functional with the exchange-hole dipole moment (XDM) dispersion correction[50,51] was employed in this work for high-level geometry optimizations on molecular crystals. B86bPBE is a GGA that combines the B86b[52] exchange functional and the PBE[53] correlation functional. Additionally, the B3LYP[54,55] hybrid functional, again with the XDM dispersion correction, was used to perform the QM calculation of conformer energy for DMACRYS optimizations.

2.3 Evolutionary Algorithms for Structure Generation

While not unique to structure generation methods, evolutionary algorithms have become popular in CSP protocols, more specifically genetic algorithms applied to optimization problems. The ideas behind these algorithms are heavily inspired by their namesake concepts in biology. In general, genetic algorithms follow a "survival of the fittest" approach, where a population of individuals are modified from generation to generation according to a chosen fitness parameter. Variation operators based on recombination and mutation are applied to create the next generation of individuals, and so on and so forth. By selecting the most fit individuals when applying the variations, ideally the fitness of the population will gradually increase with each generation and find the global minimum. This process of selection and variation can be continually repeated until some halting criteria where a "solution" with the highest fitness is found, or as many structures as the user wants have been collected. In the context of the structure generation problem, candidate structures are the individuals, and the fitness parameter can be any computed crystal property. Volume, hardness, and order are potential options; however, the parameter of choice in this work and many others is the energy. Thus, an increase in the fitness implies a lower energy, or more stable structure.

Figure 2.4 shows a flow chart of the steps involved in the algorithm employed in this work. In the initialization, an inaugural set of random structures are produced. These can be constrained

to include only specific space groups, such as those most common for molecular crystals which account for over 83% of crystal structures in the CCDC database.[56] Using the methods described previously, these structures are then optimized by an external energy minimization program. This allows the algorithm to rank the produced structures according to their fitness, where high fitness means a low energy. Based on these results, variation operators are applied to the population and the next generation of structures is produced. To find the global minimum, these variation operators, described in Section 2.3.1, are designed to create suitable offspring from the most fit structures in each generation. Randomly generated structures continue to be produced and included in each subsequent generation, and anti-seeds are applied when creating these subsequent generations to avoid trapping the algorithm in a local minimum. The anti-seeds, described in Section 2.3.2, penalize structures previously sampled so they are less likely to be sampled again when new generations are being created. This process of generating, evaluating, and selecting structures continues until a halting criteria is achieved – either the same structure has been identified as the minimum for many generations in a row, or in this case, a certain number of structures has been produced.

2.3.1 VARIATION OPERATORS

With a solid understanding of the evolutionary algorithm in a more general sense, we can delve further into the mechanisms for variation and how they work. The descriptions of the variation operators are adapted from three main papers from the developers of USPEX,[19–22] the structure generator used throughout this work.

2.3.1.1 HEREDITY

The first operator to consider is heredity, where two parent structures are chosen and combined to produce one child structure. This is done by taking two spatially coherent slabs from the parents, and combining the slabs to create the child. In the first parent, a lattice vector with an arbitrary length of 1 is chosen. The atoms along this vector are shifted by a randomly generated number between zero and one, and any atoms that are shifted out of the unit cell are returned. While the shifted and non-shifted unit cells are physically identical, the shift increases the diversity of child structures. A second number, x, is then generated between zero and one. This time, every atom from the first parent with a coordinate value on the chosen lattice vector from 0 to x is taken, and likewise the second parent from x to 1. These slabs are combined, and the lattice parameters of the child are determined by the randomly weighted average of the parent parameters.

In order to choose which slabs to take from each parent structure, an order parameter is used. This is applied in both the heredity and mutation operators, so that more ordered structures for

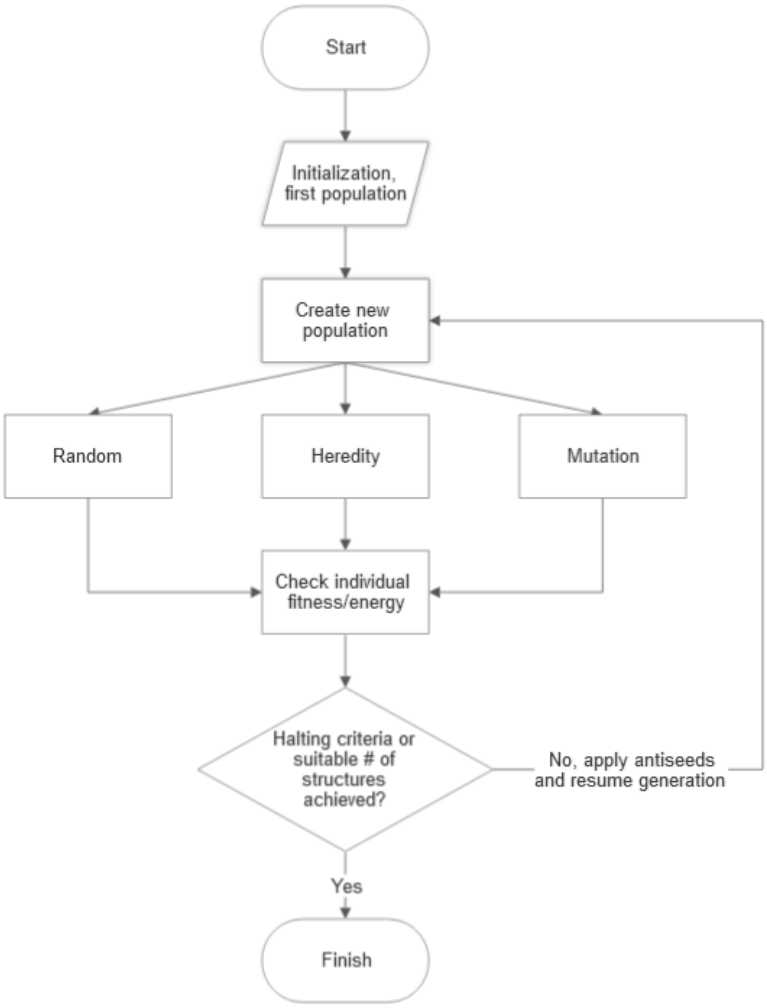


Figure 2.4: Flow chart of a typical evolutionary algorithm for structure generation.

heredity, and more disordered structures for mutations, are chosen. To do this, the correlation between atomic order and energy of the structure is calculated, and attempts are made to cut N_s slabs from the parent structures based on this correlation. In a total absence of correlation, only one slab is cut from that particular parent. Depending on the size of the unit cell, a correlation closer to 1 (or -1, if interested in disorder), meanwhile, results in more cuts attempted.

2.3.1.2 MUTATION

The second significant operator is mutation, where one individual structure is used to make one new individual. This can be applied to the lattice, molecules, or both in the case of softmutation. For a lattice mutation, a strain matrix is applied to the lattice parameters, while the atom positions or fractional coordinates within the lattice remain the same. The atom positions will likely change upon relaxation, but are identical between the unrelaxed child structure and the parent. This

operator allows for the structural “neighbourhood" surrounding good individuals to be searched. For molecular crystals, coordinate and rotational mutations are used. These replace the atom position mutation for ionic crystals, where random pairs of atoms in the crystal are swapped. The coordinate mutation displaces molecules in high-order structures to perturb these structures and once again search in the neighbourhood surrounding them. The rotational mutation works similarly, where molecules are selected and rotated at random. This operator is important to include, since the rotation of molecules can provide a significant push to help the system jump out of a local minimum.[22]

While still a mutation operator, softmutation is not based on the same principles as the lattice or atom mutations. Instead, it is based on minima-hopping, by attempting to cross low-energy barriers as opposed to high-energy ones. This helps find new basins of local minima quickly, and is good for the general exploration of the crystal energy landscape.[57] Atoms are moved along the eigenvector of the softest (or lowest frequency) mode in both positive and negative directions, with a user-defined displacement amplitude. One important feature is that the structure does not need to move exactly along the eigenvector – an approximate direction and large enough displacement amplitude is enough to find new low-energy structures. Because of this, an exact *ab initio* dynamical matrix is not necessary, and it is constructed from bond hardness coefficients.[21] In the prediction of molecular crystals, this combines aspects of the coordinate and rotational mutations and modifies both the molecular positions and orientations simultaneously.

2.3.2 FINGERPRINTS, ANTISEEDS, AND PREVENTING TRAPPING

Working with evolutionary algorithms, the biggest risk is that the method will become trapped around one or many local minima, and has no mechanism in place to escape them. This is common, especially for landscapes with local minima surrounded by high energy barriers. Thus, there must be some way of identifying similar structures and penalizing them to make the set of individuals more diverse. The first method, the fingerprint function, compares geometric properties of the structures based on interatomic distances. This function has the form

$$f(R) = \sum_i \sum_{j \neq i} \frac{Z_i Z_j}{4\pi R_{ij}^2} \frac{V}{N} \delta(R - R_{ij}) \,, \tag{2.45}$$

where Z is the atomic number, R_{ij} is the interatomic distance, V is cell volume, and N is the number of atoms in the cell. $\delta(R - R_{ij})$ is a Gaussian-smeared delta-function, which absorbs numerical errors and ensures the fingerprint function is smooth. This is summed over all atoms in the unit cell i, and all other atoms j within a certain cutoff distance. In order to compare fingerprints, the function is normalized, then discretized to represent it as a vector. Using these vectors for the

two structures being compared, the distance between them can be measured to determine similarity. Of the three potential distance measures considered, the cosine distance gives the best results.[58,59] This comparison yields a value between 0 and 1, where more similar structures will have a distance score closer to 0. This similarity factor is then used to ensure structures are "different" when participating in creating the next generation of individuals. Two parents chosen to be used in the heredity operator, for example, must be classified as different in order to produce a child structure.

Another method to avoid trapping is to apply antiseed techniques to penalize structures already sampled, and encourage the algorithm to diversify the population. By adding Gaussian potentials to sampled parts of the energy landscape, and storing their widths and heights, a time-dependent fitness function can be used to guide the algorithm. This fitness function has the form

$$f = f_0 + \sum_a W_a e^{(-d_{ia}^2/2\sigma_a^2)}\,, \tag{2.46}$$

where f is time-dependent fitness, f_0 is the true fitness property of choice (energy), W_a and σ_a are the height and width of the Gaussian, and d_{ia} is the fingerprint distance.[21] This can be applied by specifying structures to be penalized in the search, or more commonly by applying antiseeds to all structures sampled. By replacing the fitness parameter with the time-dependent one, structures that have been previously sampled become less likely to be selected as parents for heredity or mutation, and forces new areas of the energy landscape to be explored.

2.4 Comparing Crystal Structures

A variety of crystal structure comparison methods exist and were employed in this work to identify matches between predicted and experimental structures. The first of these, the powder pattern difference (POWDIFF) implemented in critic2,[60] calculates the powder X-ray diffraction (PXRD) pattern for a given pair of crystal structures and compares them using de Gelder's cross-correlation algorithm.[61] Powder patterns are represented as two continuous functions $f(x)$ and $g(x)$, where $x = 2\theta$, and the similarity is determined by the overlap between the two functions. This is given by the cross-correlation function $c_{fg}(\delta)$,

$$c_{fg}(\delta) = \int f(x)g(x+\delta)\,dx \tag{2.47}$$

with offset shift δ. The cross-correlation function is normalized such that absolute comparisons can be made between patterns, eliminating the need to pre-scale the powder patterns being compared. The inclusion of the shift, δ, accounts for details in the spectra that are similar but slightly shifted,

since this can have a significant and unwanted effect on the similarity measurement. The similarity between functions $f(x)$ and $g(x)$, S_{fg}, can then be calculated. The dissimilarity, or POWDIFF value, is given by $1 - S_{fg}$ if normalized to unity. Thus, any given pair of crystal structures are more similar when the POWDIFF value approaches zero.

While POWDIFF is a useful metric for comparing structures from analogous sources (i.e. two experimental structures, or relaxed with the same method), difficulties arise when comparing CSP-generated structures to an experimental one. Because CSP typically ignores thermal effects, it is considered “static lattice", and generally gives a more compact unit cell compared to experimental structures.[62] This is to the detriment of the reliability in the POWDIFF comparison, since PXRD peak positions are sensitive to changes in cell volume. Thus, the variable-cell powder difference (VC-PWDF) comparison method, developed by Mayo and coworkers, has been used to more effectively identify truly similar structures.[63,64] It begins by converting both structures being compared to their Niggli reduced cells, checking that the volume and lattice parameters are within user-specified tolerances (usually 20%), and applying transformations to account for inconsistencies in unit cell description. The VC-POWDIFF is calculated between all the transformed variations and the reference structure, and the variation with the lowest VC-POWDIFF is kept. The volume correction is applied by replacing the lattice parameters of the structure being compared to those of the reference, and finally the VC-POWDIFF value for the volume-corrected structures are determined via the POWDIFF metric described at the beginning of this section.

The last structure comparison method is the COMPACK algorithm developed by Chisholm and Motherwell.[65] As opposed to the diffraction pattern comparison methods discussed previously, COMPACK measures structure similarity by assessing the interatomic distances and molecular structures of a specified cluster. An adequately-sized cluster is necessary for the COMPACK comparison to be representative of the structure. For molecular crystals, a cluster size of 20 molecules is sufficient,[66] and a cluster of one (1) would only be comparing a single molecule. Working with two clusters, the interatomic distances and angles are calculated between an origin atom and all other atoms in the two crystal structures. Based on how well these values match, COMPACK can determine the number of molecules that match in the clusters, ideally a 20/20 match. The root-mean-squared-deviation, RMSD(N), of the measurements between the two structures is reported, where N is the number of molecules in the cluster. This provides a quantitative measure of similarity in cases where multiple 20/20 matches may exist. Thus, an RMSD(20) of zero would indicate an exact match for the 20 molecules.

CHAPTER 3

RESULTS AND DISCUSSION

3.1 CCDC-BT7 ENTRY

The first major project of this thesis encompassed a submission to the structure generation phase of the Cambridge Crystallographic Data Centre's 7th CSP Blind Test. This was the experimentally assisted challenge, target XXIX. This compound, also known as methyl anthranilate, is small, relatively inflexible, and used primarily as a grape flavouring agent.[67] Since there are two possible molecular conformers, seen in Figure 3.1, we had to consider that either conformer could be be possible in the experimental structure, and complete our calculations accordingly. Although one conformer was more stable than the other, intermolecular interactions play too big a role in crystallization to ignore the higher-energy conformer.[68] A simulated PXRD pattern for the experimental structure (Figure 3.8) was provided, and was indexed by R. Alex Mayo to help guide the following work. The quality of the PXRD pattern, however, was quite poor – given only as an image of the spectrum, not as specific data that could be plotted. This made it difficult to index, as well as difficult to compare with simulated PXRD of structures generated throughout this project.

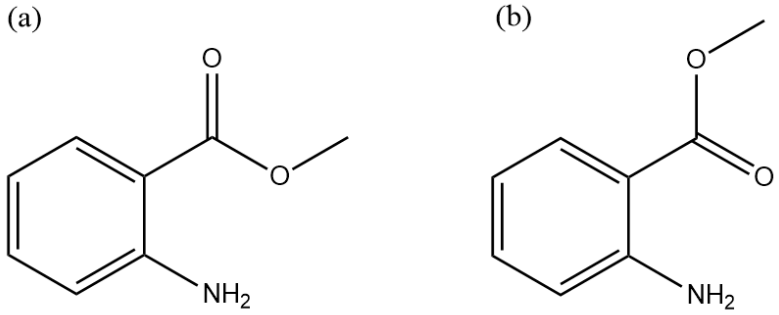


Figure 3.1: The two conformers of compound XXIX, methyl anthranilate, considered for crystal structure generation. Conformer (b) was predicted to be more stable by 2.9 kcal/mol via B3LYP-XDM/6-31+G* geometry optimization.

3.1.1 METHODS

Two conformers of compound XXIX were optimized using the B3LYP-XDM[50,54,55] functional and the 6-31+G* basis set with Gaussian 09 E.01.[47] Initial structure generation was performed using USPEX, version 10.4.[19–22] Six runs were performed, taking $Z = 4$ or $Z = 8$ with both conformers, and a 1:1 ratio of the two. A minimum of 10,000 structures were generated for each of the $Z = 4$ runs and a minimum of 6,000 were generated for each of the $Z = 8$ runs, using 100 structures per generation (30% from heredity, 50% random, and 10% each from soft mutation and rotation). The random structures were generated for the P1, $P\bar{1}$, $P2_1$, Pc, C2, Cc, C2/c, $P2_1/c$, P222, $P222_1$, $P2_12_12$, $P2_12_12_1$, and $Pca2_1$ space groups, which are common in molecular crystals. Antiseeds were used starting from generation 2, with "antiseedsMax" set to 0.01 and "antiseedsSigma" to 0.005. Volume estimates were 800 Å^3 for $Z = 4$ and 2000 Å^3 for $Z = 8$. The "TINY" force field in Tinker[29] version 8.9.1 was used for rigid-molecule relaxation of the initial structures.

Subsequent rigid-molecule relaxation was then performed on all generated structures with DMACRYS[27] version 2.3.0 using the FIT potential. The B3LYP-XDM/6-31+G* conformational energy difference was added to the DMACRYS lattice energies to allow energetic comparison of all structures. All unique structures with DMACRYS energies within 3.5 kcal/mol of the minimum were carried forward to single-point energy evaluation with periodic-boundary DFT using Quantum ESPRESSO[69] versions 6.5 and 6.8. Structures were deemed to be duplicates if their volumes were identical to within 0.1 Å^3, their energies identical to within 0.01 eV, and their PXRD difference (POWDIFF) was less than 0.07. The POWDIFF values were determined from critic2[60] using the de Gelder's cross-correlation algorithm,[61] for 2θ between 5-50° and ideal Cu Kα radiation. The Gaussian broadening parameter was set to 0.05° and the triangle weighting used $\ell = 1°$.

All structures within 2.0 kcal/mol of the DFT minimum were carried forward to full DFT relaxations. The DFT calculations used the projector augmented-wave (PAW) approach,[70] the B86bPBE functional,[52,53] and the XDM dispersion correction.[51,71] Plane-wave cutoffs were set to 80 and 800 Ry for the wavefunction and density, respectively. Regular k-point meshes were selected automatically for each crystal using an R_k length parameter of 50 Bohr. The number of points (N_i) along each direction was determined from the reciprocal lattice vectors ($\mathbf{b}_i$, for $i = 1, 2, 3$) via the formula $N_i = \mathrm{int}[\max(1, R_k|\mathbf{b}_i| + 0.5)]$. For geometry relaxation the convergence thresholds on the forces and energy were set to 10^{-4} Ry/bohr and 10^{-5} Ry, respectively, as in our previous work.[13,72]

3.1.2 USPEX PARAMETER BENCHMARK

Prior to generating the structures, a short investigation of USPEX parameters was done to find different ways to boost diversity in the population, and to check that the program succeeds in

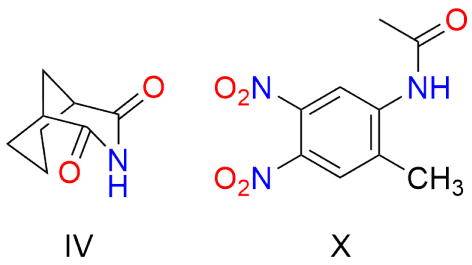


Figure 3.2: Diagrams of compounds IV and X, from the second and third blind tests.

finding the experimental structure(s) of known compounds. These tests were completed using target IV from the second blind test and target X from the third. The structures of these compounds are shown in Figure 3.2. USPEX runs were completed for structures with $Z = 4$ using the TINY FF in Tinker 8.10.1 and USPEX 10.3. In these runs, the number of structures per generation, ratio of heredity to random variation operator, use of the “AutoFrac" setting, and constraining random space groups were changed. Other than the options noted in table 3.1, all other settings (such as antiseeds) were kept the same. Since they are only incorporated in the algorithm while the generations are being produced, changing the number of structures per generation would influence how antiseeds are applied. Thus, if 5,000 structures were generated with 1,000 structures per generation, antiseeds would only be applied 4 times, as opposed to 49 times if generating the same number of structures with only 100 structures per generation. The smaller generation size allows the time-dependent penalty to be more consistently applied to the population, and should lead to more diverse structures. The ratio of the heredity- and randomly-generated structures with each new generation was also changed, with the use of mutation operators remaining the same. The “AutoFrac" option allows USPEX to alter the ratios of variation operators throughout the run in order to find the global minimum more quickly. Keeping this setting on is recommended, however, omitting it allowed for investigation of any improvement in diversity.

The final parameter checked was constraining the space group of the randomly-generated structures. Either there was no constraint applied, or the search space was limited to include only the P1, $P\bar{1}$, $P2_1$, Pc, C2, Cc, C2/c, $P2_1/c$, P222, $P222_1$, $P2_12_12$, $P2_12_12_1$, and $Pca2_1$ space groups, which are common in molecular crystals. It is important to note that this symmetry constraint is only applied to the randomly-generated structures – those produced as a result of heredity or mutation can be from any space group. The landscapes from these runs were examined based on how diffuse the CSP landscape was, and if there were any structures with the same molecular energy and volume as the experimentally known structure for the targets. Ultimately, there were only very small differences observed in the landscapes, with no parameter standing out as having a large effect on population diversity. Structures per generation was chosen to be 100, with a

Table 3.1: Summary of parameters investigated in the USPEX parameter benchmark.

Run	Structures/ Generation	Heredity/ Random	AutoFrac?	Symmetry Constraint	Target
1	50	0.5/0.3	n	n	IV
2	100	0.5/0.3	n	n	IV
3	250	0.5/0.3	n	n	IV
4	500	0.5/0.3	n	n	IV
5	1000	0.5/0.3	n	n	IV
6	100	0.5/0.3	y	n	IV
7	100	0.7/0.1	n	n	IV
8	100	0.3/0.5	n	n	IV
9	100	0.2/0.6	n	n	IV
10	100	0.3/0.5	y	n	IV
11	50	0.5/0.3	n	y	X
12	100	0.5/0.3	n	y	X
13	250	0.5/0.3	n	y	X
14	500	0.5/0.3	n	y	X
15	1000	0.5/0.3	n	y	X

heredity/random ratio of 0.3/0.5, “AutoFrac" on, and the symmetry constraint applied.

This short assessment of USPEX parameters also allowed us to test the ability of USPEX to identify the known experimental polymorphs of the targets investigated. While this is investigated in much more detail in Section 3.2, the crystal energy landscapes generated for targets IV and X, with halting criteria of 2500, 5000, and 10,000 structures, are shown in Figure 3.3. The experimental structures were found for both targets and are highlighted on the landscapes. For target IV, the experimental structure lies ca. 1.5 kcal/mol above the minimum and was found within the first 2500 candidates. For target X, the experimental form lies ca. 5 kcal/mol above the minimum and was only found when the search was extended from 5000 to 10,000 candidates. These results confirmed that USPEX and the TINY force field could be successful in finding the experimental structures of small organic molecules. Given the pseudo-random nature of the evolutionary algorithm, enough time must be given for the structure generator to optimize towards the experimental structure. Thus, a minimum of 10,000 structures with 100 structures per generation (100 generations) was chosen as the optimal number of structures and iterations required for the experimental structure to be found.

3.1.3 Preliminary Energy Ranking

The first crystal energy landscape generated for compound XXIX, shown in Figure 3.4, combined results from all six USPEX runs. This figure shows the landscape after the TINY-optimized structures were pruned to eliminate duplicates and re-optimized by DMACRYS. The relative

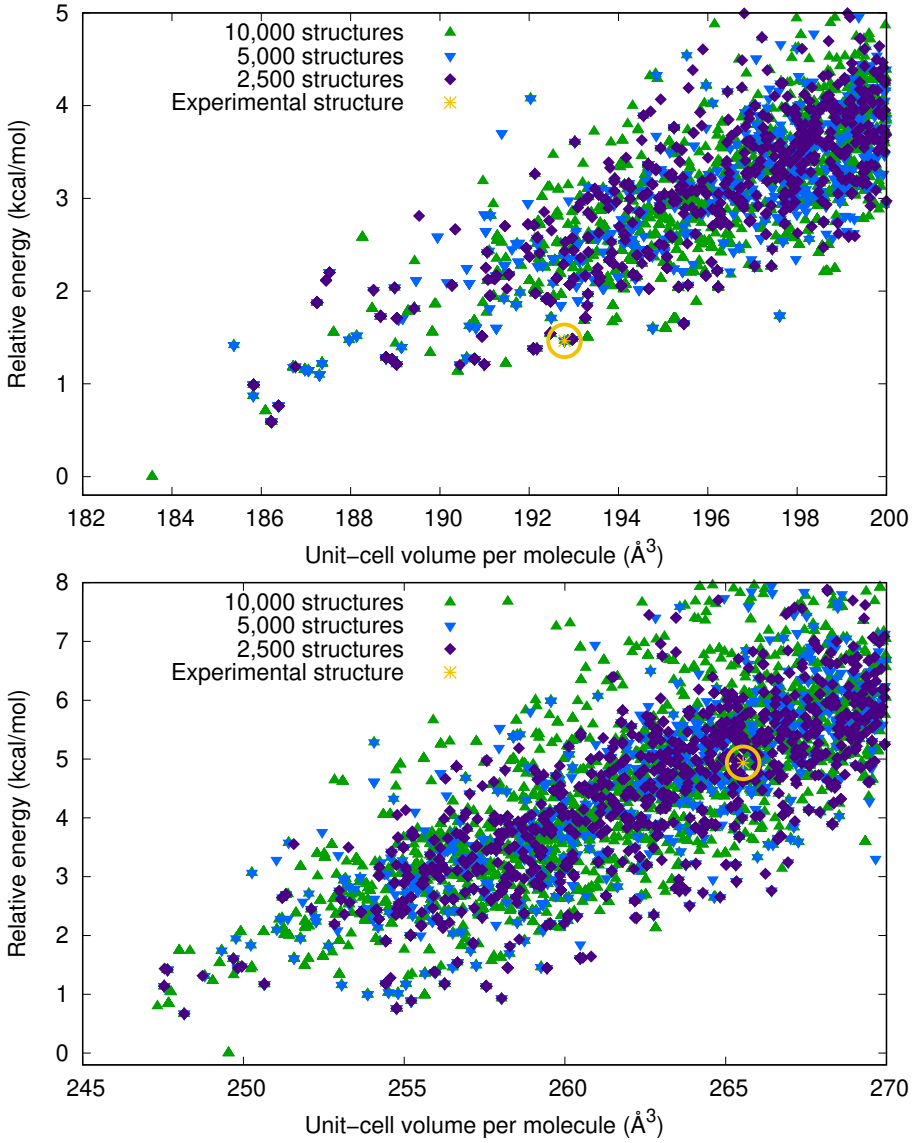


Figure 3.3: Crystal energy landscapes for targets IV (top) and X (bottom), showing where the experimental structure ranks amongst the candidates generated.

energies are corrected to account for differences in molecular conformation. Of the two potential conformers for target XXIX, shown in Figure 3.1, conformer (b) was predicted to be more stable by 2.9 kcal/mol. Thus, the landscapes containing only conformer (a) were shifted up by this value, and the landscapes with the one-to-one ratio of conformers shifted up by half of that. Omitted are higher bands of structures with relative energies between 15 and 20 kcal/mol, as they are likely caused by USPEX finding high-energy local minima in the sampling landscape. The vast majority of structures were not in these bands, indicating that the algorithm was not trapped there for any significant time and the structures can be ignored.

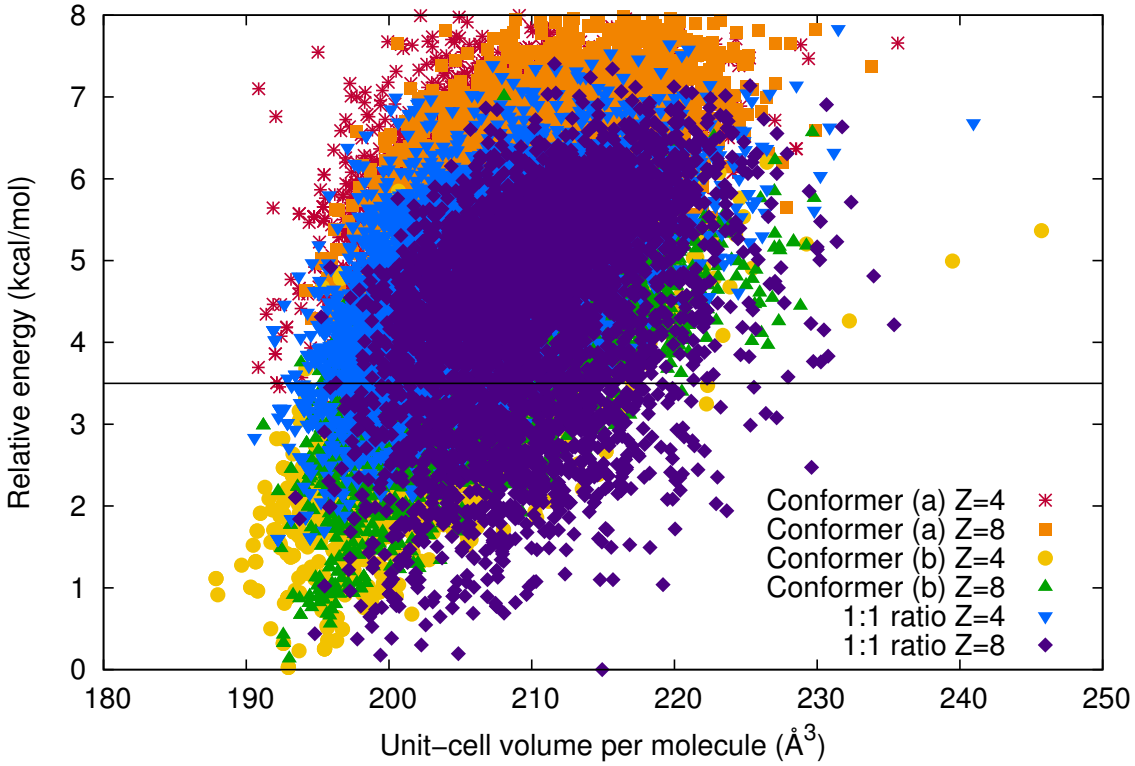


Figure 3.4: Crystal energy landscape of methyl anthranilate obtained with DMACRYS. Only structures with relative energies falling below the horizontal line were carried forward to single-point energy evaluation with DFT.

Upon first inspection, it is apparent that the conformer energy difference has a significant impact on the landscape, as none of the lowest-energy structures contain only conformer (a). In particular for relative energies below 1 kcal/mol of the minimum, the majority of structures either contain solely conformer (b) or are mixed with (a) in the 1:1 ratio. This separation, as we will see, became more pronounced as the level of theory increased throughout the protocol.

All structures below the 3.5 kcal/mol cutoff, for a total of 4609 structures across the six USPEX runs, were carried forward for DFT single-point energy evaluation. This particular cutoff was selected based on the results of a previous (unpublished) energy-ranking benchmark performed by Alberto Otero-de-la-Roza, which assessed the mean absolute errors (MAEs) in relative energies of a range of crystal structures provided by various low-cost energy methods with respect to high-level DFT results for selected compounds taken from the first five CSP blind tests, as well as several helicene compounds. Analysis of the results provided a series of confidence intervals, indicating the probability that the true DFT minimum-energy structure would lie below a particular cutoff on the low-cost landscape. For DMACRYS at 99% confidence, this cutoff was 3.5 kcal/mol. It was in this same benchmark that the TINY, MMFF94, and OPLS-AA force fields were first tested, with

TINY having the lowest MAE of these three methods for the largest variety of molecules. This was a convenient result but, as the discussion for the second project shows, we eventually found that TINY was a poor choice of FF overall.

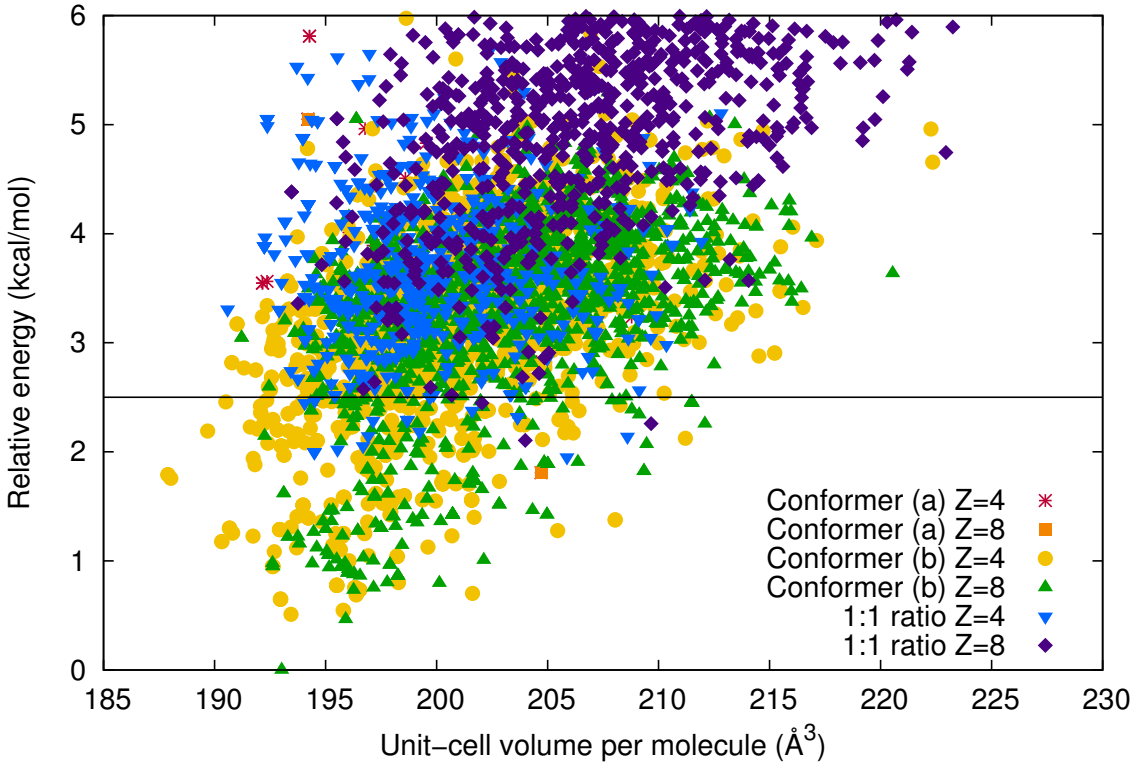


Figure 3.5: Crystal energy landscape of methyl anthranilate obtained with B86bPBE-XDM/PAW single-point energy evaluation at the DMACRYS geometries. Only structures with relative energies falling below the horizontal line were carried forward to DFT geometry optimization. Note that the majority of the DFT single-point energy calculations were performed by Adrian Rumson.

The landscape resulting from evaluating the DFT single-point energies of DMACRYS geometries is presented in Figure 3.5. Here, we see one structure with a significantly lower energy differentiate itself from the pack for the first time. This structure eventually becomes the lowest-energy structure in the DFT geometry optimization stage as well. Another feature of note in the single-point landscape is the different relative energies of structures. While they are similar for the structures consisting only of conformer (b), the conformer (a) and mixed structures are generally higher in energy. The relative energies changing between steps is natural in a CSP protocol, especially given the differences between methods. As mentioned previously, the gap only widened in the competition between conformers. This is summarized in table 3.2, where the numbers of structures from each USPEX run that passed onto the next stages of CSP are shown.

Table 3.2: Numbers of structures that fall below the energy threshold at each step and were carried forward in the CSP protocol. The notation QE//DMACRYS indicates that single-point energy calculations were performed using B86bPBE-XDM with Quantum ESPRESSO at the DMACRYS geometries.

Conformer	Z	DMACRYS < 3.5 kcal/mol	QE//DMACRYS < 2.5 kcal/mol
OMe-NH_2	4	27	1
OMe-NH_2	8	7	-
CO-NH_2	4	1380	177
CO-NH_2	8	1451	133
1:1 Ratio	4	668	13
1:1 Ratio	8	1076	3

3.1.4 DFT Optimization and Comparison to Experiment

Based on the single-point energies, the 327 structures with relative energies below 2.5 kcal/mol of the minimum were carried forward to have their geometries fully optimized by DFT. As indicated in Table 3.2, only one structure containing only conformer (a), and a total of 16 structures involving the mix of conformers, were beneath the threshold.

A comparison of the single-point landscape and geometry-optimized landscape is seen in Figure 3.6. These landscapes also show a heat map of the relative energies, enabling us to visualize whether there was significant reordering in the relative energies between the steps. Illustrated with the gradient of dark purple representing the lowest energy, to red, then finally yellow for the highest energy, we can see in the single-point landscape that the colour scheme is generally kept. The minimum-energy structure is also the same in both cases. This is ideal, as a large rearrangement would mean that there was a significant flaw in our methods. There are some structures in the single-point landscape that optimized to lower-energy structures, however, this is reasonable since DFT is a more robust method for these types of optimizations. One way to address this is to find a better energy-ranking method for the initial structure generation with USPEX. Contrarily, seeing low-energy structures from the single-point landscape re-optimize to much higher energies would also be a cause for concern, as it would indicate error in the FF geometries. Overall, the agreement between the single-point and DFT-optimized results was good, with no major changes in the landscape. Because of this, we can be confident in our methods up to this point that no low-energy structures had been missed. While the landscapes shown are based solely on thermodynamics, the DFT-optimized landscape is specifically static-lattice, so thermal expansion is ignored. Including vibrational effects from thermal expansion would shift the structures to larger volumes, and cause a slight reordering in relative energies. This reordering

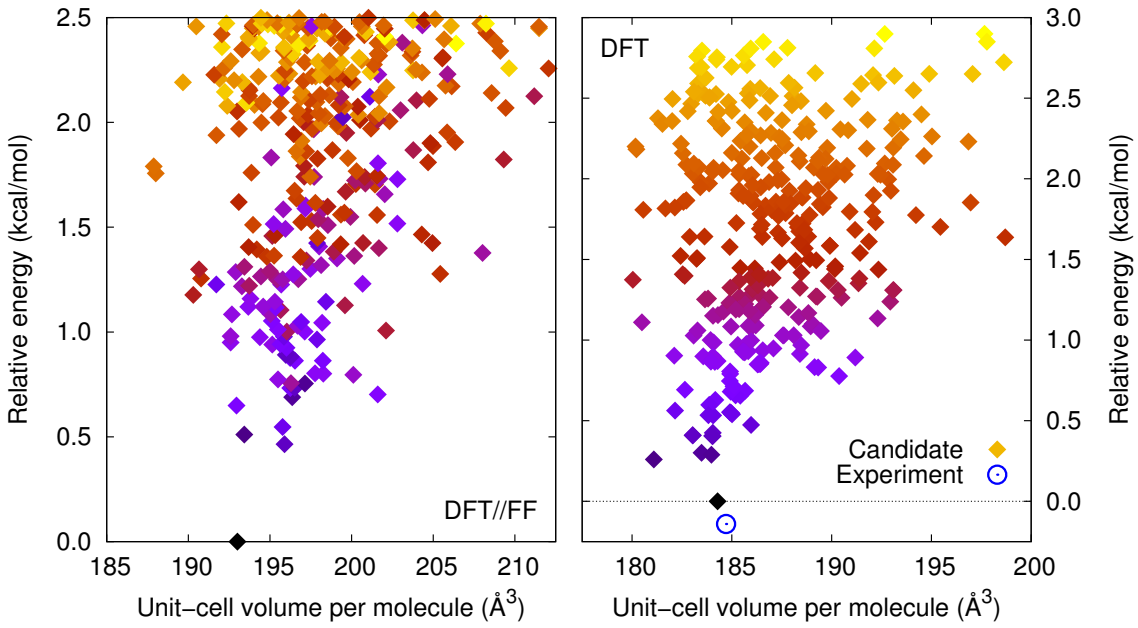


Figure 3.6: Crystal energy landscape of methyl anthranilate obtained from B86bPBE-XDM/PAW//DMACRYS (left) and from full B86bPBE-XDM/PAW geometry optimization (right). The points are coloured according to the relative energies obtained from the DFT geometry optimizations, performed by Adrian Rumson, Erin Johnson, and Alberto Otero-de-la-Roza.

is typically small, resulting in energy differences of less than 0.5 kcal/mol.[4] Since kinetic effects are ignored completely, the true experimental crystallization behaviour cannot be predicted from the landscape alone as this would also depend on factors such as solvent choice, and the specifics of the crystal nucleation and growth method, in addition to the thermodynamics.[73] While kinetic effects will not change the crystal energy landscape itself, it does change which structures can be experimentally accessed – the crystallization of a metastable structure on the landscape may be kinetically favourable, but investigating this would require synthesizing the crystal in a lab.

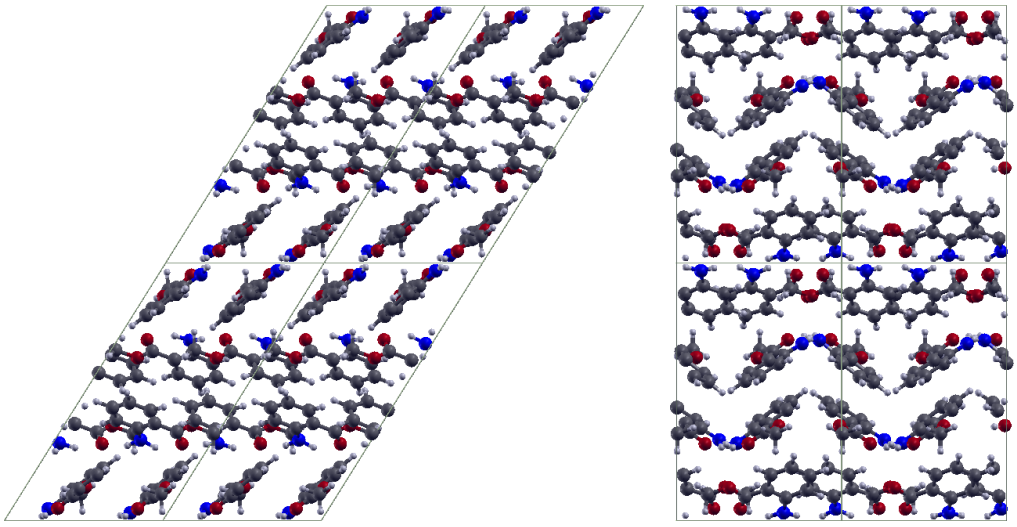
From the DFT results, a ranked list of the ten lowest-energy candidates and a landscape of 1500 structures were sent to the CCDC as our submission for the blind test. These are summarized in Table 3.3. While the accuracy of B86bPBE-XDM is expected to be quite high, the range of energies spanned by these candidates roughly corresponds to the expected uncertainty of the method. It has previously been used to confirm the relative stabilities of diamond and graphite, where the calculated ΔG for the conversion of diamond to graphite was within $2\times$ the experimental uncertainty when compared to the experimental value of -3170$\pm$150 J/mol.[74] Additionally, the ability of B86bPBE-XDM to reproduce the relative lattice energies of the EE14 benchmark set of homo- and hetero-chiral crystals was assessed in a study previously published by the Johnson group.[1] The MAE of the relative energies was determined to be 2.1 kJ/mol (0.50 kcal/mol), making

Table 3.3: Ranked list of the 10 lowest-energy structures submitted to the CCDC, as well as the results of geometry optimization of the experimental crystal structure.

Structure	Space Group	Z	Volume (Å^3)	ΔE (kcal/mol)
Experimental	$P2_1/c$	12	184.71	-0.141
Predicted	$P2_1/c$	8	184.28	0.000
	$P\bar{1}$	4	181.09	0.260
	$P2_1$	8	183.98	0.289
	$P\bar{1}$	4	183.48	0.301
	$P2_1$	8	184.04	0.404
	$P2_1/c$	8	184.02	0.406
	$P2_1$	4	183.04	0.409
	$P2_1/c$	4	183.03	0.410
	P1	8	184.05	0.425
	$P2_1$	4	185.96	0.474

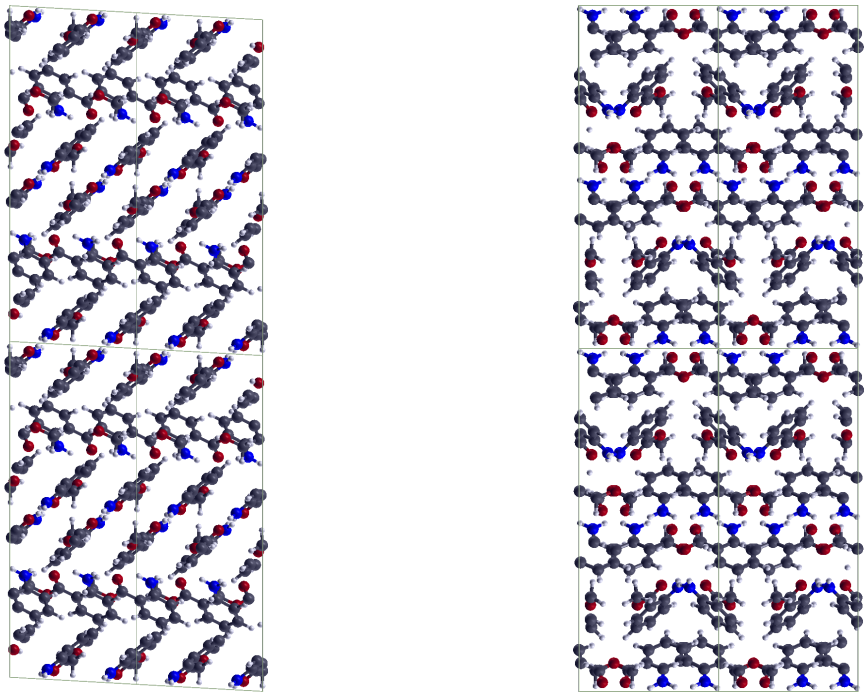
B86bPBE-XDM an effective choice of functional when used in a multi-level method akin to the one employed in this work.

The DFT-optimized minimum was a $Z = 8$ structure with the $P2_1/c$ space group with all molecules being conformer (b). The experimentally observed polymorph was revealed by the CCDC to be a $Z = 12$ structure, with the same space group and conformer (refcode: FASMEV). A B86bPBE-XDM/PAW geometry optimization of this experimental structure showed that it was more stable than our predicted minimum-energy structure by 0.14 kcal/mol per molecule. The packing motifs, however, were quite similar. Shown in Figure 3.7, the similarities between the predicted xy-plane and the experimental yz-plane, and vice versa, are evident. Two hydrogen-bonding patterns can be observed, as well as π-stacking. The similarity in packing can also be seen in Figure 3.9. Using the COMPACK algorithm for structure comparison gives a 20/20 molecule match with an RMSD(20) value of 0.441 Å. Only when the cluster size is expanded to 50 molecules does the structural mismatch become apparent. With this size, only a 45/50 match is obtained, with an RMSD(45) of 0.536 Å. A comparison of the simulated PXRD spectra provided by the CCDC, as well as those simulated in critic2 for our predicted structure and the true experimental structure are shown in Figure 3.8. From the overlay of the spectra, a clear similarity can be identified between the PXRD pattern of the reference provided by the CCDC and that of the experimental match. The predicted structure in this work, however, matches the reference and experimental structure patterns quite poorly. Given the structural similarities between our predicted structure and the experimental one, as well as the lower energy according to B86bPBE-XDM/PAW, we believe that this experimental structure could have been a leading candidate if our search had been expanded from $Z = 8$ to $Z = 12$.



a) Predicted xy-plane

b) Predicted yz-plane



c) Experimental yz-plane

d) Experimental xy-plane

Figure 3.7: Comparison of our predicted minimum-energy structure (a,b) and the experimental crystal structure (c,d) of methyl anthranilate (CCDC refcode: FASMEV).

Considering a larger search space may be simple to implement, but it is difficult to effectively execute. Just under 70 core years were spent on this project, with 64 of these core years spent on the DFT structure optimizations alone. Expanding the search space makes the high-dimensional sampling landscape even more complex, introducing new local minima that the structure generator could explore. It would therefore need to be run for longer in order to account for the increased

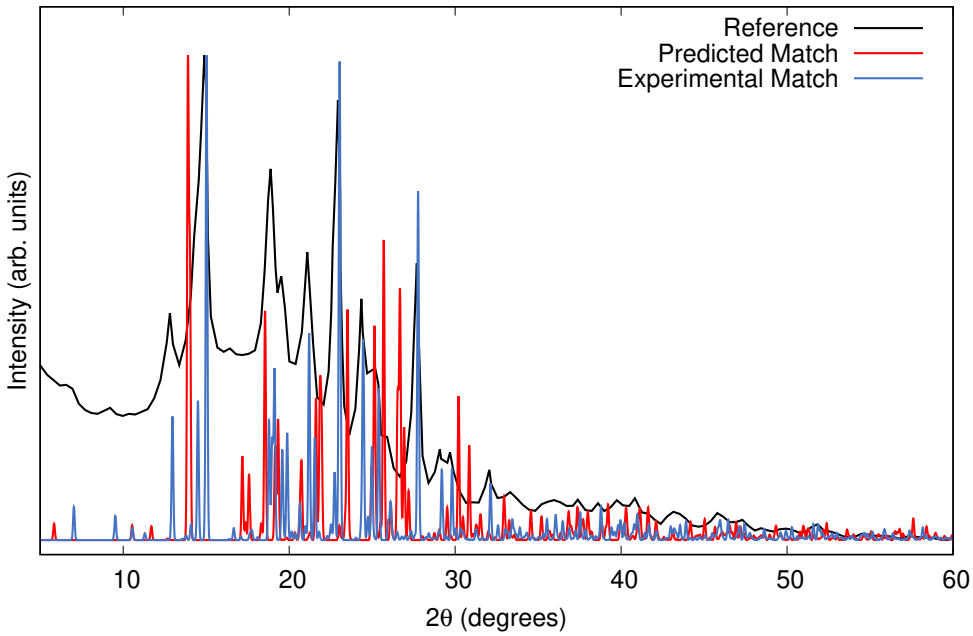


Figure 3.8: Overlay of simulated PXRD spectra of our predicted minimum-energy structure and the experimental crystal structure of methyl anthranilate (CCDC refcode: FASMEV) with the reference spectrum provided by the CCDC.

search space, likely doubling the time required for the structure generation step. However, the brunt of the computational cost would once again lie with the cost of DFT optimizations, which would now need to be performed on many more candidates with larger unit cells ($Z = 12$ vs. $Z = 8$). The system sizes in this work pushed the limits of what is feasible with the planewave implementation of our DFT method, so systems with $Z = 12$ would likely be unmanageable. The new implementation of B86bPBE-XDM in FHI-aims, where computational time scales much more reasonably with number of atoms, could be employed to address this problem.[75]

3.2 FF Benchmark

While there is a strong research precedent for the energy ranking step of CSP both within and outside of the Johnson group,[4,13] an effective structure generation protocol remains key. Here, the sampling and preliminary energy ranking methods must be accurate enough to ensure that the experimental structure is both found, and ranked with sufficiently low energy that it will move forward in the overall CSP protocol. One such method of energy ranking that has grown in popularity is creating a tailored force field for the specific molecule of interest, however this is not a trivial undertaking.[76] Especially given the increasingly complex nature of CSP targets and the high number of well-developed pre-existing force fields, this is not always a desirable or feasible option.

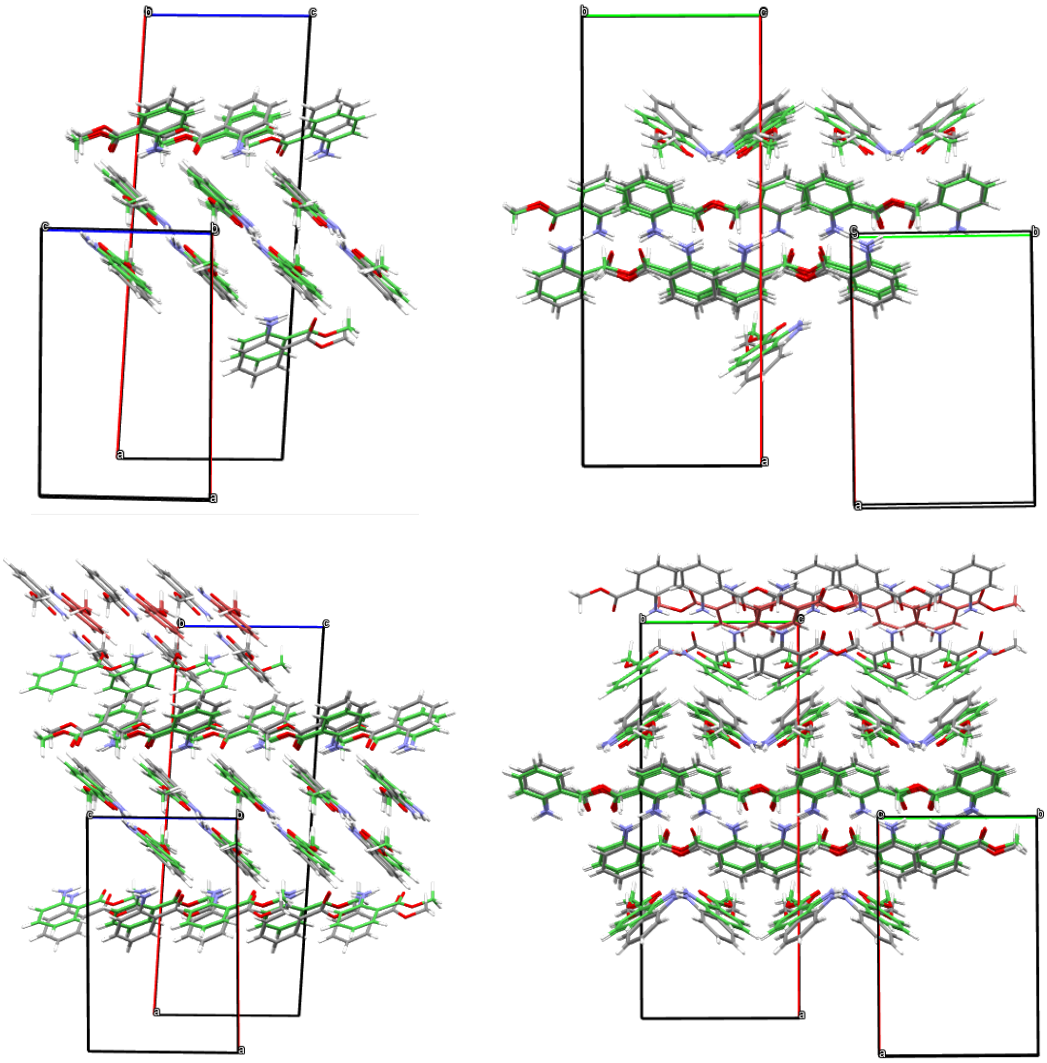


Figure 3.9: COMPACK overlays of the minimum-energy structure generated by our CSP protocol and the experimental crystal structure of methyl anthranilate (CCDC refcode: FASMEV). Top: a 20/20 overlay for cluster size of 20 molecules; bottom: a 45/50 overlay for a cluster size of 50 molecules. Matching molecules are shown in green and non-matching in red. Two orientations are shown for each overlay.

Thus, we have chosen seven accessible general or molecular (i.e. not parameterized specifically for biomolecules) force fields to evaluate the crystal landscapes of eighteen compounds. The selected compounds are from the PV17 benchmark set, plus 5-fluorouracil, as shown in Figure 3.10.[23] Each landscape produced was then analyzed to determine if either of two experimental polymorphs was found by USPEX coupled with the chosen force field.

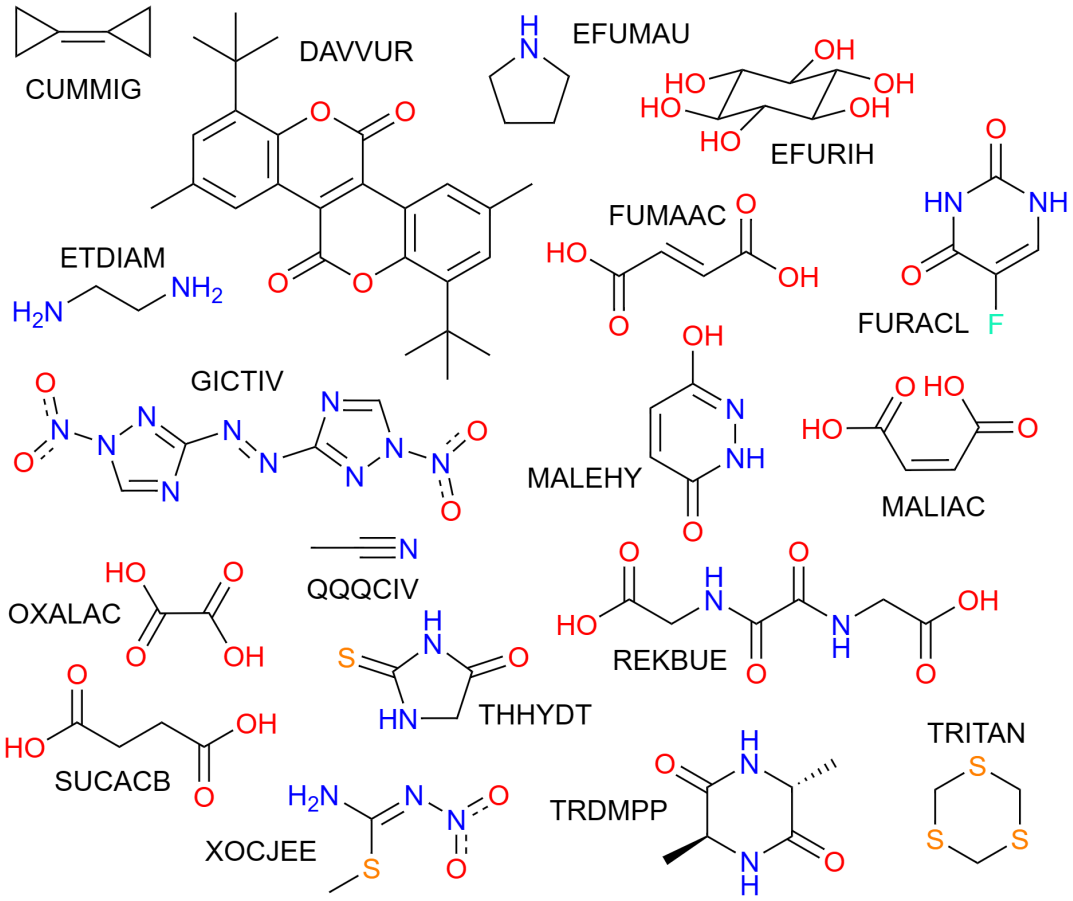


Figure 3.10: Structures of the 18 compounds forming our benchmark set, along with their CCDC refcodes.

3.2.1 Methods

For each structure and force field in the benchmark set, structure generation was completed with version 10.4 and 10.5 of USPEX. Due to the improved interface, version 10.5 was used for the GULP-implemented FFs. 126 runs were completed in total, each using equivalent USPEX inputs. A minimum of 10,000 structures were generated, with 100 structures per generation. The same collection of variation operators, random space groups, and antiseed settings from the BT7 runs was applied here. The number of formula units per unit cell, however, was varied to include $Z = 2, 4, 6,$ and 8. Molecules were kept rigid when generated by USPEX, then allowed flexibility during the FF geometry optimization. The TINY, MM3, MMFF94 and OPLS-AA force fields were used with Tinker version 8.10.2, and DREIDING, UFF, and GAFF were used with GULP version 5.2.

3.2.2 Atom Typing

There are existing codes, such as LigParGen or Antechamber, that can be used to automatically atom type molecules for specific FFs.[77,78] The specificity of the available FFs and nature of the

output made by these codes, however, made them ineffective for use in this benchmark set. Thus, atom types were manually assigned for each of the compounds, for all the FFs used. Force field atom types (FFATs) are used to specify parameters for elements in different chemical environments, so the complexity of assigning FFATs is dependent upon how the force field was designed. For general-purpose force fields like TINY or UFF, atom types are only distinguished by element and the hybridization, or number of attached atoms. In others, such as OPLS-AA, there are over 10 atom types for the oxygen in a carbonyl depending on whether it is part of a carboxylic acid, ester, amide, nitrogenous base, etc. Each molecule was examined and the best atom types were selected based on their descriptions in the atom type parameter files. While tedious, this only needs to be done once at the outset in preparation for the USPEX run, since the atom types will be automatically assigned to the generated structures as long as the molecule has been atom typed.

To prepare properly atom-typed molecular inputs for USPEX, molecule structures were converted to the format required using the Zmatrix 2.0 tool provided by USPEX. This is called the MOL_1 file, and it provides the building block for crystal structures to be generated. The general format of the MOL_1 file has the first two lines being the name and number of atoms in the molecule, followed by the specific structural information. The first column denotes the element, then the next three contain the Cartesian coordinates for atom positions. The following three columns contain a Z-matrix to define connectivity, followed by a column indicating torsional flexibility. Molecules were kept rigid while USPEX generated structures, then allowed to relax during optimization, so the first three flags were set to “1” and the rest as “0”, as per the USPEX manual. When Tinker is used, an additional column is added to set the atom type, and these had to be manually inputted. Atom types for the TINY FF are determined solely based on element and number of attached atoms. The next force fields, MM3 and MMFF94, have atom types defined by hybridization and functional group, so the more specific functional group types were selected where applicable.

When interfaced with GULP, the atom typing procedure is slightly more complex. Atoms with the same type must be noted in the element column of the MOL_1 file, and this is done to assign atom types in the GULP input. For example, all equivalent oxygen atoms, such as all carbonyl oxygens, are represented by O_1 in the MOL_1 file, and all alcohol oxygen atoms are represented by O_2. These designations, of course, depend on the specific atom types defined in the force field. The atom types used in the MOL_1 file must also be listed in the GULP input file in the “Specific" directory described in the next section. For atom typing with GAFF, the partial charges must be generated by the user. These are obtained by running an RESP calculation with the Gaussian software,[47] and inserting the partial charges as an extra column in the MOL_1 file. A summary of the number of atom types used with each force field, for each of the 18 compounds considered,

is given in Table 3.4. The relative complexities in atom typing can be observed based on the number of different atom types required by different FFs to model any given molecule. MALEHY, for example, contains 12 atoms in its molecular structure, and requires 10 unique atom types in OPLS-AA and only 5 unique types in UFF.

Table 3.4: Number of atom types used with each of the force fields for selected compounds.

Refcode	TINY	MM3	MMFF94	OPLS	DREIDING	UFF	GAFF
CUMMIG	3	3	3	3	3	3	3
DAVVUR	5	6	6	8	5	6	7
EFUMAU	3	3	4	4	4	3	4
EFURIH	3	4	4	4	4	3	4
ETDIAM	3	4	5	4	4	3	4
FUMAAC	4	6	6	6	5	4	6
FURACL	5	8	7	8	6	5	7
GICTIV	5	7	8	10	5	5	6
MALEHY	6	9	9	10	6	5	9
MALIAC	4	6	6	7	5	4	6
OXALAC	4	4	4	4	4	4	4
QQQCIV	4	5	4	4	4	4	4
REKBUE	6	9	9	10	7	6	8
SUCACB	5	6	5	6	6	5	6
THHYDT	6	8	9	7	7	6	7
TRDMPP	5	6	6	7	6	5	6
TRITAN	3	3	3	3	3	3	3
XOCJEE	7	8	8	9	8	7	9
Number of atom types available	46	164	212	906	49	127	83

3.2.3 IMPROVEMENT OF INTERFACES

Throughout the work with USPEX, improvements needed to be made with regards to the interface between structure generator and energy minimizer, as well as with the inputs used to perform the geometry optimizations. USPEX has been interfaced with a selection of different energy minimization codes, and is packaged with scripts to perform these geometry optimizations in tandem. Unfortunately, these scripts do not give meaningful results for any systems, even the test systems for which they were created. Thus, they were rewritten to perform geometry optimizations that met our standards. In doing this, the primary challenge was ensuring that these rewritten scripts would still be compatible with the program, as there is no documentation describing how the various interfaces work.

When running an USPEX job, a directory (called “Specific") is made to contain all the files that will be necessary to run the geometry optimizations, besides the specific structure file. This directory includes the script for running the optimization, the force field parameter file, and any other input files that will be required by the minimization program. Then, when the directory where the optimization will be performed is created (the calculation folder), the contents of the *Specific* directory will be copied into it. Finally, the structure to be optimized is added and the optimization script is run. Once the geometry optimization has completed, USPEX reads the output files to extract the structural and energetic information and convert it to the USPEX output format.

In all recent versions of USPEX, the source code and most additional scripts are hidden, and unable to be read or edited. This includes the scripts written to read the optimized outputs, so the only way to rewrite the optimization scripts effectively is to mimic the format of those provided by USPEX. Otherwise, there is no way to see what specific files or information USPEX reads from the output. This was especially problematic when there were multiple output files created, such as the fractional and Cartesian structure outputs given by Tinker. The only other possibility for viewing the code was to download an outdated version of USPEX that still contained human-readable code. This was done to help guide the script-writing process, but is only useful if the specific interfaces were not updated in the subsequent versions. The interfaces are extremely specific, so any deviation from precisely what USPEX expects in an output causes the program to terminate with an error. One example concerns the Tinker structure file written by critic2 – this file was unreadable to USPEX because the line containing the lattice parameters did not have enough empty space trailing it, even though the file is readable by Tinker itself. While error codes are shown that cite specific lines of scripts and processes, none of these are accessible for troubleshooting. Additionally, Python scripts meant to be run “as is" are provided to run the geometry optimizations on a remote machine, but these needed to be completely rewritten in order to function. This difficulty dealing with the source code is one of the reasons why USPEX is sub-optimal for CSP.

3.2.4 STRUCTURE COMPARISON

An in-house script was used to screen the output from USPEX. The output is given by a concatenated CIF-format file containing all structures generated, and an “individuals" file that lists structure ID, composition, energy, density, and other information. Data from the individuals file was used to remove unrealistic structures (density < 0.7 g/mL, lattice energy $>$ 5,000 kcal/mol). Additionally, any groups of structures with identical Z, density, and lattice energy values were pruned to one representative. Subsequently, nearest neighbours on a list of energy-ordered structures were compared using the critic2 program.[60] If a pair had a powder diffraction pattern difference (POWDIFF) < 0.002, the structure with the higher energy was removed. The remaining structures

were screened, in order of increasing energy, for a match to the known experimental polymorph crystal structures (listed in table 3.5) using the variable-cell powder difference method (VC-PWDF),[64] also implemented in critic2.

Structures with a VC-PWDF value < 0.03 were checked by comparing the VC-corrected structure to the target structure with the COMPACK algorithm using Mercury.[79] The selected COMPACK settings were a tolerance of $\pm 30\%$ for distances and $\pm 30^\circ$ for angles, ignoring bond types, bond counts, and numbers of bonded hydrogen atoms. If the comparison yielded a 20/20 match with RMSD(20) < 0.3 Å, then the candidate structure was recorded as a match and screening for that target was ceased. Cases where RMSD(20) values were between 0.3 and 0.5 Å were checked manually to ensure a match by geometry optimization of the candidate structure with FHI-aims.[80] These calculations used the B86bPBE-XDM dispersion-corrected density functional, with the recommended combination of light basis set and dense integration grid (lightdense), and a convergence criterion of 0.01 eV/Å in the forces.[81] If the DFT-optimized candidate structure clearly matched the target structure, with RMSD(20) < 0.3 Å, then it was recorded as a "hit" in table 3.5. Alternatively, if it was found that the candidate structure did not match the target structure adequately, then screening of the candidate list resumed.

3.2.5 RESULTS

While finding the experimental match is important, it is just as important for the relative energy of that match to be low enough that the structure will be carried forward in the CSP protocol. Summarized in table 3.5 are the identified matches for the benchmark, reported with the energies relative to the minimum-energy structure from that landscape. Additionally, structures that became identified as matches with our criteria following a DFT optimization are noted in the table. On first glance, the FF with by far the greatest success was GAFF. This method found a match for 28 out of the 36 polymorphs, all of which were within 8 kJ/mol of the minimum. The next most successful FFs in identifying matches were OPLS-AA and MMFF94, with 24 and 22 matches, respectively. These FFs, however, had cases were the matches were ranked with extremely high energies. The OPLS-AA matches for EFURIH are particularly egregious cases, with one match being found 40.30 kJ/mol above the minimum, and the second match being found over 80 kJ/mol above the first, and 126.1 kJ/mol above the minimum overall. This is an extremely large energy difference and, unless there was a second energy ranking analysis of all structures in the landscape with another method, it is highly unlikely that these would make the threshold for further analysis. While the DREIDING and UFF force fields yielded good energies for their matches, they were only found for just over 30% of the polymorphs.

Table 3.5: Summary of results for identification of experimental polymorphs within the structures generated using USPEX and selected force fields. Cells with dark green shading indicate cases where a match was identified within the specified criteria of VC-PWDF < 0.03 and VC-RMSD(20) < 0.3 Å. Cells with light green shading indicate cases with VC-PWDF < 0.03 but VC-RMSD(20) > 0.3 Å, where subsequent geometry optimization with FHI-aims resulted in a structure match. In both cases, the reported number is the energy of the matching structure, in kJ/mol per molecule, above the minimum-energy structure found with that force field. Exes (✗) indicate no match to the target structure was found.

Refcode	TINY	MM3	MMFF94	OPLS-AA	DREIDING	UFF	GAFF
CUMMIG01	3.50	0.82	5.41	0.00	0.04	✗	0.10
CUMMIG02	4.10	3.27	4.72	3.84	0.48	✗	✗
DAVVUR01	0.00	0.00	3.39	7.92	✗	✗	2.58
DAVVUR	13.01	4.64	3.73	14.05	✗	✗	✗
EFUMAU03	✗	✗	✗	✗	✗	✗	✗
EFUMAU	✗	✗	✗	0.41	✗	✗	0.05
EFURIH04	✗	14.17	23.07	126.10	0.94	✗	7.71
EFURIH	25.33	0.00	13.25	40.30	0.25	0.56	7.73
ETDIAM16	✗	22.97	6.07	11.67	0.15	✗	0.79
ETDIAM18	1.39	19.42	3.96	✗	0.12	0.56	0.68
FUMAAC01	17.04	3.02	4.76	7.23	0.07	0.18	0.36
FUMAAC	✗	✗	✗	✗	✗	✗	✗
FURACL01	✗	✗	✗	✗	✗	✗	✗
FURACL02	✗	✗	✗	✗	✗	✗	✗
GICTIV01	✗	4.80	✗	✗	✗	✗	5.57
GICTIV	✗	5.67	✗	17.55	✗	✗	5.21
MALEHY10	8.27	35.67	0.00	14.77	0.13	✗	0.03
MALEHY12	✗	✗	1.48	✗	✗	✗	0.15
MALIAC12	✗	✗	37.84	✗	✗	✗	1.52
MALIAC13	✗	58.69	25.52	50.84	✗	✗	1.37
OXALAC03	✗	✗	✗	✗	✗	✗	✗
OXALAC04	✗	✗	31.21	17.34	0.04	0.14	0.11
QQQCIV01	✗	3.46	0.04	2.30	✗	0.09	0.05
QQQCIV08	✗	✗	0.03	2.12	0.02	0.05	0.00
REKBUE01	✗	✗	✗	14.10	✗	✗	2.10
REKBUE	✗	✗	16.27	0.71	✗	✗	0.56
SUCACB02	✗	2.40	7.84	1.11	✗	0.52	0.28
SUCACB07	✗	✗	13.18	9.53	✗	✗	0.37
THHYDT02	2.41	19.36	8.68	4.42	✗	2.05	0.16
THHYDT	8.50	✗	✗	✗	0.00	✗	0.00
TRDMPP01	✗	✗	18.92	1.53	✗	✗	0.37
TRDMPP02	✗	✗	13.24	1.80	✗	✗	0.05

Refcode	TINY	MM3	MMFF94	OPLS	DREIDING	UFF	GAFF
TRITAN03	✗	3.74	✗	8.45	0.10	0.08	✗
TRITAN10	✗	5.74	✗	6.09	0.11	0.09	0.01
XOCJEE01	✗	✗	✗	✗	✗	✗	3.32
XOCJEE	✗	✗	✗	✗	✗	0.25	6.57
Total matches	10	18	22	24	13	11	28
Match rate (%)	28	50	61	67	36	31	78
Average ΔE	8.36	11.55	11.03	15.17	0.19	0.38	1.71

More instances of very high energies can be seen in the matches identified for maleic acid, MALIAC12/13. For each match identified by any FF other than GAFF, the relative energies all exceeded 25 kJ/mol. This is the *cis*-isomer of butenedioic acid, and although the atom types specified the double bond at the centre of the molecule, there were structures where a conformer change to the more stable *trans*-isomer, fumaric acid, was observed. The presence of the *trans*-isomer in the landscape for what is supposed to be only the *cis*-isomer is likely what skewed the energies. As a result of these skewed energies, USPEX chose structures with the incorrect conformer as those with the “highest fitness" upon which variation operators are used, creating a positive feedback loop. Structures with the incorrect isomer were continually favoured by the algorithm, so fewer suitable structures with the correct isomer ended up being produced. While USPEX checks that molecules are intact upon relaxation, this conformer change was not identified as the connectivity is still identical and the molecule was not “broken". Ultimately, it should not have been possible for this conformer change to occur – it would require rotation around a double bond, which is physically impossible. Figure 3.11 shows one example of such a conformer change, where the resulting structure is a co-crystal containing both conformers. Given the high relative energies for the MALIAC12/13 matches for MM3, MMFF94, and OPLS-AA, it can be inferred that all three FFs allowed this forbidden rotation. This is disheartening for MMFF94 and OPLS-AA in particular, since they performed quite well if one only considers the number of polymorph matches found. Unfortunately, the relative energies of these matches were wildly inconsistent compared to the more successful GAFF, with an average ΔE of over 10 kJ/mol compared to GAFF's average of < 2 kJ/mol. Given that the matches for MALIAC12/13 with GAFF did not have an absurdly high energy, it is safe to conclude that there were no major physically-forbidden conformer changes with this run, and this can be confirmed by visualizing the structures.

Certain structures had very few matches identified throughout the benchmark, or none at all. Both polymorphs of 5-fluorouracil went unfound, as well as EFUMAU03, FUMAAC, and OXALAC03. An insufficient description of hydrogen bonding is likely the cause of the poor across-the-board performance for these latter three cases, particularly for EFUMAU03, where it is expected to be a

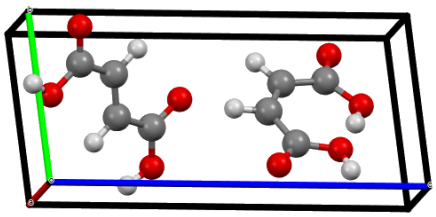


Figure 3.11: Example of an “impossible" co-crystal of maleic (right) and fumaric (left) acids generated by USPEX and relaxed with OPLS.

more important contributor to stability compared to EFUMAU. For the 5-fluorouracil polymorphs, FURACL01/02, the poor performance is most likely due to the treatment of fluorine in the FFs. When fluorine is present in organic molecules, halogen bonding can occur – this refers to the interaction between the lone electron pair of a Lewis base and the positively charged σ hole of a halogen atom.[82] Comparable to hydrogen bonding, halogen bonding has been seen in crystalline materials, but is poorly modeled by classical force fields. The anisotropy and resulting σ hole in the electrostatic potential around the fluorine atom is missing from the FF descriptions.[83] Even GAFF, with the most sophisticated description of electrostatics, was unable to find a match to either experimental polymorph for this molecule. Additionally, weakly stabilizing halogen-halogen interactions are poorly described by the FFs. This is more important for FURACL01, where these interactions contribute more to the overall stability of the polymorph.[84]

As mentioned, the sophistication of the non-bonded terms in an FF has the greatest impact on its accuracy. Each of the seven FFs used in this benchmark involves a different combination of van der Waals approximation and description of electrostatics. Whether or not cross-terms were included and if hydrogen bonding was specifically accounted for will also have an impact on the accuracy for this benchmark set, particularly since many of the experimental structures in the set have hydrogen bonding present. Finally, the quality of the reference data used to parameterize the FF should also be considered. The FFs described were primarily developed between the 1980s and early 2000s, so most of the parameters were derived from experimental data or poor (by today’s standards) computational methods like HF or MP2. These varying descriptions are summarized in table 3.6, and can be used to interpret the relative successes of each FF.

3.2.6 Overview and Recommendations for CSP

Unsurprisingly, the TINY force field showed the worst performance of the methods considered. This FF was designed only for crude structural optimizations, and despite its early success being used for the BT7 target compound, it struggled to reliably find experimental polymorphs for this larger benchmark set. In addition, the relative energies of the identified matches were very inconsistent. These ranged from the minimum-energy structure to over 25 kJ/mol above the minimum. While

Table 3.6: Summary of non-bonded descriptions, inclusion of cross-terms, and quality of reference data for the seven FFs used in this study.

Force field	vdW potential	Electrostatic description	Explicit H-bonding?	Cross terms?	Reference data
TINY	LJ 6-12	None	No	No	N/A
MM3	Hill Exp-6	Coulombic, bond dipoles	Yes	Yes	Experiment
MMFF94	Buffered 14-7	Buffered Coulombic, partial charges	Modified vdW	Yes	Experiment, MP2/6-31G*
OPLS-AA	LJ 6-12	Coulombic, partial charges	In vdW	No	Experiment, HF/6-31G*
DREIDING	LJ 6-12	None	Yes	No	Experiment
UFF	LJ 6-12	Coulombic, charge equilibration partial charges	No	No	Experiment
GAFF	LJ 6-12	Coulombic, RESP partial charges	In vdW	No	MP2/6-31G*, MP4/6-311G*

TINY is extremely simple and convenient to atom type, it is at the cost of poor performance. The DREIDING and UFF force fields showed similarly poor performances, as they are not specialized for any particular material, and are designed for predicting reasonably accurate properties for novel combinations of most elements in the periodic table. Both include only generic parameters based on experimental data. While the performance of DREIDING and UFF was tested for organic molecules when they were being originally developed, they were also tested for their performances modelling organometallic compounds and certain biomolecule complexes. It was similarly trivial to atom type molecules with DREIDING or UFF since the atom types are only determined by element and hybridization, but their poor performances overall severely outweighs the simplicity.

It should be noted, however, that the matches that *were* found by DREIDING and UFF had very low relative energies, with an average ΔE less than 0.5 kJ/mol for both. Finding quality, low-energy structures should guide USPEX towards finding more, so it is interesting that there were so few matches considering the energies. Due to the lack of high-energy matches, it can be inferred that the matches identified were generated primarily from the variation operators, and were not found by chance via random generation. This indicates that more matches could have been found if, perhaps, the search had gone on for longer. In true CSP this could be feasible, but in comparison to the similarly-implemented GAFF that was much more successful at identifying low-energy matches in the same amount of time, it makes more sense to use GAFF over either DREIDING or UFF for organic molecules. If metals or heavier atoms were included, however, DREIDING or UFF would

likely be a feasible choice since GAFF is only parameterized for elements H, C, N, O, S, P, and the halogens.

The next most successful FF was MM3 which, alongside MMFF94 and OPLS-AA, has been shown to give good results for small organic molecules.[85,86] It makes sense that MM3 showed a jump in the number of matches identified compared to the general FFs, since it was designed specifically to replicate properties of organic molecules. Likewise, the relative success of MMFF94 was expected, since it is meant to be an improvement upon the MM3 force field. These both have specific descriptions of hydrogen bonding, with either a designated energy term or specialized parameters incorporated into the vdW potential. These were also the only two FFs to include cross-terms, which should yield more realistic geometries. Ultimately, the performances of these FFs could make them useful for CSP, but the relative energies were generally quite poor, averaging about 10 kJ/mol.

While we saw that GAFF had the greatest success, likely due to its description of electrostatics, it also used the highest-quality reference data. Torsions were parameterized using energy data from MP4/6-311G*, which was the highest level of theory for any of the parameters computationally determined. There was no explicit hydrogen bonding potential used, however GAFF incorporates these interactions in the vdW parameters so it is not completely ignored. The atom typing was a significant undertaking in this project, but in true CSP it would only need to be done for one or two molecules. With GAFF, an extra step must be added to the workflow to generate RESP charges, but this is so trivial to complete that it should not be considered a drawback in practical use. For this reason, GAFF is the recommended force field for CSP at this time.

USPEX is an extremely popular program for structure generation, but it is not ideal for CSP with molecular crystals. In addition to the difficulties relating to the interfaces with energy minimization codes, USPEX is quite inefficient. Even when employing measures to diversify the population and reduce the number of duplicates, there were still up to 40% of structures in any given run that ended up being removed because they were duplicates. This wastes computation time and, while it is unreasonable to expect no duplicate structures whatsoever, a better way to screen for duplicates throughout the structure generation stage would be ideal. Ultimately, the benchmark provides an effective framework of comparison for a series of general FFs, so more informed decisions can be made when selecting an FF for CSP. We can see how different descriptions of non-bonded terms had an effect on accuracy, and these insights can be applied to FFs that were not included in the benchmark as well. A robust description of electrostatics and high-quality reference data parameterized for molecular systems should be key features to look for when choosing a FF to perform CSP.

CHAPTER 4

CONCLUSIONS AND FUTURE WORK

4.1 CONCLUSIONS

The work presented herein highlights both strengths and weaknesses of our crystal structure prediction (CSP) protocol, with a focus on the structure generation stage. With the submission to the 7th blind test, we aimed to assemble an effective CSP protocol that could take advantage of the high-level energy-ranking strengths of the Johnson group. With the experience in performing full CSP gained throughout this project, we identified the need for improvement in our structure generation methods. Thus, a benchmark of accessible force fields (FFs) paired with the USPEX evolutionary algorithm was completed to compare the performance of seven FFs in finding the experimental structures of molecules comprising the PV17 benchmark set plus 5-fluorouracil.

In the blind test submission, USPEX, interfaced with the TINY force field in Tinker, was used to generate crystal structures containing both conformers of compound XXIX, or methyl anthranilate. An assessment of USPEX input parameters was completed prior to this, allowing for the determination of optimal parameters to promote structure diversity. After being pruned for duplicates, structures were reoptimized with DMACRYS and all structures within 3.5 kcal/mol of the minimum had their B86bPBE-XDM/PAW single-point energies evaluated at these geometries. Based on these energies, the structures below 2.5 kcal/mol of the minimum were selected to have their geometries fully optimized by DFT, using the same method as the single-points. The predicted minimum-energy structure was a $Z = 8$ structure containing only conformer (b) of the target compound, with the $P2_1/c$ space group. While this was not an exact match of the experimental structure, a $Z = 12$ structure with the same space group and conformer makeup, the packing motifs were extremely similar, and gave a 20/20 COMPACK match with an RMSD(20) of only 0.441 Å. The experimental structure had a lower DFT energy than our predicted one, so it is believed that the experimental structure would have been found had the search been expanded to $Z = 12$. This

highlighted the strength of the DFT ranking methods, and led us to the development of the FF benchmark.

The structure generation benchmark of different FFs coupled with USPEX showed that GAFF is the optimal choice of generic FF for CSP at this time. GAFF showed both a high match rate, finding nearly 80% of experimental polymorphs in the benchmark set, as well as ranking these matches with low energies. This contrasts the MM-series FFs and OPLS-AA, which were successful in finding matches but gave high relative energies. Conversely, the DREIDING and UFF fields identified very few of the experimental polymorphs, but those found were well-ranked with low energies. The sophisticated description of electrostatics in GAFF was likely the largest contributor to its success, along with the inclusion of hydrogen bonding parameters and higher-quality reference data.

4.2 FUTURE WORK

Based on the results of the FF benchmark, there are many apparent avenues for extending the work on structure generation. The first of these is to find an alternative to USPEX. While it is the most popular non-proprietary structure generation code, it is extremely tedious to use. It was inefficient, produced many duplicate structures, and is poorly documented for interfacing with most energy-minimization codes. The performance of other evolutionary algorithms could be investigated, as well as other sampling methods such as basin-hopping.

The scope of FFs included in the benchmark could also be expanded. We limited our study to FFs implemented in either Tinker or GULP, but FFs implemented in LAMMPS[87] or CP2K[88] could be assessed with USPEX as well. There are numerous FFs available through many different implementations, so there are many combinations of structure generator and relaxation method theoretically available. The creation of a tailor-made force field is another option for CSP, and the feasibility of parameterizing a high-quality non-transferable FF could be investigated. One glaring deficiency in the FFs benchmarked was the poor-quality reference data. While the methods used to acquire reference structural information and force constants were advanced at the time the FFs were developed, computational methods have seen significant advances in the past 30 years. With the recent interest in machine learning, FFs could be reparameterized with higher-quality reference data, while still employing the effective functional forms described throughout Chapter 2.

BIBLIOGRAPHY

[1] LeBlanc, L. M.; Otero-de-la Roza, A.; Johnson, E. R. Composite and Low-Cost Approaches for Molecular Crystal Structure Prediction. *J. Chem. Theory Comput.* **2018**, *14*, 2265–2276.

[2] Abramov, Y. A. Current Computational Approaches to Support Pharmaceutical Solid Form Selection. *Org. Process Res. Dev.* **2013**, *17*, 472–485.

[3] Bauer, J.; Spanton, S.; Henry, R.; Quick, J.; Dziki, W.; Porter, W.; Morris, J. Ritonavir: an extraordinary example of conformational polymorphism. *Pharm. Res.* **2001**, *18*, 859–866.

[4] Nyman, J.; M. Day, G. Static and lattice vibrational energy differences between polymorphs. *CrystEngComm* **2015**, *17*, 5154–5165.

[5] Aaltonen, J.; Allesø, M.; Mirza, S.; Koradia, V.; Gordon, K. C.; Rantanen, J. Solid form screening – A review. *Euro. J. Pharm. Biopharm.* **2009**, *71*, 23–37.

[6] Reilly, A. M.; Cooper, R. I.; Adjiman, C. S.; Bhattacharya, S.; Boese, A. D.; Brandenburg, J. G.; Bygrave, P. J.; Bylsma, R.; Campbell, J. E.; Car, R.; Case, D. H.; Chadha, R.; Cole, J. C.; Cosburn, K.; Cuppen, H. M.; Curtis, F.; Day, G. M.; DiStasio Jr, R. A.; Dzyabchenko, A.; van Eijck, B. P.; Elking, D. M.; van den Ende, J. A.; Facelli, J. C.; Ferraro, M. B.; Fusti-Molnar, L.; Gatsiou, C.-A.; Gee, T. S.; de Gelder, R.; Ghiringhelli, L. M.; Goto, H.; Grimme, S.; Guo, R.; Hofmann, D. W. M.; Hoja, J.; Hylton, R. K.; Iuzzolino, L.; Jankiewicz, W.; de Jong, D. T.; Kendrick, J.; de Klerk, N. J. J.; Ko, H.-Y.; Kuleshova, L. N.; Li, X.; Lohani, S.; Leusen, F. J. J.; Lund, A. M.; Lv, J.; Ma, Y.; Marom, N.; Masunov, A. E.; McCabe, P.; McMahon, D. P.; Meekes, H.; Metz, M. P.; Misquitta, A. J.; Mohamed, S.; Monserrat, B.; Needs, R. J.; Neumann, M. A.; Nyman, J.; Obata, S.; Oberhofer, H.; Oganov, A. R.; Orendt, A. M.; Pagola, G. I.; Pantelides, C. C.; Pickard, C. J.; Podeszwa, R.; Price, L. S.; Price, S. L.; Pulido, A.; Read, M. G.; Reuter, K.; Schneider, E.; Schober, C.; Shields, G. P.; Singh, P.; Sugden, I. J.; Szalewicz, K.; Taylor, C. R.; Tkatchenko, A.; Tuckerman, M. E.; Vacarro, F.; Vasileiadis, M.; Vazquez-Mayagoitia, A.; Vogt, L.; Wang, Y.; Watson, R. E.; de Wijs, G. A.; Yang, J.; Zhu, Q.; Groom, C. R. Report on the sixth blind test of organic crystal structure prediction methods. *Acta Cryst. B* **2016**, *72*, 439–459.

[7] Glass, C. W.; Oganov, A. R.; Hansen, N. USPEX—Evolutionary crystal structure prediction. *Comput. Phys. Commun.* **2006**, *175*, 713–720.

[8] Yamashita, T.; Kanehira, S.; Sato, N.; Kino, H.; Terayama, K.; Sawahata, H.; Sato, T.; Utsuno, F.; Tsuda, K.; Miyake, T.; Oguchi, T. CrySPY: a crystal structure prediction tool accelerated by machine learning. *Sci. Technol. Adv. Mater.* **2021**, *1*, 87–97.

[9] Curtis, F.; Li, X.; Rose, T.; Vazquez-Mayagoitia, A.; Bhattacharya, S.; Ghiringhelli, L. M.; Marom, N. GAtor: a first-principles genetic algorithm for molecular crystal structure prediction. *J. Chem. Theory Comput.* **2018**, *14*, 2246–2264.

[10] Oganov, A. R., Ed. *Modern methods of crystal structure prediction*; Wiley-VCH: Weinheim, Germany, 2011.

[11] Wang, Y.; Lv, J.; Gao, P.; Ma, Y. Crystal Structure Prediction via Efficient Sampling of the Potential Energy Surface. *Acc. Chem. Res.* **2022**, *55*, 2068–2076.

[12] L. Price, S. Is zeroth order crystal structure prediction (CSP_0) coming to maturity? What should we aim for in an ideal crystal structure prediction code? *Faraday Discuss.* **2018**, *211*, 9–30.

[13] Whittleton, S. R.; Otero-de-la-Roza, A.; Johnson, E. R. The exchange-hole dipole dispersion model for accurate energy ranking in molecular crystal structure prediction. *J. Chem. Theory Comput.* **2017**, *13*, 441–450.

[14] Lommerse, J. P. M.; Motherwell, W. D. S.; Ammon, H. L.; Dunitz, J. D.; Gavezzotti, A.; Hofmann, D. W. M.; Leusen, F. J. J.; Mooij, W. T. M.; Price, S. L.; Schweizer, B.; Schmidt, M. U.; Eijck, B. P. v.; Verwer, P.; Williams, D. E. A test of crystal structure prediction of small organic molecules. *Acta Cryst. B* **2000**, *56*, 697–714.

[15] Motherwell, W. D. S.; Ammon, H. L.; Dunitz, J. D.; Dzyabchenko, A.; Erk, P.; Gavezzotti, A.; Hofmann, D. W. M.; Leusen, F. J. J.; Lommerse, J. P. M.; Mooij, W. T. M.; Price, S. L.; Scheraga, H.; Schweizer, B.; Schmidt, M. U.; Eijck, B. P. v.; Verwer, P.; Williams, D. E. Crystal structure prediction of small organic molecules: a second blind test. *Acta Cryst. B* **2002**, *58*, 647–661.

[16] Day, G. M.; Motherwell, W. D. S.; Ammon, H. L.; Boerrigter, S. X. M.; Della Valle, R. G.; Venuti, E.; Dzyabchenko, A.; Dunitz, J. D.; Schweizer, B.; van Eijck, B. P.; Erk, P.; Facelli, J. C.; Bazterra, V. E.; Ferraro, M. B.; Hofmann, D. W. M.; Leusen, F. J. J.; Liang, C.; Pantelides, C. C.; Karamertzanis, P. G.; Price, S. L.; Lewis, T. C.; Nowell, H.; Torrisi, A.; Scheraga, H. A.; Arnautova, Y. A.; Schmidt, M. U.; Verwer, P. A third blind test of crystal structure prediction. *Acta Cryst. B* **2005**, *61*, 511–527.

[17] Day, G. M.; Cooper, T. G.; Cruz-Cabeza, A. J.; Hejczyk, K. E.; Ammon, H. L.; Boerrigter, S. X. M.; Tan, J. S.; Della Valle, R. G.; Venuti, E.; Jose, J.; Gadre, S. R.; Desiraju, G. R.; Thakur, T. S.; van Eijck, B. P.; Facelli, J. C.; Bazterra, V. E.; Ferraro, M. B.; Hofmann, D. W. M.; Neumann, M. A.; Leusen, F. J. J.; Kendrick, J.; Price, S. L.; Misquitta, A. J.; Karamertzanis, P. G.; Welch, G. W. A.; Scheraga, H. A.; Arnautova, Y. A.; Schmidt, M. U.; van de Streek, J.; Wolf, A. K.; Schweizer, B. Significant progress in predicting the crystal structures of small organic molecules – a report on the fourth blind test. *Acta Cryst. B* **2009**, *65*, 107–125.

[18] Bardwell, D. A.; Adjiman, C. S.; Arnautova, Y. A.; Bartashevich, E.; Boerrigter, S. X. M.; Braun, D. E.; Cruz-Cabeza, A. J.; Day, G. M.; Della Valle, R. G.; Desiraju, G. R.; van Eijck, B. P.; Facelli, J. C.; Ferraro, M. B.; Grillo, D.; Habgood, M.; Hofmann, D. W. M.; Hofmann, F.; Jose, K. V. J.; Karamertzanis, P. G.; Kazantsev, A. V.; Kendrick, J.; Kuleshova, L. N.; Leusen, F. J. J.; Maleev, A. V.; Misquitta, A. J.; Mohamed, S.; Needs, R. J.; Neumann, M. A.; Nikylov, D.; Orendt, A. M.; Pal, R.; Pantelides, C. C.; Pickard, C. J.; Price, L. S.; Price, S. L.; Scheraga, H. A.; van de Streek, J.; Thakur, T. S.; Tiwari, S.; Venuti, E.; Zhitkov, I. K. Towards crystal structure prediction of complex organic compounds – a report on the fifth blind test. *Acta Cryst. B* **2011**, *67*, 535–551.

[19] Oganov, A. R.; Glass, C. W. Crystal structure prediction using *ab initio* evolutionary techniques: Principles and applications. *J. Chem. Phys.* **2006**, *124*, 244704.

[20] Oganov, A. R.; Lyakhov, A. O.; Valle, M. How Evolutionary Crystal Structure Prediction Works – and Why. *Acc. Chem. Res.* **2011**, *44*, 227–237.

[21] Lyakhov, A. O.; Oganov, A. R.; Stokes, H. T.; Zhu, Q. New developments in evolutionary structure prediction algorithm USPEX. *Comput. Phys. Commun.* **2013**, *184*, 1172–1182.

[22] Zhu, Q.; Oganov, A. R.; Glass, C. W.; Stokes, H. T. Constrained evolutionary algorithm for structure prediction of molecular crystals: methodology and applications. *Acta Cryst. B* **2012**, *68*, 215–226.

[23] Weatherby, J. A.; Rumson, A. F.; Price, A. J. A.; Otero de la Roza, A.; Johnson, E. R. A density-functional benchmark of vibrational free-energy corrections for molecular crystal polymorphism. *J. Chem. Phys.* **2022**, *156*, 114108.

[24] Leach, A. R. *Molecular modelling: principles and applications*, 2nd ed.; Prentice Hall: Harlow, England ; New York, 2001.

[25] Coombes, D. S.; Price, S. L.; Willock, D. J.; Leslie, M. Role of Electrostatic Interactions in Determining the Crystal Structures of Polar Organic Molecules. A Distributed Multipole Study. *J. Phys. Chem.* **1996**, *100*, 7352–7360.

[26] Wang, J.; Wolf, R. M.; Caldwell, J. W.; Kollman, P. A.; Case, D. A. Development and testing of a general amber force field. *J. Comp. Chem.* **2004**, *25*, 1157–1174.

[27] Price, S. L.; Leslie, M.; Welch, G. W. A.; Habgood, M.; Price, L. S.; Karamertzanis, P. G.; Day, G. M. Modelling organic crystal structures using distributed multipole and polarizability-based model intermolecular potentials. *Phys. Chem. Chem. Phys.* **2010**, *12*, 8478–8490.

[28] Stone, A. J. *The theory of intermolecular forces*, 2nd ed.; Oxford University Press: Oxford, 2013.

[29] Rackers, J. A.; Wang, Z.; Lu, C.; Laury, M. L.; Lagardère, L.; Schnieders, M. J.; Piquemal, J.-P.; Ren, P.; Ponder, J. W. Tinker 8: Software Tools for Molecular Design. *J. Chem. Theory Comput.* **2018**, *14*, 5273–5289.

[30] Allinger, N. L.; Yuh, Y. H.; Lii, J. H. Molecular mechanics. The MM3 force field for hydrocarbons. 1. *J. Am. Chem. Soc.* **1989**, *111*, 8551–8566.

[31] Hill, T. L. Steric Effects. I. Van der Waals Potential Energy Curves. *J. Chem. Phys.* **1948**, *16*, 399–404.

[32] Williams, D. E. Representation of the molecular electrostatic potential by atomic multipole and bond dipole models. *J. Comp. Chem.* **1988**, *9*, 745–763.

[33] Allinger, N. L. Conformational analysis. 130. MM2. A hydrocarbon force field utilizing V1 and V2 torsional terms. *J. Am. Chem. Soc.* **1977**, *99*, 8127–8134.

[34] Halgren, T. A. Merck molecular force field. I. Basis, form, scope, parameterization, and performance of MMFF94. *J. Comp. Chem.* **1996**, *17*, 490–519.

[35] Halgren, T. A. The representation of van der Waals (vdW) interactions in molecular mechanics force fields: potential form, combination rules, and vdW parameters. *J. Am. Chem. Soc.* **1992**, *114*, 7827–7843.

[36] Jorgensen, W. L.; Maxwell, D. S.; Tirado-Rives, J. Development and Testing of the OPLS All-Atom Force Field on Conformational Energetics and Properties of Organic Liquids. *J. Am. Chem. Soc.* **1996**, *118*, 11225–11236.

[37] Brooks, B. R.; Bruccoleri, R. E.; Olafson, B. D.; States, D. J.; Swaminathan, S.; Karplus, M. CHARMM: A program for macromolecular energy, minimization, and dynamics calculations. *J. Comp. Chem.* **1983**, *4*, 187–217.

[38] D. Gale, J. GULP: A computer program for the symmetry-adapted simulation of solids. *J. Chem. Soc. Faraday Trans.* **1997**, *93*, 629–637.

[39] Gale, J. D.; Rohl, A. L. The General Utility Lattice Program (GULP). *Mol. Simul.* **2003**, *29*, 291–341.

[40] Mayo, S. L.; Olafson, B. D.; Goddard, W. A. DREIDING: a generic force field for molecular simulations. *J. Phys. Chem.* **1990**, *94*, 8897–8909.

[41] Williams, D. E. Improved intermolecular force field for molecules containing H, C, N, and O atoms, with application to nucleoside and peptide crystals. *J. Comp. Chem.* **2001**, *22*, 1154–1166.

[42] Gasteiger, J.; Marsili, M. A new model for calculating atomic charges in molecules. *Tetrahedron Lett.* **1978**, *19*, 3181–3184.

[43] Rappe, A. K.; Casewit, C. J.; Colwell, K. S.; Goddard, W. A.; Skiff, W. M. UFF, a full periodic table force field for molecular mechanics and molecular dynamics simulations. *J. Am. Chem. Soc.* **1992**, *114*, 10024–10035.

[44] Rappe, A. K.; Goddard, W. A. Charge equilibration for molecular dynamics simulations. *J. Phys. Chem.* **1991**, *95*, 3358–3363.

[45] Wang, J.; Wolf, R. M.; Caldwell, J. W.; Kollman, P. A.; Case, D. A. Development and testing of a general amber force field. *J. Comp. Chem.* **2004**, *25*, 1157–1174.

[46] Bayly, C. I.; Cieplak, P.; Cornell, W.; Kollman, P. A. A well-behaved electrostatic potential based method using charge restraints for deriving atomic charges: the RESP model. *J. Phys. Chem.* **1993**, *97*, 10269–10280.

[47] Frisch, M. J.; Trucks, G. W.; Schlegel, H. B.; Scuseria, G. E.; Robb, M. A.; Cheeseman, J. R.; Scalmani, G.; Barone, V.; Mennucci, B.; Petersson, G. A.; Nakatsuji, H.; Caricato, M.; Li, X.; Hratchian, H. P.; Izmaylov, A. F.; Bloino, J.; Zheng, G.; Sonnenberg, J. L.; Hada, M.; Ehara, M.; Toyota, K.; Fukuda, R.; Hasegawa, J.; Ishida, M.; Nakajima, T.; Honda, Y.; Kitao, O.; Nakai, H.; Vreven, T.; Montgomery, J. A., Jr.; Peralta, J. E.; Ogliaro, F.; Bearpark, M.; Heyd, J. J.; Brothers, E.; Kudin, K. N.; Staroverov, V. N.; Kobayashi, R.; Normand, J.; Raghavachari, K.; Rendell, A.; Burant, J. C.; Iyengar, S. S.; Tomasi, J.; Cossi, M.; Rega, N.; Millam, J. M.; Klene, M.; Knox, J. E.; Cross, J. B.; Bakken, V.; Adamo, C.; Jaramillo, J.; Gomperts, R.; Stratmann, R. E.; Yazyev, O.; Austin, A. J.; Cammi, R.; Pomelli, C.; Ochterski, J. W.; Martin, R. L.; Morokuma, K.; Zakrzewski, V. G.; Voth, G. A.; Salvador, P.; Dannenberg, J. J.; Dapprich, S.; Daniels, A. D.; Farkas, O.; Foresman, J. B.; Ortiz, J. V.; Cioslowski, J.; Fox, D. J. Gaussian 09 Revision E.1. Gaussian Inc. Wallingford CT 2013.

[48] Jakalian, A.; Bush, B. L.; Jack, D. B.; Bayly, C. I. Fast, efficient generation of high-quality atomic charges. AM1-BCC model: I. Method. *J. Comp. Chem.* **2000**, *21*, 132–146.

[49] Jakalian, A.; Jack, D. B.; Bayly, C. I. Fast, efficient generation of high-quality atomic charges. AM1-BCC model: II. Parameterization and validation. *J. Comp. Chem.* **2002**, *23*, 1623–1641.

[50] Otero-de-la-Roza, A.; Johnson, E. R. Non-Covalent Interactions and Thermochemistry using XDM-Corrected Hybrid and Range-Separated Hybrid Density Functionals. *J. Chem. Phys.* **2013**, *138*, 204109.

[51] Otero-de-la-Roza, A.; Johnson, E. R. Van der Waals interactions in solids using the exchange-hole dipole moment. *J. Chem. Phys.* **2012**, *136*, 174109.

[52] Becke, A. D. On the large-gradient behavior of the density functional exchange energy. *J. Chem. Phys.* **1986**, *85*, 7184.

[53] Perdew, J. P.; Burke, K.; Ernzerhof, M. Generalized gradient approximation made simple. *Phys. Rev. Lett.* **1996**, *77*, 3865.

[54] Becke, A. D. Density-functional thermochemistry. III. The role of exact exchange. *J. Chem. Phys.* **1993**, *98*, 5648–5652.

[55] Lee, C.; Yang, W.; Parr, R. G. Development of the Colle-Salvetti correlation-energy formula into a functional of the electron density. *Phys. Rev. B* **1988**, *37*, 785.

[56] Somov, N. V.; Chuprunov, E. V. On the Distribution of Molecular Crystals of Organic and Elementoorganic Compounds over Symmetry Space Groups. *Crystallogr. Rep.* **2021**, *66*, 361–366.

[57] Roy, S.; Goedecker, S.; Hellmann, V. Bell-Evans-Polanyi principle for molecular dynamics trajectories and its implications for global optimization. *Phys. Rev. E* **2008**, *77*, 056707.

[58] Oganov, A. R.; Valle, M. How to quantify energy landscapes of solids. *J. Chem. Phys.* **2009**, *130*, 104504.

[59] Valle, M.; Oganov, A. R. Crystal structures classifier for an evolutionary algorithm structure predictor. 2008 IEEE Symposium on Visual Analytics Science and Technology. 2008; pp 11–18.

[60] Otero-de-la-Roza, A.; Johnson, E. R.; Luaña, V. Critic2: A program for real-space analysis of quantum chemical interactions in solids. *Comput. Phys. Commun.* **2014**, *185*, 1007–1018.

[61] de Gelder, R.; Wehrens, R.; Hageman, J. A. A generalized expression for the similarity of spectra: application to powder diffraction pattern classification. *J. Comput. Chem.* **2001**, *22*, 273–289.

[62] Heit, Y. N.; Beran, G. J. O. How important is thermal expansion for predicting molecular crystal structures and thermochemistry at finite temperatures? *Acta Cryst. B* **2016**, *72*, 514–529.

[63] Mayo, R. A.; Johnson, E. R. Improved quantitative crystal-structure comparison using powder diffractograms *via* anisotropic volume correction. *CrystEngComm* **2021**, *23*, 7118–7131.

[64] Mayo, R. A.; Otero de la Roza, A.; Johnson, E. R. Development and assessment of an improved powder-diffraction-based method for molecular crystal structure similarity. *CrystEngComm* **2022**, *24*, 8326–8338.

[65] Chisholm, J. A.; Motherwell, S. *COMPACK* : a program for identifying crystal structure similarity using distances. *J. Appl. Crystallogr.* **2005**, *38*, 228–231.

[66] Sacchi, P.; Lusi, M.; Cruz-Cabeza, A. J.; Nauha, E.; Bernstein, J. Same or different – that is the question: identification of crystal forms from crystal structure data. *CrystEngComm* **2020**, *22*, 7170–7185.

[67] Brown, J. E.; Luo, W.; Isabelle, L. M.; Pankow, J. F. Candy Flavorings in Tobacco. *N. Engl. J. Med.* **2014**, *370*, 2250–2252.

[68] Desiraju, G. R. Crystal Engineering: From Molecule to Crystal. *J. Am. Chem. Soc.* **2013**, *135*, 9952–9967.

[69] Giannozzi, P.; Andreussi, O.; Brumme, T.; Bunau, O.; Nardelli, M. B.; Calandra, M.; Car, R.; Cavazzoni, C.; Ceresoli, D.; Cococcioni, M. Advanced capabilities for materials modelling with Quantum ESPRESSO. *J. Phys.: Condens. Mat.* **2017**, *29*, 465901.

[70] Blöchl, P. E. Projector augmented-wave method. *Phys. Rev. B* **1994**, *50*, 17953.

[71] Johnson, E. R. In *Non-covalent Interactions in Quantum Chemistry and Physics*; Otero-de-la-Roza, A., DiLabio, G. A., Eds.; Elsevier, 2017; Chapter 5, pp 169–194.

[72] Whittleton, S. R.; Otero-de-la-Roza, A.; Johnson, E. R. The exchange-hole dipole dispersion model for accurate energy ranking in molecular crystal structure prediction II: Non-planar molecules. *J. Chem. Theory Comput.* **2017**, *13*, 5332–5342.

[73] Dirksen, J.; Ring, T. Fundamentals of crystallization: kinetic effects on particle size distributions and morphology. *Chem. Eng. Sci.* **1991**, *46*, 2389–2427.

[74] White, M. A.; Kahwaji, S.; Freitas, V. L. S.; Siewert, R.; Weatherby, J. A.; Ribeiro da Silva, M. D. M. C.; Verevkin, S. P.; Johnson, E. R.; Zwanziger, J. W. The Relative Thermodynamic Stability of Diamond and Graphite. *Angewandte Chemie* **2021**, *133*, 1570–1573.

[75] Price, A. J.; Otero-de-la Roza, A.; Johnson, E. R. XDM-corrected hybrid DFT with numerical atomic orbitals predicts molecular crystal lattice energies with unprecedented accuracy. *Chem. Sci.* **2023**,

[76] Neumann, M. A. Tailor-Made Force Fields for Crystal-Structure Prediction. *J. Phys. Chem. B* **2008**, *112*, 9810–9829.

[77] Dodda, L. S.; Cabeza de Vaca, I.; Tirado-Rives, J.; Jorgensen, W. L. LigParGen web server: an automatic OPLS-AA parameter generator for organic ligands. *Nucleic Acids Res.* **2017**, *45*, W331–W336.

[78] Wang, J.; Wang, W.; Kollman, P. A.; Case, D. A. Automatic atom type and bond type perception in molecular mechanical calculations. *J. Mol. Graph. Model.* **2006**, *25*, 247–260.

[79] Macrae, C. F.; Sovago, I.; Cottrell, S. J.; Galek, P. T. A.; McCabe, P.; Pidcock, E.; Platings, M.; Shields, G. P.; Stevens, J. S.; Towler, M.; Wood, P. A. *Mercury 4.0*: from visualization to analysis, design and prediction. *J. Appl. Cryst.* **2020**, *53*, 226–235.

[80] Blum, V.; Gehrke, R.; Hanke, F.; Havu, P.; Havu, V.; Ren, X.; Reuter, K.; Scheffler, M. Ab initio molecular simulations with numeric atom-centered orbitals. *Comp. Phys. Comm.* **2009**, *180*, 2175–2196.

[81] Price, A. J. A.; Otero de la Roza, A.; Johnson, E. R. XDM-corrected hybrid DFT with numerical atomic orbitals predicts molecular crystal lattice energies with unprecedented accuracy. *Chem. Sci.* **2023**, *14*, 1252–1262.

[82] Politzer, P.; Lane, P.; Concha, M. C.; Ma, Y.; Murray, J. S. An overview of halogen bonding. *J. Mol. Model.* **2007**, *13*, 305–311.

[83] Kolář, M.; Hobza, P. On Extension of the Current Biomolecular Empirical Force Field for the Description of Halogen Bonds. *J. Chem. Theory Comput.* **2012**, *8*, 1325–1333.

[84] M. LeBlanc, L.; R. Johnson, E. Crystal-energy landscapes of active pharmaceutical ingredients using composite approaches. *CrystEngComm* **2019**, *21*, 5995–6009.

[85] Lewis-Atwell, T.; Townsend, P. A.; Grayson, M. N. Comparisons of different force fields in conformational analysis and searching of organic molecules: A review. *Tetrahedron* **2021**, *79*, 131865.

[86] Paton, R. S.; Goodman, J. M. Hydrogen Bonding and π-Stacking: How Reliable are Force Fields? A Critical Evaluation of Force Field Descriptions of Nonbonded Interactions. *J. Chem. Inf. Model.* **2009**, *49*, 944–955.

[87] Thompson, A. P.; Aktulga, H. M.; Berger, R.; Bolintineanu, D. S.; Brown, W. M.; Crozier, P. S.; in 't Veld, P. J.; Kohlmeyer, A.; Moore, S. G.; Nguyen, T. D.; Shan, R.; Stevens, M. J.; Tranchida, J.; Trott, C.; Plimpton, S. J. LAMMPS - a flexible simulation tool for particle-based materials modeling at the atomic, meso, and continuum scales. *Comp. Phys. Comm.* **2022**, *271*, 108171.

[88] Kühne, T. D.; Iannuzzi, M.; Del Ben, M.; Rybkin, V. V.; Seewald, P.; Stein, F.; Laino, T.; Khaliullin, R. Z.; Schütt, O.; Schiffmann, F.; Golze, D.; Wilhelm, J.; Chulkov, S.; Bani-Hashemian, M. H.; Weber, V.; Borštnik, U.; Taillefumier, M.; Jakobovits, A. S.; Lazzaro, A.; Pabst, H.; Müller, T.; Schade, R.; Guidon, M.; Andermatt, S.; Holmberg, N.; Schenter, G. K.; Hehn, A.; Bussy, A.; Belleflamme, F.; Tabacchi, G.; Glöß, A.; Lass, M.; Bethune, I.; Mundy, C. J.; Plessl, C.; Watkins, M.; VandeVondele, J.; Krack, M.; Hutter, J. CP2K: An electronic structure and molecular dynamics software package - Quickstep: Efficient and accurate electronic structure calculations. *J. Chem. Phys.* **2020**, *152*, 194103.