Carnegie Mellon University
MELLON COLLEGE OF SCIENCE

THESIS

SUBMITTED IN PARTIAL FULFILLMENT OF THE REQUIREMENTS
FOR THE DEGREE OF

DOCTOR OF PHILOSOPHY IN THE FIELD OF PHYSICS

TITLE: "Structure Prediction of Molecular Crystals from First Principles with the GAtor Genetic Algorithm Package"

PRESENTED BY: Farren Curtis

ACCEPTED BY THE DEPARTMENT OF PHYSICS

Noa Marom	1/4/18
NOA MAROM, CHAIR PROFESSOR	DATE
Scott Dodelson	1/8/18
SCOTT DODELSON, DEPT HEAD	DATE

APPROVED BY THE COLLEGE COUNCIL

Rebecca Doerge	1/11/18
REBECCA DOERGE, DEAN	DATE

Structure Prediction of Molecular Crystals from First Principles with the GAtor Genetic Algorithm Package

by

Farren Shawna Curtis

Submitted in partial fulfillment of the
requirements for the degree of
Doctor of Philosophy
at
Carnegie Mellon University
Department of Physics
Pittsburgh, Pennsylvania

Advised by Professor Noa Marom

January 4, 2018

Abstract

Molecular crystals are a versatile class of materials with applications ranging from pharmaceuticals to organic electronics. Because molecular crystals are bound by weak dispersion interactions they often crystallize in more than one solid form, a phenomenon known as polymorphism. Understanding polymorphism has become an increasingly important issue because different crystal forms may display vastly different physical properties, which affects their functionality for a given application. Crystal structure prediction (CSP), or the prediction of a molecule's putative crystal structures solely from its chemical composition, is a coveted computational tool as it can predict previously unobserved polymorphs and serve as complementary tool for experimental investigations. CSP is difficult in part because one needs to sample a large configuration space for even the simplest molecules. Furthermore, the differences between polymorphs can be even lower than 1 kJ/mol, making reliable CSP an extremely challenging task. In this thesis, I develop and apply a first principles genetic algorithm (GA) for CSP called GAtor, which finds the most stable crystal structures for small (semi-)rigid molecules solely from their chemical composition. State-of-the-art dispersion-inclusive density functional theory (DFT) is applied for the final ranking of putative crystal structures. A preliminary version of GAtor was used to participate in the Cambridge Crystallographic Data Centre's sixth blind test of organic CSP methods. The relative stabilities and electronic properties of potential polymorphs of tricyano-1,4-dithiino[c]-isothiazole generated therein are investigated in an additional study. The methodology of the production version of GAtor, and its corresponding initial pool generation package Genarris, are presented and applied to a chemically diverse set of four past blind test targets: 3,4-cyclobutylfuran, 5-cyano-3-hydroxythiophene, 1,3-dibromo-2-chloro-5-fluorobenzene, and tricyano-1,4-dithiino[c]-isothiazole. GAtor successfully predicts the experimental crystal structure(s) for all four targets, as well as other important low-energy structures. Notably, the lowest energy putative crystal structure for 5-cyano-3-hydroxythiophene has not been reported in any previous investigations of this molecule. This may motivate additional computational and experimental studies of this molecule.

Acknowledgments

I would like to thank Dr. Noa Marom for your mentorship throughout my PhD work and for being a positive female role model. I'd also like to thank the rest of my thesis committee: Dr. Mike Widom, Dr. Di Xiao, and Dr. Alan McGaughey. Thanks to John Paul Kilecdi-Li and Dr. Álvaro Vázquez-Mayagoitia for your important contributions to this work.

I would like to thank my parents for always encouraging my academic pursuits. Thanks to Roscoe Charlemagne for offering perspective and getting me outside. Last but not least, I'd like to thank Matt for your unconditional love and support throughout this entire process– this thesis is dedicated to you.

Contents

1 Introduction **1**

1.1 Molecular Crystal Polymorphism 1

1.2 Crystal Structure Prediction 3

1.2.1 Molecular Conformation Generation 4

1.2.2 Configuration Space Exploration 5

1.2.3 Final Energetic Stability Ranking 6

1.2.4 Cambridge Crystallographic Data Center (CCDC) Blind Tests 7

1.3 Overview 7

2 Methods **9**

2.1 Density Functional Theory 9

2.1.1 Density Functional Theory Formalism 9

2.1.2 The Local Density Approximation 11

2.1.3 Generalized Gradient Approximations 12

2.1.4 Hybrid Functionals 14

2.2 Dispersion Corrections 15

2.2.1 Functionals based on Non-Local Correlation 16

2.2.2 Semi-Local Meta-GGA Functionals 17

2.2.3 Pairwise Dispersion Approaches 17

2.2.4 The Tkatchenko-Scheffler Pairwise Dispersion Correction . . . 19

2.2.5 Many-body Dispersion Correction 20

3 Papers **24**

3.1 Published Paper: Report on the sixth blind test of organic crystal structure prediction methods 25

3.2 Published Paper: Effect of packing motifs on the energy ranking and electronic properties of putative crystal structures of tricyano-1, 4-dithiino[c]-isothiazole 55
3.3 Accepted Manuscript: Genarris: Random Generation of Molecular Crystal Structures and Fast Screening with a Harris Approximation . 66
3.4 Submitted Manuscript: GAtor: A First Principles Genetic Algorithm for Molecular Crystal Structure Prediction 94

4 Summary and Outlook **131**

A Appendix **136**
A.1 GAtor Genetic Algorithm User Manual 136

References **164**

Chapter 1

Introduction

1.1 Molecular Crystal Polymorphism

Molecular crystals are a unique class of materials with diverse applications in pharmaceuticals, organic electronics, pigments, and explosives [1–11]. The molecules comprising these crystals are bound by weak dispersion (van der Waals) interactions. These long-range, intermolecular interactions are much weaker than covalent bonds and arise from electrostatic interactions between instantaneous multipole moments generated by quantum mechanical fluctuations in the electron density. Due to the weak nature of dispersion interactions, many organic molecules may crystallize in more than one solid form, a phenomenon known as polymorphism. Because a molecular crystal's structure governs its physical properties, polymorphism may drastically affect the crystal's desired functionality for a given application.

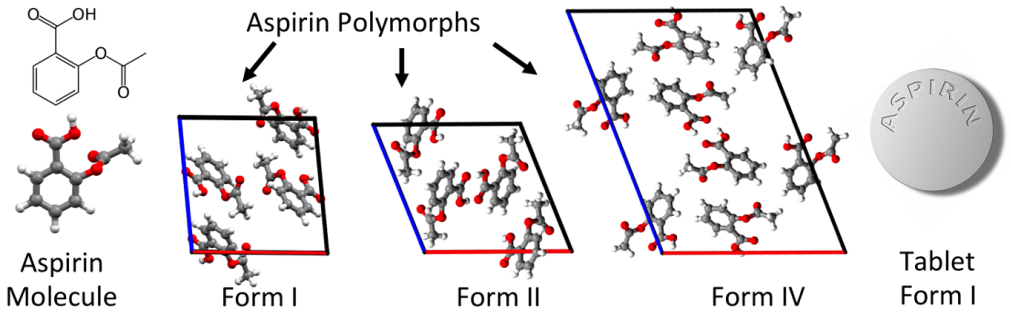


Figure 1.1: Different ambient polymorphs of the common pharmaceutical acetylsalicylic acid (Aspirin).

Understanding polymorphism is crucial for the pharmaceutical industry, since different polymorphs may display varying stability, solubility, and compressibility, affecting the drug's manufacturability, bioavailability, and efficacy [1, 12, 13]. Pharmaceutical companies must be able to predict and control which form is being produced in order to patent and distribute consistent products. To demonstrate the ubiquity of polymorphism in pharmaceuticals, consider one of the most commonly used drugs worldwide, Aspirin. Aspirin in fact has three ambient polymorphs, as shown in Fig. 1.1. Form I is the crystal structure that is produced for distribution. It was first discovered in 1853 [14] and was the only form known for over 100 years. Form II was suspected to exist since the 1960s, but was not synthesized until 2005 [15] following a crystal structure prediction study (introduced in Section 1.2) published in 2004 [16]. Form II was found to be considerably softer than Form I [17], an important mechanical property for tabletization. The third ambient polymorph, Form IV, was not reported until 2017 [18] and its structure was determined using a combination of X-ray diffraction and crystal structure prediction algorithms. If this form can be stored, it is predicted to have faster bioavailability than the other forms, which is the rate at which the drug reaches the circulatory system. Evidently, the possible polymorphs of this very common drug is still the subject of ongoing investigations 160 years after its discovery. Nowadays, when a pharmaceutical company develops a new drug, they perform extensive (and expensive) experimental solid form screening. This involves the production of numerous crystal forms for a given compound by varying experimental crystallization techniques and conditions. Different crystallization techniques include solvent evaporation, cooling of the solution (fast or slow), crystallization from the melt, heat or pressure induced transformations, vapor diffusion, etc. [19]. Once the different solid forms of a given compound have been produced, their physical properties are investigated in order to deduce the form that has the best properties for further development.

While polymorphism has long been a critical issue for the pharmaceutical community, it has recently received attention for its role in the development of organic electronic devices, including organic light-emitting diodes (OLEDs), organic field effect transistors (OFETs), and organic photovoltaic cells (OPVs). Organic electronics hold many advantages over their inorganic counterparts because they are cheap to manufacture, lightweight, made from earth-abundant elements, and flexible [20]. Importantly, their properties can be tuned for a given application by modifying their

chemical composition and crystal structure. Different organic semiconductor polymorphs may display markedly different band structures, optoelectronic properties, electronic couplings, and electron-phonon couplings that can drastically modify their performance in organic electronic devices [21–24]. For example, Rubrene is one of the most widely studied organic semiconductors due to its excellent charge transport properties. Single crystals of orthorhombic rubrene have demonstrated charge mobilities up to 20 $cm^2/(V{\cdot}s)$ [6, 25, 26], which far surpasses amorphous silicon. The monoclinic and triclinic forms have much lower mobilities [27], but may exhibit higher singlet-fission efficiencies [20], an important property for next-generation organic solar cell applications. Therefore, understanding polymorphism is crucial for the development of organic electronic devices.

1.2 Crystal Structure Prediction

One of the continuing scandals in the physical sciences is that it remains impossible to predict the structure of the simplest crystalline solids from a knowledge of their chemical composition. -Sir John Maddox (1988) [28]

Because molecular crystals have a wide range of applications, there has been increasing interest in the fundamental challenge of crystal structure prediction (CSP), or the computation of a molecule's putative crystal structure(s) solely from its 2D chemical diagram. Example 2D diagrams, which were used as the starting point for the CSP studies performed within this thesis, are shown in Fig. 1.2 and discussed in detail in Sections 3.3 and 3.4. CSP is a coveted computational tool as it can predict the existence of new polymorphs and serve as a complementary tool for experimental investigations [13, 29, 30]. Once considered impossible [28, 31], CSP is still an extremely challenging task as it requires combining highly accurate dispersion-inclusive electronic structure methods with efficient search algorithms that perform optimization over the complex, multidimensional potential energy surfaces of molecular crystals. The goal of a CSP study is not just to locate the most stable (global minimum) structure, but also any potential polymorphs that may be close in energy and thus potentially accessible experimentally. In general, a CSP study is divided into 3 parts: molecular conformation generation, configuration space exploration, and a final stability ranking of putative structures, briefly discussed below.

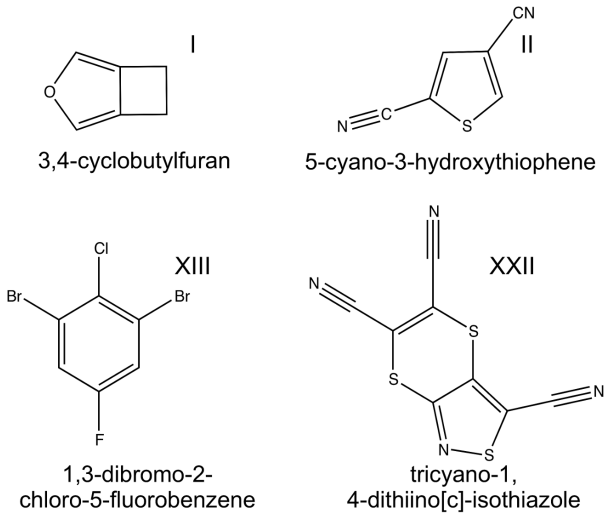


Figure 1.2: 2D diagrams of the small organic molecules investigated within this thesis.

1.2.1 Molecular Conformation Generation

The first step of a CSP study is to generate the molecule's stable gas phase conformers as a reasonable starting point for the conformation(s) the molecule will form in the solid state. For simple molecules with only a few degrees of freedom the low-energy conformers may be found by varying the molecular geometry along a given degree of freedom and then locating the local minima that correspond to the stable conformation(s) using molecular dynamics or first principles total energy evaluation methods. Such is the case for the (semi-)rigid molecule Target XXII, shown in Fig. 1.3, which has only one degree of freedom that allows for the molecule to bend along the S-S axis of the six-membered ring. The energy versus bending angle curve possesses two local minima corresponding to the two stable conformations that are mirror images of

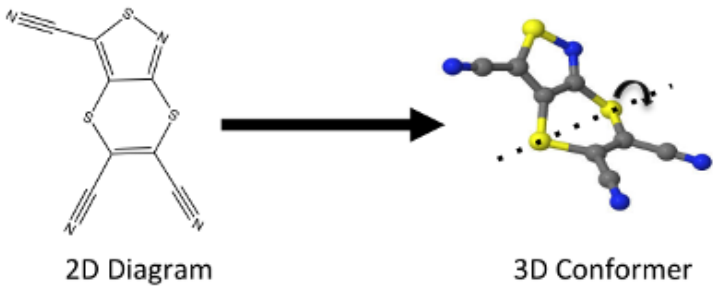


Figure 1.3: 2D chemical diagram and one 3D conformation of Target XXII.

one another, one of which is shown in Fig. 1.3. For the molecules investigated in this thesis, shown in Fig. 1.2, only Target XXII has partial flexibility, while the others are completely rigid. In the context of CSP a rigid molecule is defined as a molecule that has no rotational or torsional degrees of freedom. For more complex molecules, full conformation exploration studies may need to be performed using, e.g., molecular dynamics or Monte Carlo methods [32], which significantly adds to the overall complexity of the CSP study and is beyond the scope of the present thesis.

1.2.2 Configuration Space Exploration

Once the 3D conformer(s) have been generated for a given molecule, one needs to sample the crystalline configuration space by generating a large number of putative crystal structures and then evaluating their total energies. This task is highly nontrivial because the search space is enormous: it is not known *a priori* how many molecules will be in the unit cell, what space group the molecule will crystallize in, or what the lattice parameters will be. Furthermore, one needs to consider the accuracy and computational cost of the total energy method used since a large number of structures needs to be considered initially. Typically, a less intensive computational method is employed for sampling the configuration space, such as a tailor-made force field [33] or lower-level dispersion-inclusive density functional theory (DFT) (See Chapter 2). Classic "generalized" force fields [34, 35] cannot be used because they are not accurate enough for the purposes of reliable molecular crystal structure prediction [36].

Approaches to configuration space exploration in CSP include molecular dynamics [37, 38], Monte Carlo methods [29, 39], particle swarm optimization [40], basin-hopping [41], and (quasi)-random searches [42, 43]. In addition, genetic algorithms (GAs) are a versatile class of optimization algorithms inspired by the evolutionary principle of survival of the fittest [44–46]. GAs are suitable for molecular CSP because they can be applied robustly to complex multidimensional search spaces, including those with many extrema or discontinuous derivatives. They provide a good balance between exploration and exploitation by introducing randomness in the mating step followed by local optimization. Furthermore, they are conceptually simple algorithms, ideal for parallelization, and can lead to unbiased and unintuitive solutions. In the context of structure prediction, the target function being optimized is typically the total energy. A detailed description of the methodology and applications of the

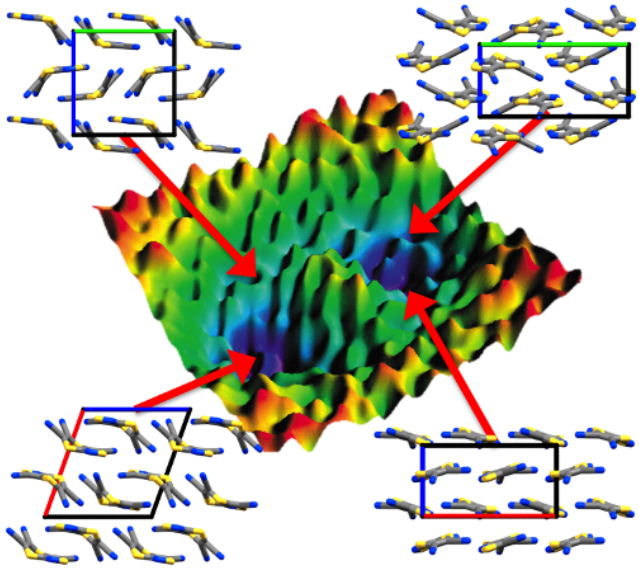


Figure 1.4: A schematic representation of the exploration of Target XXII's potential energy surface showing four putative crystal structures.

genetic algorithm GAtor, developed within this thesis, is provided in Section 3.4.

1.2.3 Final Energetic Stability Ranking

Once the configuration space for a given molecule has been sampled for a large number of structures with a given energy method, the next step involves reevaluating the energies of the best structures produced with more accurate and computationally demanding methods. This procedure is known as a hierarchal approach [36]. The final reranking step is crucial since the energy differences between molecular crystal polymorphs are typically within a few kJ/mol [47–50], which makes computing the relative stabilities of putative crystal structures particularly challenging. Reaching the required accuracy has become more practical thanks to modern computing power and

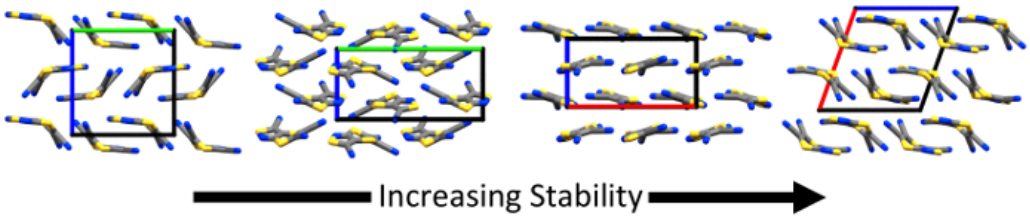


Figure 1.5: The final reranking step of a CSP study.

a decade of development in dispersion-inclusive DFT, including improved exchange-correlation functionals [51–62] and dispersion methods [63–77]. Chapter 2 provides more details on DFT and dispersion-inclusive DFT methods, with Sections 2.2.3 and 2.3.4 detailing the specific dispersion corrections used within this thesis.

1.2.4 Cambridge Crystallographic Data Center (CCDC) Blind Tests

Since 1999, the CCDC has held periodic CSP blind tests [36, 78–82] to assess the advances and remaining challenges of state-of-the-art methods in CSP. In these blind tests, 2D chemical diagrams of molecules with unpublished crystal structures are distributed to participating groups who then have a year to submit 100 putative crystal structures using a variety of CSP methods. There are several categories, ranging from small (semi-)rigid molecules to large flexible molecules. The blind tests have shown steady progress, but the accurate stability ranking of polymorphs remains a challenge. For example, in the sixth and most recent blind test, the experimental structures of many of the targets were ranked as the most stable structure by several methods, but no one method was able to rank all experimental structures as the most stable consistently. Our group participated in the small (semi-rigid) molecule category of the sixth blind test using a preliminary version of the GAtor genetic algorithm. Section 3.1 provides the main sixth blind test publication as well the supporting information detailing our individual participation.

1.3 Overview

In this thesis I develop and apply a massively parallel, first principles genetic algorithm for molecular crystal structure prediction, called GAtor. Chapter 2 introduces the main theory and concepts behind dispersion-inclusive DFT. Chapter 3 provides the full manuscripts of papers published (or submitted) chronologically throughout the course of this thesis. Specifically, Section 3.1 includes the main publication from the sixth blind test of organic CSP methods, followed by the supporting information of our individual submission using a preliminary version of GAtor. While our group was unable to predict the experimental crystal structure for Target XXII, we generated several other important low-energy structures. Section 3.2 includes a publication that

presents an analysis of the relative stabilities and electronic properties of putative crystal structures of Target XXII produced during the sixth blind test. We show that a potential polymorph, possessing a layered packing motif, exhibits markedly different electronic and optical properties from the experimental structure, including a narrower band gap, enhanced band dispersion and broader optical absorption. Section 3.3 includes an accepted manuscript that details the implementation and application of the Python package Genarris, which performs fast configuration space screening of molecular crystals using a combination of fragment-based DFT with unsupervised clustering methods from machine learning. Genarris is used to generate the initial pool structures for GAtor. Finally, Section 3.4 includes a submitted manuscript that details the methodology and application of the most recent version of GAtor, the main subject of this thesis. Herein, the experimentally observed crystal structures and other low-energy structures are generated for the four chemically-diverse targets shown in Fig. 1.2. For Target II, the top ranked putative crystal structure possesses a scaffold packing motif unlike the layered motif of the experimental and was predicted for the first time using GAtor. This may motivate further computational and experimental investigations of Target II. Chapter 4 provides a summary and outlook the future development of GAtor. The user manual for GAtor is included as an appendix.

Chapter 2

Methods

This chapter briefly describes the general formalism of density functional theory (DFT) as well as modern approaches for incorporating dispersion into common DFT strategies. Specifically, Sections 2.2.4 and 2.2.5 describe the Tkatchenko-Sheffler (TS) pairwise dispersion correction and the many-body dispersion correction (MBD) used for evaluating the dispersion energy of molecular crystals within this thesis. The methodology for the Genarris structure generation package and the GAtor genetic algorithm are provided in the manuscripts in Sections 3.3-3.4 and are not further included here.

2.1 Density Functional Theory

2.1.1 Density Functional Theory Formalism

The time-independent Shrödinger equation (1926) for a system of N interacting particles is given by

$$\left(\sum_{j=1}^{N} -\frac{1}{2}\nabla^2 + V(\mathbf{r}_1, \mathbf{r}_2, ..., \mathbf{r}_N)\right) \Psi(\mathbf{r}_1, \mathbf{r}_2, ..., \mathbf{r}_N) = E\Psi(\mathbf{r}_1, \mathbf{r}_2, ..., \mathbf{r}_N) \tag{2.1}$$

where $\mathbf{r}_j$ is the position of particle j, V is the potential of the system, and E is the total energy. Thus in principle, given the potential V for any system, one can solve for its total energy using Eq. 2.1. However, this equation depends on the position of every particle in the system, each additional particle adding four degrees of freedom, three

for position and one for spin. Therefore, Schrödinger's equation becomes intractable for any system consisting of more than a few electrons and cannot be solved directly for any realistic system of interest.

In 1965, Walter Kohn and Lu Jeu Sham developed a set of equations [83] that replaced the system of many *interacting* electrons moving in an external potential with an equivalent system consisting of *noninteracting* particles moving in an effective potential. The Kohn-Sham system has the same ground state density as the fully interacting system. In this formulation the ground state density $n(\mathbf{r})$ is represented as a sum of single particle orbitals, called Kohn-Sham orbitals $\psi_j(\mathbf{r})$, and is given by

$$n(\mathbf{r}) = \sum_{j=1}^{N} |\psi_j(\mathbf{r})|^2 \tag{2.2}$$

where each particle's spin has been absorbed into the index j. The self-consistent, single particle equations known as the Kohn-Sham (KS) equations are given by

$$\left(-\frac{1}{2}\nabla^2 + v_{ext}(\mathbf{r}) + v_H(\mathbf{r}) + v_{xc}(\mathbf{r})\right)\psi_j(\mathbf{r}) = \epsilon_j\psi_j(\mathbf{r}). \tag{2.3}$$

where ψ_j are the single-particle, or Kohn-Sham, orbitals with corresponding energy ϵ_j. Eq. 2.3 contains the external (ion-electron) potential $v_{ext}(\mathbf{r})$ as well as the classical electrostatic Hartree potential v_H given by

$$v_H(\mathbf{r}) = \int \frac{n(\mathbf{r}')}{|\mathbf{r} - \mathbf{r}'|} d\mathbf{r}'. \tag{2.4}$$

The last part of the effective potential, v_{xc}, is called the exchange-correlation potential. It is the functional derivative of what is known as the exchange-correlation energy E_{xc},

$$v_{xc} = \frac{\partial E_{xc}[n]}{\partial n(\mathbf{r})} \tag{2.5}$$

Exchange-correlation accounts for electron-electron interactions beyond those that would arise from an effectively classical charge distribution. Exchange incorporates the effects of the Pauli exclusion principle for identical fermions while correlation incorporates the non-classical Coulombic effects beyond exchange.

Eq. 2.3 is solved self-consistently. First, an initial guess for $\psi_j(\mathbf{r})$ is used to construct the ground state density $n(\mathbf{r})$ as given in Eq. 2.2 and then input into Eq.

2.3, which is solved for a new set of orbitals $\psi_j(\mathbf{r})'$. The process repeats until the final orbitals obtained are the same as the initial guess, or in other words the solution is self-consistent. The total ground state energy of the system is given by

$$E[n] = \sum_j \epsilon_j - \frac{1}{2}\int\int \frac{n(\mathbf{r})}{|\mathbf{r}-\mathbf{r}'|} d\mathbf{r}d\mathbf{r}' + E_{xc}[n] - \int v_{xc}(\mathbf{r})n(\mathbf{r})d\mathbf{r}. \tag{2.6}$$

Eq. 2.6 is exact provided the form of exchange-correlation energy functional $E_{xc}[n]$ is known. However, the exact form of exchange-correlation energy functional is not known and thus must be approximated.

2.1.2 The Local Density Approximation

The simplest approximation to the exchange-correlation energy functional is known as the local density approximation (LDA), suggested by Kohn and Sham [83]. LDA treats a nonuniform system, e.g. a molecule or solid, as if it is composed of infinitesimal volume elements, each of which consists of a uniform electron gas. The LDA exchange-correlation functional is given by

$$E_{xc}^{LDA}[n] = \int n(\mathbf{r})\epsilon_{xc}^{unif}(n(\mathbf{r}))d^3\mathbf{r} \tag{2.7}$$

where $\epsilon_{xc}^{unif}(n(\mathbf{r}))$ is the exchange-correlation energy per electron of a uniform electron gas of density $n(\mathbf{r})$ and can be partitioned into a sum of an exchange and a correlation contribution

$$\epsilon_{xc}^{unif}(n(\mathbf{r})) = \epsilon_x^{unif}(n(\mathbf{r})) + \epsilon_c^{unif}(n(\mathbf{r})). \tag{2.8}$$

The exchange energy of the uniform electron gas is known exactly and is given by the expression [83]:

$$\epsilon_x^{unif}(n(\mathbf{r})) = -e^2\left(\frac{3}{n}(\mathbf{r})/\pi\right)^{1/3}. \tag{2.9}$$

Analytical expressions for the correlation energy only exist for the high density [84] and low density [85, 86] limits. The intermediate range correlation energy has been calculated using Monte Carlo simulations [87]. LDA is parametrized by putting together the exchange-correlation energy for the various density regimes [87–90].

LDA was expected to work well for systems with slowly varying electron densities, relative to the local Fermi wavelength. Most realistic systems have nonuniform densi-

ties with rapidly changing densities in certain regions of space (e.g. the charge density of molecules). LDA turned out to be much more accurate than was expected from its initial approximations, and was not confined to systems with approximately uniform electron density, as in bulk metals and surfaces. The surprisingly good performance of LDA for many systems is attributed to systematic error cancellation between the exchange and correlation energies. The shortcomings of LDA can be found, for example, by looking at the atomization energies of molecules. In this case LDA overbinds, producing an absolute error of about 1 eV [91], which is significantly larger than the desired 0.05 eV chemical accuracy [92–94]. Additionally, LDA typically overestimates the bond strength in solids, which underestimates bond lengths and favors close-packed structures. LDA severely overbinds hydrogen-bonded systems [95]. To obtain better accuracy, the exchange-correlation functional needs to go beyond the simple dependency on the local electron density.

2.1.3 Generalized Gradient Approximations

Gradient approximations are based on the idea that the exchange-correlation functional may be improved by additionally including the gradient of the electron density, $\nabla n(\mathbf{r})$, and potentially higher-order derivatives to account for density variations in nonuniform systems. The Gradient Expansion Approximation (GEA) yields a functional of the form

$$E_{xc}^{GEA}[n] = \int n(\mathbf{r})\epsilon_{xc}(n(\mathbf{r}), |\nabla n(\mathbf{r})|)d^3\mathbf{r}. \tag{2.10}$$

The early implementations of GEA performed worse than LDA for most realistic materials. This is due to the large gradients produced near rapidly changing regions, which produces significant errors in the exchange-correlation functional. LDA is exact for a uniform charge density, but GEA is not based on a physical system and does not satisfy several physical constraints [96, 97].

In the generalized gradient approximation (GGA) the exchange-correlation functional is given by

$$E_{xc}^{GGA}[n] = \int n(\mathbf{r})\epsilon_{xc}^{unif}(n(\mathbf{r}))F_{xc}(n(\mathbf{r}), |\nabla n(\mathbf{r})|)d^3\mathbf{r}. \tag{2.11}$$

where F_{xc} is a dimensionless enhancement factor that controls the behavior of the functional when the gradient grows large. Specifically, the exchange and correlation

functionals are written as

$$E_x^{GGA}[n] = A_x \int n(\mathbf{r})^{4/3} F_x(s) d^3\mathbf{r}. \tag{2.12}$$

$$E_c^{GGA}[n] = \int n(\mathbf{r})[\epsilon_c^{unif}(n(\mathbf{r})) + F_c(t)] d^3\mathbf{r} \tag{2.13}$$

where A_x is a coefficient, and $F_x(s)$ and $F_c(t)$ are the exchange and correlation enhancement factors, respectively, which are functions of the reduced gradients s and t given by

$$s = \frac{|\nabla n|}{2k_F n} \tag{2.14}$$

$$t = \frac{|\nabla n|}{2k_s n} \tag{2.15}$$

where k_F is the Fermi wavevector and $1/k_s$ is the Thomas-Fermi screening length. GGA has several implementations that differ in their construction of the enhancement factors, which obey different physically limiting cases. Common GGA implementations include those of Perdew, Burke, and Ernzerhof (PBE) [98, 99]; Perdew and Wang [100]; the Becke exchange [101], and the Lee, Yang, and Parr (LYP) correlation [102]. The GGA functionals show improvement over LDA for predicting molecular atomization energies, producing an absolute error of molecular atomization energies on the order of 0.3 eV, but this is still an order of magnitude larger than the desired chemical accuracy. For solids, the PBE functional significantly improves LDA cohesive energies and lattice constants, but shows a tendency to underbind [103], the opposite trend of LDA. GGAs and LDA both underestimate band gaps. In general, the GGA functionals produce better results than LDA, with a similar computational cost. However, LDA and GGAs both fail qualitatively for systems containing highly localized charge densities (e.g. transition-metal oxides) due to the self-interaction error (SIE), the spurious interaction of an electron with itself. In order to correct the SIE, the self-repulsion part of the classsical Hartree term in Eq 2.6. must be exactly canceled out by E_{xc}. In the Hartree-Fock wavefunction model this is indeed the case, but in DFT the self-interaction error persists due to the various approximations to E_{xc}.

2.1.4 Hybrid Functionals

The idea behind a hybrid functional is to mix in a fraction of exact exchange energy (from Hartree Fock theory) with the exchange and correlation energy of a semi-local functional to mitigate the effects of the SIE and obtain more accurate results. A global hybrid functional takes the form:

$$E_{xc}^{hybrid} = \alpha E_x^{exact} + (1-\alpha)E_x^{GGA} + E_c^{GGA} \tag{2.16}$$

$$E_x^{exact} = -\frac{1}{2}\sum_{ij}^{N}\int\int \frac{\psi_i^*(\mathbf{r})\psi_j(\mathbf{r}')\psi_j^*(\mathbf{r})\psi_i(\mathbf{r}')}{|\mathbf{r}-\mathbf{r}'|}d^3\mathbf{r}d^3\mathbf{r}' \tag{2.17}$$

where α controls the fraction of exact exchange and $\psi_{i,j}$ are the Kohn-Sham orbitals. The parameter α is typically set between 0.2-0.3 and can be set semi-empirically by fitting to experimental data or derived from first principles. One popular semi-empirical hybrid functional used commonly in the chemistry community is the Becke-three-parameter-Lee-Yang-Parr (B3LYP) [104] functional, in which a three-parameter functional is used that combines Hartree Fock exchange, LDA exchange, LYP correlation [102], and Vosko-Wilk-Nusair (VWN) [88] LDA correlation given by

$$E_{xc}^{B3LYP} = (1-a_0)E_x^{LDA} + a_0E_x^{exact} + a_x\Delta E_x^{B88} + a_cE_c^{LYP} + (1-a_c)E_c^{VWN} \tag{2.18}$$

where ΔE_x^{B88} is the exchange gradient correction developed by Becke in 1988 [101] and $a_0 = 0.2$, $a_x = 0.72$, and $a_c = 0.81$. These parameters were fit to reproduce important quantities, such as atomization energies, of the molecular G2 test set of Pople [105]. B3LYP successfully describes many properties of molecules, but is not as suitable for extended systems as they were not part of the data set the functional was fitted to.

Another popular hybrid functional is the nonempirically-derived PBE0 [106, 107], which is based on the PBE GGA functional [98, 99]. In PBE0, the fraction of exact exchange, α from Eq. 2.15, is set to 0.25 such that the exchange-correlation functional is given by:

$$E_{xc}^{PBE0} = 0.25E_x^{exact} + 0.75E_x^{PBE} + E_c^{PBE}. \tag{2.19}$$

In PBE0 the value of $\alpha = 0.25$ is determined nonempirically using perturbation theory and the adiabatic connection theorem, which connects the fictitious noninteracting

Kohn-Sham reference system to the fully interacting system adiabatically, through a continuum of partially interacting systems all sharing a common density. As compared to PBE, PBE0 significantly improves the equilibrium lattice constants and the bulk moduli for solids [63, 64] and the atomization energies of small molecules [108]. Most importantly, the inclusion of a fraction of Fock exchange mitigates the SIE. However, PBE0 is much more computationally expensive than PBE ($O(N^4)$ scaling versus $O(N^3)$) due to the four orbital terms in the expression for exact exchange.

2.2 Dispersion Corrections

The structure of a molecular crystal is governed by dispersion (van der Waals) interactions. These interactions do not involve significant overlap of charge densities (as is the case for a chemical bond) but rather emerge from quantum fluctuations of the electron density that lead to the formation of instantaneous dipoles and higher order multipoles. The electrostatic interaction between these multipoles generates a weak but long-ranged attractive force, as depicted in Fig. 2.1. The inherently non-local

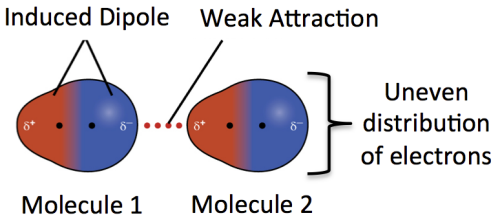


Figure 2.1: A cartoon depicting the weak dispersion interactions that arise between two molecules.

electronic correlation effects responsible for these long-ranged dispersion forces are not accounted for in standard semi-local approximations to the exchange-correlation functional, or in standard hybrid functionals based on semi-local correlation [109, 110]. Therefore, additional measures must be taken to properly account for dispersion.

There are many modern approaches for including dispersion in density functional theory. These can be broadly categorized into three main categories: (i) Non-local density-based functionals, (ii) semi-empirical one-electron approaches, and (iii) pairwise approaches. Each general approach will be discussed briefly in Sections 2.2.1-2.2.3. Particular emphasis will be placed on the Tkatchenko-Scheffler (TS) pairwise

dispersion correction in Section 2.2.4 as this is employed in Chapter 3 of this thesis. A recently developed Many-body Dispersion (MBD) correction (incorporating parts of the TS method) is detailed in Section 2.2.5 and is also used in Chapter 3.

2.2.1 Functionals based on Non-Local Correlation

In the spirit of traditional density functional theory, non-local density-based functionals share the feature that only the electron density is needed to compute the dispersion energy. In these approaches the *long-range* correlation energy is given by

$$E_c^{nl} = \frac{1}{2} \int \int n(\mathbf{r}) \Phi(\mathbf{r}, \mathbf{r}') n(\mathbf{r}') d^3\mathbf{r} d^3\mathbf{r}' \tag{2.20}$$

In Eq. 2.20 the correlation integral kernal Φ is a function of two electron coordinates (in contrast to semi-local functionals which depend locally on a single coordinate in space). Non-local functionals are seamless, in the sense that Eq. 2.20 is added to the rest of the functional without having to specify atomic fragments or atomic partitioning [111, 112]. Furthermore they are typically constructed with little no or empiricism, typically relying on the adiabatic connection theorem.

The most successful and commonly used representation of a non-local density-based approach for including dispersion is the vdW-DF family of functionals [51, 52, 113, 114]. In these functionals the correlation energy is given by

$$E_c^{vdW\text{-}DF} = E_c^{sr} + E_c^{nl} \tag{2.21}$$

where the short-range (sr) local part E_c^{sr} is given by LDA, and the long-range part is given by Eq. 2.20. The exchange energy is based on a GGA (revPBE [115] for vdW-DF and rPW86 [52, 116] for vdW-DF2). The main field of application for the vdW-DF functional is systems with extensive electron–electron delocalization such as, e.g., physisorption and metal surfaces. Although vdW-DF is a successful approach for many solid state applications, it underestimates hydrogen bond strengths and overestimates molecular separations [52, 117]. The improved functional, vdW-DF2, better captures, e.g., the lattice energies and geometries of molecular crystals [67].

2.2.2 Semi-Local Meta-GGA Functionals

Although dispersion arises due to the correlated movement between electrons, it may be described to a certain extent using effective one-electron approaches. These semilocal density functionals are heavily parametrized and capable of describing dispersion interactions in the intermediate range. A popular approach is the Minnesota (M0) family of functionals [54–56, 118–124]. These functionals are a class of meta-GGA functionals, meaning they incorporate the second-derivative-based kinetic energy term of the electron density. Arguably, this extra term is needed to allow sufficient flexibility in the functional form to capture additional effects that cannot be captured by a GGA. A functional form is adopted in which the correlation kernel is expressed as a finite expansion of the local density, which decays exponentially [125]. These functionals contain a large amount of parameters (≈20-50), some of which are fixed by applying physical constraints, but most of which are set by fitting to high-level reference data sets describing systems with various levels of noncovalent interactions [111, 112]. The Minnesota functionals perform well for thermochemistry and kinetics on large data sets [126], but their accuracy at describing, e.g. molecular crystals, is average at best [112]. This can be attributed to their inability at capturing the long-range ($1/R^6$) component of dispersion interactions. Furthermore, they come at a large computational cost, as compared to the pairwise methods described in the following section, but still lower than functionals with non-local correlation.

2.2.3 Pairwise Dispersion Approaches

In this set of approaches the dispersion energy is simply added to the electronic energy of the base DFT functional:

$$E_{tot} = E_{DFT} + E_{disp}. \tag{2.22}$$

For two neutral atoms (A and B) separated by a large distance R, the dispersion energy is always attractive and given approximately by

$$E_{disp}^{AB} \approx -\frac{3}{2}\frac{I_A I_B}{I_A + I_B}\frac{\alpha_A^0 \alpha_B^0}{R^6} = -\frac{C_{6,approx}^{AB}}{R^6} \tag{2.23}$$

where α and I are each atoms static dipole polarizability and atomic ionization potential, respectively. These constants are combined into the pairwise C_6 coefficient, which determines the strength of the attractive interaction for the two atoms. This well-known equation is known as London's formula and can be derived quantum mechanically using second-order perturbation theory [127]. This simple equation represents the basic concept behind a popular class of dispersion corrections that compute the C_6 coefficients and higher order dispersion coefficients, given generally by

$$E_{disp} = -\sum_{n=6,8,10,\ldots}\sum_{A>B} f_n(R_{AB})\frac{C_n^{AB}}{R_{AB}^n} \tag{2.24}$$

where f_n are damping functions that are required in order to avoid the divergence of the $1/R^6$ term in the short-range and to avoid overbinding in typical covalent-bonding regimes. Currently, the most widely used pairwise approaches include the exchange-dipole moment (XDM) method by Becke and Johnson [66, 128, 129], the D1 [130], D2 [64], and D3 [65, 131] approaches by Grimme *et al.*, and the Tkatchenko-Scheffler [73] method (described in Section 2.2.4). In the density-dependent XDM model, the dispersion energy is derived from the interaction of the real-space electrostatic distributions generated by the electrons and their associated exchange holes. This model computes the C_6, C_8, and C_{10} coefficients strictly from first principles without empirical parameters. The DFT+D methods are extensively parametrized, and sacrifice strong adherence to physical principles in exchange for flexibility and simplicity in the implementations [111]. The D1 approach is not as commonly used anymore, due to the small data set of dimers to which it was fitted [112]. The D2 method is still commonly used, where the C_6 parameters are determined using *in vacuo* atomic static polarizabilities. The disadvantages to the D2 approach include the fact that the effect of the local chemical environment is not accounted for in the construction of the C_6 parameters and the fact that it lacks higher-order dispersion coefficients. The most recent D3 approach, takes the local chemical environment into account by the empirical concept of fractional coordination numbers (CNs), in which atoms in different bonding/hybridization situations have different CNs. DFT-D3 is extensively based on pre-computed quantities using time-dependent density functional theory (TDDDFT) and *ad hoc* methods like the CN in order to determine, e.g. the C_6 and C_8 coefficients, which it makes use of. Advantages of the D3 method include its accuracy and

computational efficiency.

2.2.4 The Tkatchenko-Scheffler Pairwise Dispersion Correction

The TS method [73] is a pairwise dispersion correction which computes the C_6 two-body dispersion energy term from Eq. 2.23. The TS method is novel in the sense that the C_6 parameters are determined from the electron density and thus change dynamically to account for the local chemical environment. The C_6 parameters are computed according to the formula:

$$C_6^{AB} = \frac{2C_6^{AA}C_6^{BB}}{(\alpha_B^0/\alpha_A^0)C_6^{AA} + (\alpha_A^0/\alpha_B^0)C_6^{BB}} \tag{2.25}$$

where α_A^0 is the static polarizability of atom A in its chemical environment and C_6^{AA} is its respective homoatomic coefficient. These are computed from the following relationships

$$C_6^{AA} = \left(\frac{V_A^{eff}}{V_A^{free}}\right)^2 C_{6,free}^{AA} \tag{2.26}$$

$$\alpha_A^0 = \left(\frac{V_A^{eff}}{V_A^{free}}\right) \alpha_{A,free}^0 \tag{2.27}$$

where the free values $\alpha_{A,free}^0$ and $C_{6,free}^{AA}$ that are taken from the database of Chu and Dalgarno [132]. The ratio between the effective volume of an atom in its local environment and the free atom volume, is given by

$$\frac{V_A^{eff}}{V_A^{free}} = \frac{\int w_A(\mathbf{r}) n(\mathbf{r}) r^3 d^3\mathbf{r}}{\int n_A^{free}(\mathbf{r}) r^3 d^3\mathbf{r}} \tag{2.28}$$

where r is the distance from the nucleus of atom A, $n(\mathbf{r})$ is the computed electron density in its local environment, $n_A^{free}(\mathbf{r})$ is the computed electron density of the free atom, and $w_A(\mathbf{r})$ is the Hirshfeld atomic partitioning [133, 134] weight for atom A given by

$$w_A(\mathbf{r}) = \frac{n_A^{free}(\mathbf{r})}{\sum_B n_B^{free}(\mathbf{r})} \tag{2.29}$$

As a result, changes in the polarizability of an atom A due to the local environment are estimated by changes in its atomic volume and different hybridization states are taken into account automatically.

In the TS scheme, the damping function from Eq. 2.23 is given by the Fermi-type function

$$f_{damp}(R_{AB}, R^0_{AB}) = \frac{1}{1 + \exp\left(-d\left(\frac{R_{AB}}{s_R R^0_{AB}} - 1\right)\right)} \tag{2.30}$$

where R_{AB} is the interatomic distance for atoms A and B and R^0_{AB} is the sum of the equilibrium vdW radii for the pair, given by

$$R^0_{AB} = R^0_A + R^0_B = \left(\frac{V^{eff}_A}{V^{free}_A}\right)^{1/3} R^{free}_A + \left(\frac{V^{eff}_B}{V^{free}_B}\right)^{1/3} R^{free}_B \tag{2.31}$$

where the free vdW radii R^{free}_A are taken from the literature [135]. The parameter d was set to 20 after fitting to the S22 database of Jurečka *et al* [136], which contains binding energies of 22 different weakly bound systems. The s_R parameter determines the onset of the dispersion correction for a particular exchange-correlation functional and is set to 0.94 for PBE and 0.96 for PBE0 [73].

2.2.5 Many-body Dispersion Correction

The pairwise approaches reviewed in Sections 2.2.3-2.2.4 represent efficient solutions for including the missing dispersion interactions in the context of semilocal DFT. However, pairwise methods are unable to capture the inherently many-body nature of dispersion interactions. The inclusion of beyond-pairwise-additive many-body dispersion (MBD) plays a key role in, e.g., achieving chemical accuracy (1 kcal/mol) for the interaction energies of molecular crystals [112] and accurately computing the relative energies and structures of certain molecular crystal polymorphs [47, 137]. In this section, an overview of the MBD method by Tkatchenko *et al* [75–77] will be presented in three main steps.

Representing the Atoms as Quantum Harmonic Oscillators

In the MBD formalism, the atoms comprising a molecular system are represented as a collection of spherical quantum harmonic oscillators (QHOs). Each QHO is

characterized by an effective, frequency-dependent dipole polarizability given by

$$\alpha_A(i\omega) = \frac{\alpha_A^0}{1 + (\omega/\omega_A)^2} \tag{2.32}$$

where α_A^0 is each atoms static polarizability (from Eq. 2.26 in the TS method) and ω_A is each atom's characteristic (resonant) frequency defined by

$$\omega_A = \frac{4}{3}\frac{C_6^{AA}}{(\alpha_A^0)^2}. \tag{2.33}$$

Computing Each Atom's Screened Dynamic Polarizability

In the next step, the screened polarizabilities for each atom are calculated. This is necessary because the dynamic response of each atom is influenced by the other atoms in the system. The screened polarizabilities $\alpha_A^{SCS}(i\omega)$ are obtained solving the electrostatic self-consistent screening (SCS) equation given by

$$\alpha_A^{SCS}(i\omega) = \alpha_A(i\omega) + \alpha_A(i\omega)\sum_{A\neq B}^{N} \mathcal{T}_{AB}^{SR}\alpha_B^{SCS}(i\omega) \tag{2.34}$$

where $\mathcal{T}_{AB}^{SR}$ is the short-range dipole-dipole interaction tensor. The short-range aspect comes from the fact that the Coulomb interaction is split into a short-range and a long-range contribution to avoid double counting of the short-range correlation energy coming from DFT. Formally, $\mathcal{T}_{AB}^{SR}$ is given by

$$\mathcal{T}_{AB,lm}^{SR} = (1 - f(r_{AB}))\mathcal{T}_{AB,lm}^{GG} \tag{2.35}$$

where $\mathcal{T}_{AB,lm}^{GG}$ is defined as

$$\mathcal{T}_{AB,lm}^{GG} = \partial_{r_A^l}\partial_{r_B^m} v_{GG}(r_{AB}). \tag{2.36}$$

In Eq. 2.36 r_A^l is the lth Cartesian component of $\mathbf{r}_i$ and v_{GG} is the Coulomb potential due to a spherical Gaussian charge distribution given by

$$v_{GG}(r_{AB}) = \frac{\mathrm{erf}(r_{AB}/\sigma_{AB})}{r_{AB}}. \tag{2.37}$$

In Eq. 2.35, $f(r_{AB})$ is given by the Fermi-type damping function given by Eq. 2.30.

The resonant frequency of each atom due to screening is given by

$$\omega_A^{SCS} = \frac{4}{3}\frac{C_6^{SCS}}{|\alpha_A^{SCS}|^2} \tag{2.38}$$

where the screened C_6 coefficients are computed from the Casimir-Polder equation

$$C_{6,AB}^{SCS} = \frac{3}{\pi}\int_0^\infty \alpha_A^{SCS}(i\omega)\alpha_B^{SCS}(i\omega)d\omega. \tag{2.39}$$

Computing the MBD Dispersion Energy

Next, the many-body dispersion energy is computed using the coupled fluctuating dipole model (CFDM) [138, 139] for a collection of coupled isotropic three-dimensional QHOs representing the atoms of the system. The CFDM Hamiltonian can be written as

$$\mathcal{H}_{CFDM} = -\sum_A^N \frac{\nabla_{\boldsymbol{\chi}}^2}{2} + \sum_A^N \frac{1}{2}\omega_A^2\boldsymbol{\chi}_A^2 + \sum_{A>B}^N \omega_A\omega_B\sqrt{\alpha_A\alpha_B}\boldsymbol{\chi}_A\mathcal{T}_{AB}^{LR}\boldsymbol{\chi}_B \tag{2.40}$$

where $\boldsymbol{\chi}_A = \sqrt{m_A}\mu_A$, with μ_A being the displacement of oscillator A from equilibrium. In Eq. 2.40 all polarizabilities and resonant frequencies are the self-consistent (SCS) ones computed in Eqs. 2.34 and 2.38, with the superscripts dropped for simplicity. The first two terms in Eq. 2.40 represent the kinetic and potential energy for an individual QHO, while the last term corresponds to the long-range dipole-dipole interaction between the coupled QHOs. The long-range dipole-dipole tensor is given by

$$\mathcal{T}_{AB,lm}^{LR} = f(r_{AB})\frac{-3r_{AB}^l r_{AB}^m + r_{AB}^2\delta_{lm}}{r_{AB}^5} \tag{2.41}$$

where r_{AB}^l and r_{AB}^m denote the l and m Cartesian components of $\mathbf{r}_{AB}$. Diagonalizing the $3N$ X $3N$ dipole-dipole interaction matrix from Eq. 2.40 leads to the following expression for the MBD dispersion energy:

$$E_{disp}^{MBD} = \frac{1}{2}\sum_{p=1}^{3N}\sqrt{\lambda_p} - \frac{3}{2}\sum_A^N \omega_A^{SCS} \tag{2.42}$$

where λ_p is the pth eigenvalue of the coupling matrix. Eq. 2.42 is the difference between the zero-point energies of the coupled and uncoupled QHOs. For a system with N atoms the many-body dispersion energy will contain terms up to the Nth

order, i.e. 2-body up to N-body contributions. Eq. 2.42 can also be derived from the adiabatic connection theorem [140, 141].

Chapter 3

Papers

In this chapter, the full length documents of two published papers and two manuscripts are provided. Section 3.1 provides the main publication from the sixth blind test of organic CSP methods, followed by the supporting information for our individual submission using a preliminary version of GAtor. Section 3.2 details an analysis of the energetic ranking and electronic properties of putative crystal structures of Target XXII produced during the sixth blind test. Section 3.3 includes an accepted manuscript that details the Python package Genarris, which can be used for fast configuration space screening of molecular crystals by employing a Harris approximation and clustering techniques from machine learning. Genarris is used to create a diverse set of initial pool structures for GAtor. Finally, Section 3.4 includes a submitted manuscript that details the latest development and applications of the Python package GAtor, the main subject of this thesis.

3.1 Published Paper: Report on the sixth blind test of organic crystal structure prediction methods

The main publication from the sixth blind test of organic CSP is provided followed by the supporting information discussing our group's individual approach. Our group participated for the first time in the small, rigid category (Target XXII). My contributions to this work include developing and applying the preliminary version GAtor used for CSP and being the team leader of our group's submission. This involved writing core GA modules for molecular crystal manipulation including crossover, mutation, selection, duplicate checks, and parallel workflow management on HPC systems. I led a team of graduate and undergraduate students as well as international collaborators from the initial stages of our CSP strategy to the final stages of structure reranking and preparation of the results. I also authored our groups supplementary information document, which is included after the main text.

STRUCTURAL SCIENCE CRYSTAL ENGINEERING MATERIALS

ISSN 2052-5206

Report on the sixth blind test of organic crystal structure prediction methods

Anthony M. Reilly,[a]* Richard I. Cooper,[b] Claire S. Adjiman,[c] Saswata Bhattacharya,[d] A. Daniel Boese,[e] Jan Gerit Brandenburg,[f]‡ Peter J. Bygrave,[g] Rita Bylsma,[h] Josh E. Campbell,[g] Roberto Car,[i] David H. Case,[g] Renu Chadha,[j] Jason C. Cole,[a] Katherine Cosburn,[k,l] Herma M. Cuppen,[h] Farren Curtis,[k,m] Graeme M. Day,[g] Robert A. DiStasio Jr,[i,n] Alexander Dzyabchenko,[o] Bouke P. van Eijck,[p]§ Dennis M. Elking,[q] Joost A. van den Ende,[h] Julio C. Facelli,[r,s] Marta B. Ferraro,[t] Laszlo Fusti-Molnar,[q] Christina-Anna Gatsiou,[c] Thomas S. Gee,[g] René de Gelder,[h] Luca M. Ghiringhelli,[d] Hitoshi Goto,[u,v] Stefan Grimme,[f] Rui Guo,[w] Detlef W. M. Hofmann,[x,y] Johannes Hoja,[d] Rebecca K. Hylton,[w] Luca Iuzzolino,[w] Wojciech Jankiewicz,[z] Daniël T. de Jong,[h] John Kendrick,[aa] Niek J. J. de Klerk,[h] Hsin-Yu Ko,[i] Liudmila N. Kuleshova,[y] Xiayue Li,[k,bb] Sanjaya Lohani,[k] Frank J. J. Leusen,[aa] Albert M. Lund,[q,cc] Jian Lv,[dd] Yanming Ma,[dd] Noa Marom,[k,ee] Artëm E. Masunov,[ff,gg,hh,ii] Patrick McCabe,[a] David P. McMahon,[g] Hugo Meekes,[h] Michael P. Metz,[jj] Alston J. Misquitta,[kk] Sharmarke Mohamed,[ll] Bartomeu Monserrat,[mm,nn] Richard J. Needs,[mm] Marcus A. Neumann,[oo] Jonas Nyman,[g] Shigeaki Obata,[u] Harald Oberhofer,[pp] Artem R. Oganov,[qq,rr,ss,tt] Anita M. Orendt,[r] Gabriel I. Pagola,[t] Constantinos C. Pantelides,[c] Chris J. Pickard,[uu,vv] Rafal Podeszwa,[z] Louise S. Price,[w] Sarah L. Price,[w] Angeles Pulido,[g] Murray G. Read,[a] Karsten Reuter,[pp] Elia Schneider,[ww] Christoph Schober,[pp] Gregory P. Shields,[a] Pawanpreet Singh,[j] Isaac J. Sugden,[c] Krzysztof Szalewicz,[jj] Christopher R. Taylor,[g] Alexandre Tkatchenko,[d,xx] Mark E. Tuckerman,[ww,yy,zz] Francesca Vacarro,[k,aaa] Manolis Vasileiadis,[c] Alvaro Vazquez-Mayagoitia,[bb] Leslie Vogt,[ww] Yanchao Wang,[dd] Rona E. Watson,[w] Gilles A. de Wijs,[h] Jack Yang,[g] Qiang Zhu[qq] and Colin R. Groom[a]

[a]The Cambridge Crystallographic Data Centre, 12 Union Road, Cambridge CB2 1EZ, England, [b]Chemical Crystallography, Chemistry Research Laboratory, Mansfield Road, Oxford OX1 3TA, England, [c]Department of Chemical Engineering, Centre for Process Systems Engineering, Imperial College London, London SW7 2AZ, England, [d]Fritz-Haber-Institut der Max-Planck-Gesellschaft, Faradayweg 4-6, 14195, Berlin, Germany, [e]Department of Chemistry, Institute of Physical and Theoretical Chemistry, University of Graz, Heinrichstraße 28/IV, 8010 Graz, Austria, [f]Mulliken Center for Theoretical Chemistry, Institut für Physikalische und Theoretische Chemie, Rheinische Friedrich-Wilhelms Universität Bonn, Beringstraße 4, 53115 Bonn, Germany, [g]School of Chemistry, University of Southampton, Southampton SO17 1BJ, England, [h]Radboud University, Institute for Molecules and Materials, Heyendaalseweg 135, 6525 AJ Nijmegen, The Netherlands, [i]Department of Chemistry, Princeton University, Princeton, NJ 08544, USA, [j]University Institute of Pharmaceutical Sciences, Panjab University, Chandigarh, India, [k]Department of Physics and Engineering Physics, Tulane University, New Orleans, LA 70118, USA, [l]Department of Physics, University of Toronto, Toronto, Canada M5S 1A7, [m]Department of Physics, Carnegie Mellon University, Pittsburgh, PA 15213, USA, [n]Department of Chemistry and Chemical Biology, Cornell University, Ithaca, NY 14853, USA, [o]Karpov Institute of Physical Chemistry, Moscow, Russia, [p]Utrecht University, The Netherlands, [q]OpenEye Scientific Software, 9 Bisbee Court, Suite D, Santa Fe, NM 87508, USA, [r]Center for High Performance Computing, University of Utah, 155 South 1452 East Room 405, Salt Lake City, UT 84112-0190, USA, [s]Department of Biomedical Informatics, University of Utah, 155 South 1452 East Room 405, Salt Lake City, UT 84112-0190, USA, [t]Departamento de Física and Ifiba (CONICET) Facultad de Ciencias Exactas y Naturales, Universidad de Buenos Aires, Ciudad Universitaria, Pab. I (1428), Buenos Aires, Argentina, [u]Educational Programs on Advanced Simulation Engineering, Toyohashi University of Technology, 1-1 Hibarigaoka, Tempaku-cho, Toyohashi, Aichi 441-8580, Japan, [v]Department of Computer Science and Engineering, Graduate School of Engineering, Toyohashi University of Technology, 1-1 Hibarigaoka, Tempaku-cho, Toyohashi, Aichi 441-8580, Japan, [w]Department of Chemistry, University College London, 20 Gordon Street, London WC1H 0AJ, England, [x]CRS4, Parco Scientifico e Tecnologico, POLARIS, Edificio 1, 09010 PULA, Italy, [y]FlexCryst, Schleifweg 23, 91080 Uttenreuth, Germany, [z]Institute of Chemistry, University of Silesia, Szkolna 9, 40-006 Katowice, Poland, [aa]Faculty of Life Sciences, University of Bradford, Richmond Road, Bradford BD7 1DP, England, [bb]Argonne Leadership Computing Facility, Argonne National Laboratory, Lemont, IL 60439, USA, [cc]Department of Chemistry, University of Utah, 155 South 1452 East Room 405, Salt Lake City, UT 84112-0190, USA, [dd]State Key Laboratory of Superhard Materials, Jilin University, Changchun 130012, People's Republic of China, [ee]Department of Materials Science and Engineering and Department of Physics, Carnegie Mellon University, Pittsburgh, PA 15213, USA, [ff]NanoScience Technology Center, University of Central Florida, 12424 Research Parkway, PAV400, Orlando, FL 32826, USA, [gg]Department of Chemistry, University of Central Florida, 4111 Libra Drive

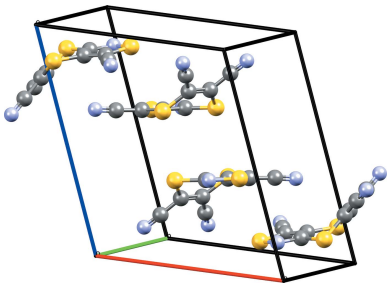
Received 15 February 2016
Accepted 4 May 2016

Edited by C. H. Görbitz, University of Oslo, Norway

‡ Present Address: Department of Chemistry, London Centre for Nanotechnology, University College London, 20 Gordon Street, London WC1H 0AJ, England.
§ Retired.

Keywords: crystal structure prediction; polymorphism; lattice energies; Cambridge Structural Database

Supporting information: this article has supporting information at journals.iucr.org/b



PSB225, Orlando, FL 32816, USA, [hh]Department of Physics, University of Central Florida, 4111 Libra Drive PSB430, Orlando, FL 32816, USA, [ii]Department of Condensed Matter Physics, National Research Nuclear University MEPhI, Kashirskoye shosse 31, Moscow 115409, Russia, [jj]Department of Physics and Astronomy, University of Delaware, Newark, DE 19716, USA, [kk]School of Physics and Astronomy, Queen Mary University of London, London E1 4NS, England, [ll]Khalifa University, PO Box 127788, Abu Dhabi, United Arab Emirates, [mm]Cavendish Laboratory, 19, J. J. Thomson Avenue, Cambridge CB3 0HE, England, [nn]Department of Physics and Astronomy, Rutgers University, Piscataway, NJ 08854-8019, USA, [oo]Avant-garde Materials Simulation, Germany, [pp]Chair for Theoretical Chemistry and Catalysis Research Center, Technische Universität München, Lichtenbergstr. 4, D-85747 Garching, Germany, [qq]Department of Geosciences, Center for Materials by Design, and Institute for Advanced Computational Science, SUNY Stony Brook, NY 11794-2100, USA, [rr]Skolkovo Institute of Science and Technology, Skolkovo Innovation Centers, Bldg. 3, Moscow Region, 143026, Russia, [ss]Moscow Institute of Physics and Technology, 9 Institutskiy Lane, Dolgoprudny City, Moscow Region 141700, Russia, [tt]International Center for Materials Discovery, School of Materials Science and Engineering, Northwestern Polytechnical University, Xi'an 710072, China, [uu]Department of Materials Science and Metallurgy, University of Cambridge, 27 Charles Babbage Road, Cambridge CB3 0FS, England, [vv]Department of Physics and Astronomy, University College London, Gower St., London WC1E 6BT, England, [ww]Department of Chemistry, New York University, New York, NY 10003, USA, [xx]Physics and Materials Science Research Unit, University of Luxembourg, L-1511 Luxembourg, [yy]Courant Institute of Mathematical Sciences, New York University, New York, NY 10012, USA, [zz]NYU-ECNU Center for Computational Chemistry at NYU Shanghai, 3663 Zhongshan Road North, Shanghai 200062, China, and [aaa]Department of Chemistry, Loyola University, New Orleans, LA 70118, USA. *Correspondence e-mail: reilly@ccdc.cam.ac.uk

The sixth blind test of organic crystal structure prediction (CSP) methods has been held, with five target systems: a small nearly rigid molecule, a polymorphic former drug candidate, a chloride salt hydrate, a co-crystal and a bulky flexible molecule. This blind test has seen substantial growth in the number of participants, with the broad range of prediction methods giving a unique insight into the state of the art in the field. Significant progress has been seen in treating flexible molecules, usage of hierarchical approaches to ranking structures, the application of density-functional approximations, and the establishment of new workflows and 'best practices' for performing CSP calculations. All of the targets, apart from a single potentially disordered $Z' = 2$ polymorph of the drug candidate, were predicted by at least one submission. Despite many remaining challenges, it is clear that CSP methods are becoming more applicable to a wider range of real systems, including salts, hydrates and larger flexible molecules. The results also highlight the potential for CSP calculations to complement and augment experimental studies of organic solid forms.

1. Introduction

The ability to predict or explore the solid-state properties of molecules has long been a central aim of computational chemistry and materials science. The ultimate goal of crystal structure prediction (CSP) methods is to be able to explore the possible polymorphs, co-crystals, salts, hydrates *etc.* of a molecule based solely on minimal information such as its two-dimensional chemical diagram. This information could be used to predict or design novel solid forms, or determine the chance of undesirable polymorphs or solid forms occurring. The latter application of CSP methods is of particular importance for active pharmaceutical ingredients, due to the time and material cost of experimental solid-form screening and the serious consequences of unforeseen polymorphism or alternative solid forms.

Progress in the development of organic CSP methods over the past 15 years has been charted in a series of blind tests, hosted by the Cambridge Crystallographic Data Centre (CCDC). Five blind tests have been held to date, in 1999 (Lommerse *et al.*, 2000), 2001 (Motherwell *et al.*, 2002), 2004 (Day *et al.*, 2005), 2007 (Day *et al.*, 2009) and 2010 (Bardwell *et*

al., 2011). Participants were provided with the two-dimensional chemical diagram and crystallization conditions of a set of target systems where the experimental structure had been determined but not yet reported.

These tests have shown many advances, with the range and size of the target systems expanding from three relatively 'simple' molecules (Lommerse *et al.*, 2000), to tackling 'drug-like' molecules, co-crystals and polymorphic systems in the most recent fifth blind test (Bardwell *et al.*, 2011). In the fourth and fifth blind tests, all systems were predicted by at least one method (Neumann *et al.*, 2008; Day *et al.*, 2009; Bardwell *et al.*, 2011). However, the tests have highlighted many challenges, including accuracy of ranking methods, their computational cost and the applicability of methods for the full range of solid-form types, with salts, hydrates and larger molecules proving challenging in previous blind tests.

For many years, the focus of CSP research and the blind tests was often on predicting 'the' crystal structure of a molecule, with participants in previous blind tests submitting only three official predictions for each target. Recently, CSP methods have moved towards understanding the solid-form landscape of the putative structures they generate, with various factors influencing which structures are likely to be found experimentally (Price, 2013). At the same time, there has been considerable interest in using CSP methods to augment and understand experimental solid-form screening of pharmaceuticals (see, for example: Bhardwaj *et al.*, 2013; Ismail *et al.*, 2013; Kuleshova *et al.*, 2013; Neumann *et al.*, 2015), organic semiconductors (Valle *et al.*, 2008) and microporous materials (Pyzer-Knapp *et al.*, 2014). Density-functional approximations (DFAs), which have been some of the most promising tools for ranking the stability of possible crystal structures have also developed considerably, with many new van der Waals (vdW)-inclusive methods (Klimeš & Michaelides, 2012) particularly suited to modelling molecular materials (Reilly & Tkatchenko, 2015; Kronik & Tkatchenko, 2014; Brandenburg & Grimme, 2014). New developments in CSP codes and algorithms have also been reported (Habgood *et al.*, 2015; Wang *et al.*, 2012; Lund *et al.*, 2015; Zhu *et al.*, 2012; Obata & Goto, 2015), while there have been a number of new insights into conformational polymorphism (Cruz-Cabeza & Bernstein, 2014; Thompson & Day, 2014).

On the basis of this shift in the focus of CSP and new methodological developments and insights, a sixth blind test of organic CSP methods was launched in 2014. The aims of this test were to provide a fair benchmark of the state-of-the-art in CSP methodology, to spur on the continued development of CSP methods, and to provide a platform to communicate progress and challenges for CSP research with the wider scientific community (Groom & Reilly, 2014). To this end, this blind test has seen more challenging and 'realistic' target systems and changes in the nature of submissions to ensure as much information and as many insights as possible can be gained from the blind test.

This paper reports the overall results of the blind test, and its structure is as follows: the blind-test procedure and selection of targets is outlined in §2, a brief report of the methods and approaches employed is given in §3 and a summary and discussion of the results is presented in §4, including a discussion of current challenges in §4.8. With 25 submissions, the volume of data and information precludes a detailed discussion of every result. However, the supporting-information documents of each submission (part of the supporting information of this paper) provide important context for the trends and general results presented in the main paper, and the interested reader is encouraged to consult these.

2. Organization and approach

Previous blind tests largely followed the same format with the number and complexity of the target systems increasing over the years. Following dialogue with the CSP community in early 2014, a number of changes were made to the organization of the sixth blind test, which are outlined in the following subsections.

2.1. Target categories and selection

In the previous blind test (Bardwell *et al.*, 2011), six target categories were employed, covering simple and more complex rigid molecules, partially flexible molecules, salts and co-crystals, flexible molecules and polymorphic systems. Finding unpublished crystal structures of small rigid molecules containing only CHNO atoms proved very difficult in the fifth blind test, as did finding a polymorphic system (Bardwell *et al.*, 2011). Therefore, the target categories for the sixth blind test were adjusted to remove the small rigid CHNO molecule target and the separate polymorphic system. In addition, co-crystals and salts, which had been a single category previously, were split into two separate categories, resulting in five target categories:

(1) Rigid molecules, with functional groups restricted to CHNO, halogens, S, P, B; one molecule in the asymmetric unit; up to about 30 atoms.

(2) Partially flexible molecules with two to four internal degrees of freedom; one molecule in the asymmetric unit; up to about 40 atoms.

(3) Partially flexible molecule with one or two internal degrees of freedom as a salt; two charged components in the asymmetric unit, in any space group; up to about 40 atoms.

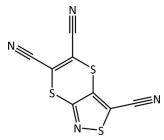
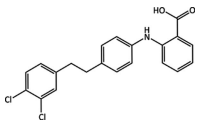
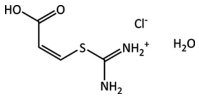
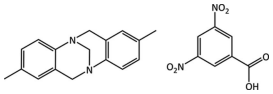
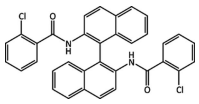
(4) Multiple partially flexible (one or two degrees of freedom) independent molecules as a co-crystal or solvate in any space group; up to about 40 atoms.

(5) Molecules with four to eight internal degrees of freedom; no more than two molecules in the asymmetric unit, in any space group; 50–60 atoms.

One of the most challenging aspects of organizing the blind tests has been finding suitable unpublished crystal structures that fit these categories. In addition to being unpublished, the structures must be of high quality and have all atoms located. As in previous blind tests, the structures were also required to be free of disorder. The collection of potential experimental structures for these categories took place in summer 2014. A number of crystallographers were contacted and asked to send

Table 1
Two-dimensional chemical diagrams, crystallization conditions for the five target systems in the sixth blind test, including information disclosed to participants initially and following queries, as well as a summary of the full predictions for each target system.

Separate lists and re-ranking submissions are not counted in these totals, but the best rank given does include re-ranking attempts. See §2.1 for more details of the categories.

Target	Chemical diagram	Crystallization conditions, remarks and clarifications	Attempted predictions	Times generated	Best rank (incl. re-ranking)
(XXII)		Crystallized from an acetone/water mixture; chiral-like character due to potential flexibility of the six-membered ring, but no chiral precursors used in synthesis.	21	12	1
(XXIII)		Five known polymorphs (*A*–*E*); three $Z' = 1$ (*A*, *B*, *D*), two $Z' = 2$ (*C* and *E*). The most stable polymorphs at 257 and 293 K are both $Z' = 1$. Crystallization conditions include slow evaporation of acetone solution and of ethyl acetate: water mixture.	*A*, *B* and *D*: 14; *C* and *E*: 3	*A*: 4, *B*: 8, *C*: 1, *D*: 3, *E*: 0	*A*: 23, *B*: 1, *C*: 6, *D*: 2, *E*: –
(XXIV)		Crystallized from 1 *M* HCl solution. The substituents of the C═C double bond are in the *cis* configuration.	8	1	2
(XXV)		Slow evaporation of a methanol solution, which contained a racemic mixture of the enantiomers of Tröger's base.	14	5	1
(XXVI)		Slow evaporation from 1:1 mixture of hexane and dichloromethane. No chiral precursors used in synthesis.	12	3	1

information on any suitable targets directly to an external referee, Professor Richard Cooper (University of Oxford). A general request for structures was also included in the announcement of the blind test (Groom & Reilly, 2014). The full experimental structures were known only to the external referee, who also made the final selection of candidates, enabling the CCDC itself to participate in the blind test.

2.1.1. Selection of suitable targets. Following the initial requests, 20 unpublished structures were submitted for consideration. Of these, ten were considered candidates for category 2, four were considered for category 4, and two fell into each of the remaining categories. A further request yielded some additional possible category 1 and 2 structures. The final targets are given in Table 1 and are numbered (XXII)–(XXVI), following on from the 21 molecules and systems studied in previous blind tests.

All three potential category 1 molecules contained one or more ring systems with more than one possible conformation. Molecule (XXII) contains no rotatable bonds but the molecule is 'hinged' about the six-membered ring, introducing some flexibility, with the flat molecule representing a saddle point *in vacuo*. However, the hinged conformation and flexibility was deemed to be predictable, although participants were not provided with the conformation.

Molecule (XXIII) was disclosed along with five known crystal structures (*A*–*E*) and experimental determination of the most stable polymorphs at 257 and 293 K through slurrying experiments. The molecule formally has five rotatable bonds but an intramolecular hydrogen bond between the amine and carboxylic acid group constrains two of these to be almost planar in the observed crystal structures, although a complete CSP calculation would need to explore the possibility of the molecule not forming such a hydrogen bond. The presence of two $Z' = 2$ polymorphs (*C* and *E*) also stretches the requirements of category 2, but given there were three other $Z' = 1$ crystal structures as potential structure prediction targets, it was decided that this would not make the target too difficult. One of the two molecules in the asymmetric unit of form *E* has significantly larger anisotropic displacement parameters than the other, particularly for the ethyl linker between the two phenyl rings (see Fig. S1 of the supporting information). While this suggests that there is potentially disorder in the structure, it was still deemed a valid target.

Structure (XXIV) was chosen from two candidates and satisfied the criteria of category 3. Although containing only 11 non-H atoms, it did contain an additional solvent of crystallization, which increases the difficulty of the structure prediction problem.

Structure (XXV) was chosen from four candidates as the best example of a co-crystal that satisfied the category 4 criteria. Both molecules in the structure appeared to be quite rigid, but the two possible hydrogen-bonding interactions between the molecules retained some of the complexity. The original experimental data for molecule (XXV) were collected

at room temperature. They were remeasured after the blind test at 100 K, which revealed that there is a significant amount of proton transfer from the carboxylic acid group to the amine. A competitive refinement determined proton occupancies of 0.58 (3) on the carboxylic acid oxygen and 0.42 (3) on the nitrogen.

Molecule (XXVI) was one of two possibilities for category 5 and contains five rotatable bonds, with each half of the topologically symmetric molecule adopting different conformations in the solid state. Molecule (XXVI) was screened for additional polymorphs by Johnson Matthey (Pharmorphix). The study found one high-temperature polymorph and several solvates.

2.2. Structure of the blind test

The primary aims of the sixth blind test were to enable the CSP community to perform a fair benchmark of their methodologies, provide a platform to communicate progress and state-of-the art in the field and to spur new development in the methodologies. To further these aims, the format and structure of this latest blind test differs from the previous one in a number of areas.

In previous blind tests, participants were allowed to submit three predicted crystal structures for each target as their principal predictions, although they were encouraged to submit extended lists of structures resulting from their predictions for further analysis. This is not in keeping with the more recent focus of CSP methods on solid form landscapes and the insight they can provide on the multiple likely solid forms of a molecule. The restriction of submitting only three structures as principal predictions also created an arbitrary cut-off point for what was considered a successful prediction. In choosing their three structures, some participants combined different analysis or ranking approaches, highlighting that various information and calculations can be complementary.

Reflecting all these points, each submission in the sixth blind test could contain up to 100 predicted structures ranked in order of their likelihood using some form of fitness function. Participants were also allowed to submit a second list of 100 structures, which could be generated or re-ranked using alternative methods. The purpose of these changes was to maximise the information and insight gained from the blind test. For this reason, re-ranking submissions, where a submission solely re-ranked structures provided by other participants, were also permitted for this blind test. This allowed a number of research groups developing ranking approaches [*e.g.* bespoke potentials, density-functional theory (DFT) and quantum-chemical methods] to apply their methods under blind-test conditions.

Participants were required to submit a supporting-information document that would provide a clear summary of their methodology at the time of submission, as opposed to optionally providing one afterwards. These changes in procedure were agreed through dialogue with potential participants in spring and summer 2014. Previous participants in blind tests and anyone who had expressed interest in any new blind tests were invited *via* email to take part in the sixth blind test, while an open invitation was published on the CCDC and IUCr websites and in *Acta Cryst. B* (Groom & Reilly, 2014).

The two-dimensional chemical diagrams and crystallization conditions (Table 1) were sent to researchers interested in participating on 12 September 2014 by the referee, with a deadline for submissions of 31 August 2015. As in previous blind tests, participants were not required to attempt all five target systems. A number of researchers expressed interest after the start date and were also allowed to participate. In the week following the submission deadline the predicted structures were compared with the experimentally known ones by the CCDC and the referee. Participants were then sent the experimental structures on 7 September 2015, and the results confirmed by mid-September 2015. A workshop was held to discuss the results in October 2015 in Cambridge, UK.

2.3. Assessment of predictions

The predicted crystal structures submitted by participants were compared with the experimentally known crystal structures using the Crystal Packing Similarity Tool (Chisholm & Motherwell, 2005), as available through the CSD Python API (Groom *et al.*, 2016) and *Mercury* 3.6 (Macrae *et al.*, 2008). The tool represents a crystal structure using a cluster of N molecules comprised of a central reference molecule and $(N - 1)$ nearest-neighbour molecules. The distances and a subset of the triangles that define the reference cluster are then used as a three-dimensional substructure-search query within the comparison structure. For this search, two molecules within the packing shells are considered to match if these distances agree within 25% and the angles of the triangles agree within 25°. Those molecules that match are then overlaid and a root mean-squared deviation (RMSD) is calculated.

The result of the comparison is a number of molecules that match, n, between the two packing shells and a corresponding RMSD_n for those matching molecules. Where multiple clusters can be defined for an input crystal (*i.e.* $Z' > 1$ or structures submitted in $P1$ symmetry) the best result is retained. The Crystal Packing Similarity Tool normally considers only heavy atoms when calculating distances and angles within clusters and for the final RMSD analysis, ignoring H-atom positions due to their limited accuracy in standard X-ray diffraction crystal structures. However, matching and overlay of the heavy atoms does require the number of H atoms bonded to them to be the same. Predicted structures were deemed to match an experimental structure when 20 out of 20 molecules matched. The largest RMSD_{20} value was approximately 0.8 Å. A single predicted structure of (XXV) approximately matched the experimental structure, but with an RMSD of more than 1.2 Å, which was deemed too far from the experimental geometry.

For (XXIII), some of the predicted crystal structures have the same heavy-atom positions as the experimental structure but place the carboxylic acid H atom on the oxygen closest to the NH group. The analysis for these systems was therefore performed twice, once requiring the H atom to be located as in

Table 2
List of members of each team/submission (* denotes corresponding author), as well as a *brief* summary of the generation and ranking methods used.

Please refer to §3 for an overview of the methods, Tables S10 and S11 of the supporting information, and each submission's supporting-information document for more details. Helmholtz free-energy contributions are denoted by F_{vib}, polarizable continuum model is abbreviated PCM, while Monte Carlo is abbreviated MC.

			Final ranking method(s)	
Team	Members	Generation method	List One (L1)	List Two (L2)
1	Chadha,* Singh	MC simulated annealing	COMPASS (2.8) force field	–
2	Cole,* McCabe, Read, Reilly, Shields	CSD analogues	Fitted exp-6 potential	–
3	Day*, Bygrave, Campbell, Case, Gee, McMahon, Nyman, Pulido, Taylor, Yang	Quasi-random search (Sobol')	Atomic multipoles and exp-6	F_{vib} contributions [(XXII) and (XXV)], PCM $\epsilon = 3$ [(XXIV) and (XXVI)]
4	Dzyabchenko	Grid search	Empirical potential	–
5	van Eijck	Random search	Atomic charges, intramolecular 6-31G** energies and exp-6	–
6	Elking, Fusti-Molnar	Random generation	Empirical potential	PBE+XDM
7	de Jong, van den Ende,* de Gelder, de Klerk, Bylsma, de Wijs, Meekes, Cuppen	Random search	*q*-GRID method	Smallest critical nucleus size from kinetic MC simulations
8	Lund, Pagola, Orendt, Ferraro, Facelli*	Genetic algorithm	PBE-D2	PBE-D2 for all stages of GA search
9	Obata, Goto*	Grid search	PBE+TS	–
10	Hofmann,* Kuleshova	Random search	Fitted potential	–
11	Lv, Wang, Ma*	Random search	optB86b-vdW	–
12	Curtis, Li, Schober, Cosburn, Lohani, Vacarro, Oberhofer, Reuter, Bhattacharya, Vázquez-Mayagoitia, Ghiringhelli, Marom*	Genetic algorithm	PBE+TS	PBE+MBD
13	Mohamed	MC simulated annealing	Atomic multipoles and exp-6	–
14	Neumann, Kendrick, Leusen	MC parallel tempering	PBE+Neumann–Perrin	Includes $Z' = 2$ structures for (XXIII) and (XXVI)
15	Sugden, Gatsiou, Vasileiadis, Adjiman,* Pantelides*	Quasi-random search (Sobol')	Atomic multipoles and exp-6	–
16	Pickard,* Monserrat, Misquitta, Needs	Random search	PBE+MBD	–
17	Jankiewicz, Metz, Podeszwa,* Szalewicz	Grid search	SAPT(DFT) fitted potential	Alternative SAPT(DFT) fitted potential
18	S. L. Price,* Hylton, L. S. Price, Guo, Watson, Iuzzolino	Quasi-random search (Sobol')	Atomic multipoles and exp-6	Different PCM treatments (all); F_{vib} for all but (XXIV)
19	Metz, Hylton, S. L. Price, Szalewicz*	Quasi-random search (Sobol')	SAPT(DFT) fitted potential	–
20	Vogt, Schneider, Metz, Tuckerman,* Szalewicz*	Random search	SAPT(DFT) fitted potential	–
21	Zhu,* Oganov, Masunov	Evolutionary algorithm	vdW-DF	–
22	Boese	Re-ranking 10	PBE+TS and BLYP-D3	–
23	Brandenburg, Grimme	Re-ranking 18	HF-3c^{atm}	TPSS-D3^{atm}
24	Metz, Guo, Szalewicz	Re-ranking 18	SAPT(DFT) fitted potential	–
25	Hoja, Ko, Car, DiStasio Jr, Tkatchenko*	Re-ranking 18	PBE+MBD	F_{vib} contributions

the experimental structure and a second time where the H-atom location and connectivity was not considered.

In the case of (XXIV), each of the three components in the asymmetric unit counts towards N, therefore a cluster of 20 components does not amount to the same physical extent as for the other systems. In addition, H-atom positions are particularly important for this system. Therefore, initial analysis was performed ignoring H-atom positions and with N = 20. If a match was found, the analysis for that structure was re-run considering H-atom positions and with N = 60 to confirm the match.

Finally, after the blind test had concluded it was discovered that the hydrogen-bonding proton in (XXV) is disordered, making the structure a mixture of a molecular salt and a co-crystal. Therefore, the analysis of (XXV) was performed twice to find both co-crystal and salt matches to the experimental heavy-atom coordinates.

3. Methodologies

There are a wide variety of approaches to predicting organic crystal structures. The larger number of submissions in this blind test has seen a number of new approaches being applied in a blind test for the first time. Broadly speaking, the CSP process can be broken down into a series of steps:

(i) Exploration of the conformational preferences of the target molecules.

(ii) Generating plausible crystal-packing arrangements of the target molecules.

(iii) Ranking the likelihood of resulting crystal structures forming using some form of scoring or fitness function.

There are, however, many variations on these steps. In this section we summarize some of the approaches used in the current blind test. Brief details of the approach used in each submission are given in Table 2, while full details are provided in the supporting information document that accompanied each submission.

3.1. Molecular structure generation and conformational analysis

For many approaches to predicting crystal structures, the first stage is to explore the conformational flexibility of the target molecules. This can help to define a set of rigid conformations that some methods use for structure generation, while in other methods this information is used to define and limit the flexible degrees of freedom explored in tandem with the unit-cell degrees of freedom. Not all approaches require this information though, with some exploring molecular degrees of freedom in the search stage in an unbiased way or with implicit limits imposed by the search strategy.

In several approaches, the initial starting conformations for molecules were determined using *ab initio* calculations of isolated molecules in the gas phase, including 'scans' of specific degrees of freedom (such as torsions), which have been used to understand the extent of flexibility of a molecule and define conformations. Information on conformational preferences from the Cambridge Structural Database (Bruno *et al.*, 2004) has been combined with *ab initio* data in some methods, and also used to directly generate conformations in one approach.

In some cases, force fields have been used for the initial stages of exploring flexibility, which allows one to apply more exhaustive methods for exploring conformational flexibility, such as low-mode conformational searches (Kolossváry & Guida, 1996), systematic grid searches and perturbations of initial conformations, including CONFLEX conformational searches (Goto & Osawa, 1989; Goto & Osawa, 1993). In many cases, the resulting conformations were then optimized using *ab initio* methods.

3.2. Crystal structure generation

There are a plethora of methods for generating possible organic crystal structures, which requires exploring the degrees of freedom of the unit cell (up to six lattice parameters), the position and orientation of molecules in the unit cell and, in some cases, internal molecular degrees of freedom. As in the previous blind test, the majority of methods employ some variation on random or quasi-random searches to generate trial crystal structures (Submissions 3, 5–7, 10, 11, 15, 16 and 18–20), with four submissions (3, 15, 18, 19) using low-discrepancy Sobol' sequences (Sobol', 1967). Monte Carlo simulated annealing (Submissions 1 and 13) and parallel tempering (Submission 14) have also been used, as have systematic grid searches (Submissions 4, 9, 17) and evolutionary and genetic algorithms (Submissions 8, 12 and 21). Shape matching of the target systems to known experimental structures in the CSD has been employed in one submission to generate analogue crystal structures (Submission 2).

An important choice in the structure-generation process is the consideration of the set of space groups or Z values to consider in the search. The majority of submissions imposed crystallographic symmetry, explicitly exploring a set of space groups, typically chosen on the basis of frequencies of occurrence in the CSD. For some submissions, parts of the ranking or generation process, including some DFT codes and MD simulations, do not fully conserve the crystallographic symmetry. Software and utilities including *PLATON* (Spek, 2009), *PyMatGen* (Ong *et al.*, 2013), *FINDSYM* (Stokes & Hatch, 2005) and *Spglib* (Spglib, 2015) have been used to detect and enforce such symmetry in the final submitted structures.

As noted above, some methods explore the molecular degrees of freedom as part of the search for putative crystal structures. This can be important, as conformers that appear unstable for the molecule *in vacuo* can be found in the stable crystal structure of the molecule (Thompson & Day, 2014), while in some cases the solid-state conformation may not even correspond to a conformer on the isolated molecule's potential-energy surface. More than half of the search methods in the present blind test allowed for some molecular flexibility while exploring the search space and many of those that performed only a rigid-conformation search used a set of likely or low-energy conformations or were attempting only molecule (XXII), which contains no rotatable bonds.

3.3. Optimization and ranking

The final stage of predicting crystal structures is to optimize or minimize the energy of the raw crystal structures generated and then rank them in order of stability or likelihood of occurrence. All of the submissions in this blind test used some form of energy-based metric to rank structures.

In a number of methods, a hierarchical approach has been adopted, in which a less intensive computational method or algorithm is used initially, for example, generic or tailor-made empirical potentials (Neumann, 2008) or 'coarse' evaluation of DFT energies, including the use of a modified Harris approximation to calculate solid-state charge densities from molecular charge densities (Submission 12). More computationally demanding methods and algorithms were then employed for the final set of structures closest to the global minimum. In a number of submissions the final ranking was performed using potentials based on distributed multipole electrostatics (Stone, 2005; Price *et al.*, 2010), *ab initio* intramolecular energies (Kazantsev *et al.*, 2011; Habgood *et al.*, 2015) and various dispersion–repulsion potentials. Other methods employed generic force fields, sometimes fitted to *ab initio* or experimental data or augmented with *ab initio* conformational energies (van Eijck *et al.*, 2001*a*), while three

submissions shared potentials derived from symmetry-adapted perturbation theory based on DFT [SAPT(DFT)] calculations (Misquitta *et al.*, 2005) of (XXII) (Submissions 17, 19 and 20).

DFT has seen extensive use with a range of vdW-inclusive density-functional approximations (DFAs) (Klimeš & Michaelides, 2012) being applied. These include the Neumann–Perrin (Neumann & Perrin, 2005), D2 (Grimme, 2006), TS (Tkatchenko & Scheffler, 2009), XDM (Becke & Johnson, 2007), D3 (Grimme *et al.*, 2010) and MBD (Tkatchenko *et al.*, 2012; Ambrosetti *et al.*, 2014) methods, as well as two vdW density functionals, vdW-DF (Dion *et al.*, 2004) and optB86b-vdW (Klimeš *et al.*, 2011). These treatments differ in the way the dispersion interaction is modelled. Many of the methods are based on C_6/R^6 terms, and differ in the origin of the C_6 coefficients and whether higher-order terms (*i.e.* C_8 and/or C_{10} term, as in D3 and XDM) are included. Many-body vdW effects, which have been shown to be increasingly important for molecular materials (Reilly & Tkatchenko, 2015) including for polymorphism (Marom *et al.*, 2013), are also modelled by some methods, either using three-body Axilrod–Teller–Muto (Axilrod & Teller, 1943) contributions (D3), or a full many-body treatment using coupled atomic response functions (MBD). Most of these have been combined with the Perdew, Burke and Ernzerhof (PBE) semi-local density functional (Perdew *et al.*, 1996), with the TPSS (Tao *et al.*, 2003) and BLYP (Lee *et al.*, 1988; Becke, 1988) functionals also used. The two vdW density functionals feature an additional density-dependent term in the functional to approximate long-range or non-local correlation. See Table 2 and the supporting-information documents for details of the methods used by each submission.

The ranking methods mentioned above are normally used to estimate a lattice-energy difference between polymorphs. In reality, the relative thermodynamic stability of polymorphs is governed by free-energy differences, which include the contributions of zero-point and thermal motion to the enthalpy and entropy of the lattice, with configurational entropy also important in cases of disorder. Such contributions can affect the rank ordering of polymorphs (van Eijck *et al.*, 2001*b*; Reilly & Tkatchenko, 2014; Nyman & Day, 2015). A number of methods have involved the use of lattice dynamics (Born & Huang, 1954; Dove, 1993) to estimate harmonic Helmholtz free energies. The effects of anharmonicity of the free energy have been captured using an extension of lattice dynamics (vibrational self-consistent field theory; Monserrat *et al.*, 2013), while molecular-dynamics (MD) simulations have been used to generate time- and ensemble-averaged structures and lattice energies at experimental temperatures and pressures. Finally, one submission considered kinetic aspects by ranking the structures generated based on the smallest critical-nucleus size determined from kinetic Monte Carlo simulations (Boerrigter *et al.*, 2004; Deij *et al.*, 2007). However, although crystallization conditions (*e.g.* solvent of crystallization) were provided as part of the blind test, none of the methods used this information as part of the CSP process.

3.4. Analysis and post-processing

Many CSP methods involve analysis and post-processing of the structures generated. The nature of search algorithms frequently leads to the same structure being generated multiple times. In some approaches this is used as a measure or indication of the search completeness (Case *et al.*, 2016), but in all cases further calculations on duplicate structures waste computational resources. Many different approaches are used to detect and remove duplicates, ranging from packing-similarity analysis (discussed in §2.3), powder-pattern similarity (de Gelder *et al.*, 2001; Hofmann & Kuleshova, 2005), fingerprint functions (Oganov & Valle, 2009) and radial distribution functions (Verwer & Leusen, 1998). In some cases, structures that were very similar (*e.g.* structures with closely related hydrogen-bonding patterns or similar gross packings) were also removed, on the basis that such structures are unlikely to exist as distinct points or minima on the free-energy solid-form landscape. Filtering of results based on CSD informatics has also been used.

Post-processing of structures has been used to investigate the sensitivity of the results to the method used to rank them, *e.g.* to different repulsion–dispersion parameters, different quality wavefunctions or a polarizable continuum model for distributed multipoles and intramolecular energy contributions. As noted above, MD simulations and lattice-dynamics calculations can be used to provide finite-temperature estimates of relative stability of different structures. Such methods also provide an indication of the inherent finite-temperature and mechanical stability of the crystal structures generated. The crystal-adiabatic free-energy dynamics method (Yu & Tuckerman, 2011) was used to explore the stability and relations of structures in one submission.

3.5. Changes in the methodologies

Comparing the present blind test with previous ones, we can see a number of changes in the approaches and methods employed. Firstly, there has been a change in the aims of some methods, which are not targeting an accurate prediction of the experimental crystal structure, but rather explicitly aiming to generate the experimental lattice somewhere within their low-energy structures. These results might then feed into other re-ranking approaches or analysis.

The protocols and workflows used by the different methods have also been developed and refined. Many approaches are now employing more exhaustive searches, considering more space groups, as well as larger regions of conformational space or a greater number of rigid conformations. In many instances, these expanded searches are guided by analysis of the results to inform on their completeness or sensitivity to levels of theory. This already feeds directly into the search process for some methods, while in others it is used to refine future searches (see individual supporting-information documents for more details).

One of the most significant changes is in the ranking methods employed. Solid-state DFT calculations have been used by 12 submissions, a significant increase compared with

Table 3
Results of each submission in the sixth blind test, broken down by target system and the two lists (L1 and L2; *cf.* Table 2) that could be submitted.

Numbers indicate the position in the submitted list at which an experimental structure was found, a dash (–) indicates that the experimental structure was not found in the submitted predicted structures, and a blank entry indicates no prediction was attempted. For re-ranking submissions, an asterisk (*) indicates that the experimental structure was not present in the set of re-ranked structures. For (XXIII) *C* and *E*, only submissions that explicitly considered $Z' = 2$ searches are noted in the table. Numbers in parentheses for (XXIII) indicate that the heavy-atom positions were predicted, but not the correct position of the H atom of the carboxylic acid.

		(XXII)		(XXIII)										(XXIV)		(XXV)		(XXVI)	
				A		B		C		D		E							
Team	Members	L1	L2	L1	L2	L1	L2	L1	L2	L1	L2	L1	L2	L1	L2	L1	L2	L1	L2
1	Chadha & Singh	–		–		–				–								–	
2	Cole *et al.*	–		–		–		–		–		–				–		–	
3	Day *et al.*	3	1	23	–	–	75			75	–			–	–	–	–	–	–
4	Dzyabchenko	1														–		–	
5	van Eijck	4		83		20				–				–		1		–	
6	Elking & Fusti-Molnar	–	–	–	–	78	–			(73)	–			–	–	–	–	8	1
7	van den Ende, Cuppen *et al.*	9	90	–	–	–	–			–	–					–	–		
8	Facelli *et al.*	–	–	–		–				–				–		–			
9	Obata & Goto	2		–		13				(66)						–			
10	Hofmann & Kuleshova	–		–		–		–		–		–		–		–		–	
11	Lv, Wang, Ma	–	–																
12	Marom *et al.*	–	–																
13	Mohamed	1		–		88				–						–		–	
14	Neumann, Kendrick, Leusen	2		26	85	2	4	–	6	11	39	–	–	2		6		1	1
15	Pantelides, Adjiman *et al.*	6		70		13				–						1		–	
16	Pickard *et al.*	–																	
17	Podeszwa *et al.*	8	3																
18	Price *et al.*	6	2	–	–	1	2			85	44			–	–	1	1	2	1
19	Szalewicz *et al.*	–																	
20	Tuckerman, Szalewicz *et al.*	4																	
21	Zhu, Oganov, Masunov	3	–	–		–				–				–		2		–	
22	Boese	*		*		*		*		*		*		*		*		*	
23	Brandenburg & Grimme	–	–	–	–	11	1			–	–			*	*	2		–	–
24	Szalewicz *et al.*													*					
25	Tkatchenko *et al.*	3	1	–	–	2	5			14	2			*		1			

the fifth blind test, where only two submissions employed DFT. Many other submissions used more computationally demanding or bespoke potentials than in the past, with the use of generic empirical potentials and simple point-charge electrostatics as a final ranking method further declining to only a few submissions. In addition to focusing on better lattice energies, more methods are calculating free energies to rank the experimental structures at finite temperatures.

4. Results and discussion

The sixth blind test has been the biggest to date: 25 distinct submissions were received, of which seven were full submissions, 14 attempted some of the targets, and four involved re-ranking structures generated using another method (by another team). This compares to 15 submissions in total in the previous blind test. Table 2 lists those who contributed to each submission along with a very brief summary of the methods employed, while Tables S10 and S11 in the supporting information provide a more detailed summary of the methods employed. The supporting-information document also contains details on access to computational data resulting from the blind test.

The overall results of the blind test are presented in Table 1, which lists for each system the number of attempts at prediction, the number of times the experimental structure was generated and the best ranking of that structure within the submitted lists. Table 3 provides the full results of each submission, broken down by target and the two lists. Tables showing the relative deviation between the lattice parameters

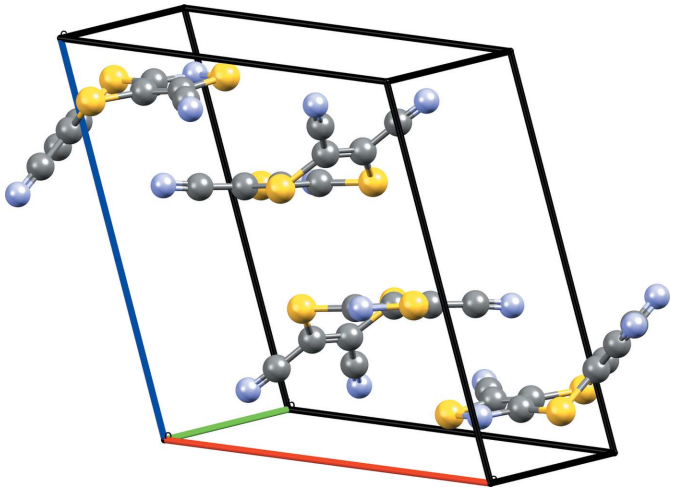


Figure 1
Experimental crystal structure of (XXII); C atoms are in grey, N in blue and S in yellow.

of the predicted and experimental structures, as well as crystal and conformational RMSD values, are provided in the supporting information.

Given the number of submissions and large volume of data produced, an exhaustive account of the results is beyond the scope of this publication. Instead, we now focus on describing the experimental structures of the target systems and the trends and challenges in predicting and modelling them. A broad discussion of the results is then presented in §4.7.

4.1. Target (XXII)

Tricyano-1,4-dithiino[*c*]-isothiazole ($C_8N_4S_3$) was crystallized from an acetone:water mixture with X-ray diffraction data collected at 150 K (Horton & Gossel, 2016). The molecule crystallizes in the monoclinic $P2_1/n$ space group. In the experimental crystal structure the molecules form rows of molecules clasped together but offset from one another.

As Fig. 1 shows, the six-membered ring containing two S atoms is hinged, with an angle between the two C═C—S planes of 44.4°. This makes the molecule chiral, although calculations suggest the barrier to interconversion may be small. As communicated to participants, no chiral precursors were used during synthesis and therefore crystallization in a centrosymmetric space group is not unexpected. A search of the CSD (Version 5.37; Groom *et al.*, 2016; *R*-factor < 0.075; no errors, disorder or polymeric systems; organics only) for the six-membered dithiino ring, finds 77 structures that contain it, the majority of which feature the molecule in the hinged conformation with an angle between the two C═C—S planes of > 40°. Around 15 molecules have angles close to or at 0°, but many sit on a symmetry element such as an inversion centre, which can result in conformational bias (Cruz-Cabeza *et al.*, 2012).

Some force fields fail to adequately represent the hinge of this molecule, instead predicting that the molecule should be completely flat. Such a flat molecule is, as noted by a number of groups, a saddle point between the S atoms being above or below the mean plane of the molecule. Even some DFT methods have difficulty with the conformation of the molecule, which can be traced back to issues with the treatment of the S atoms in some vdW approaches. As a result, a number of submissions, even fully *ab initio* ones, featured crystal structures with flat or nearly flat molecules, although intermolecular interactions will also stabilize the planar conformation in some crystal structures.

Overall though, the experimental crystal structure was successfully generated and ranked by 12 out of 21 submissions, with all but one of those ranking the known experimental structure within the top eight most likely or stable structures and four ranking it as number one. A comparison of the predicted structures with the experimental one is given in Table S1. There is no definite trend in performance, with a range of treatments from generic potentials, point and multipole electrostatics, and DFAs ranking the experimental structure as being one of the most stable. Some of the other predicted structures are similar to the experimental one (for example, featuring a shift of the inversion centre), while others have more layered structures. Interestingly, many low-energy putative structures were found by multiple submissions. Solid-form screening of (XXII) may shed light on whether these predicted crystal structures could be isolated experimentally.

A number of second lists of predicted structures were submitted for (XXII) and three submissions re-ranked other structures, which gives an insight into the sensitivity of the ranking to the method employed. Three submissions (Podeszwa *et al.*, Szalewicz *et al.*, and Tuckerman, Szalewicz *et al.*) shared a set of potentials fitted to SAPT(DFT) calculations. Different functional forms for the potential, necessitated by the different software employed by the different methods, led to significantly different rankings for the experimental structure, while the ranking was sensitive to errors in the fitting procedure. Tkatchenko *et al.* re-ranked structures provided by Price *et al.* using the PBE+MBD functional, which improved the ranking compared with that with the FIT potential and multipole electrostatics. The second lists of Day *et al.*, Price *et al.* and Tkatchenko *et al.* all employed Helmholtz free energies, which changed the rank order of the putative structures and, in all three cases, improved the ranking of the experimentally known structure. In addition to free energies, two methods (Tuckerman, Szalewicz *et al.* and Podeszwa *et al.*) used MD simulations to obtain thermally averaged structures and potential energies at 300 K. The actual temperature of the diffraction experiment (150 K) was not disclosed to participants. These simulations confirm the stability of the experimental form on the potential-energy surface of the SAPT(DFT)-fitted potential. In post-test analysis, Marom *et al.* have also explored the rank ordering of low-energy structures of (XXII) using the PBE0 hybrid functional (Adamo & Barone, 1999) alongside different dispersion contributions.

4.2. Target (XXIII)

2-((4-(3,4-Dichlorophenethyl)phenyl)amino)benzoic acid ($C_{21}H_{17}Cl_2N_1O_2$) is a former drug candidate. (XXIII) targeted β-amyloid aggregation (Simons *et al.*, 2009; Augelli-Szafran *et al.*, 2002), which is believed to play an important role in Alzheimer's disease. Five polymorphs of (XXIII) are known, three $Z' = 1$ structures [forms *A* (Samas, 2016*a*), *B* (Samas, 2016*b*) and *D* (Samas, 2016*d*)] and two $Z' = 2$ structures [forms *C* (Samas, 2016*c*) and *E* (Samas, 2016*e*)]. Forms *A* and *D* crystallize in the monoclinic $P2_1/c$ space group, while forms *B*, *C* and *E* crystallize as triclinic $P\bar{1}$ structures. Slurrying experiments have identified form *A* as being the most stable polymorph at 257 K, while at 293 K form D is the most stable polymorph (Samas, 2015).

All five polymorphs feature $R_2^2(8)$ carboxylic acid hydrogen-bond dimers and intramolecular hydrogen bonds between the NH group and the carbonyl oxygen of the carboxylic acid, which is common in many fenamate structures. Fig. 2 shows the overlay of the conformations of (XXIII) in forms *A*–*D*. Forms *B* and *D* have a similar conformation, while form *A* has the chloro-phenyl ring flipped approximately 180° compared with *B* and *D*. The two molecules in the asymmetric unit of

Figure 2
Molecular conformations found in forms *A*–*D* of (XXIII), overlaid onto the fenemate group of the molecule; form *A* is in blue, form *B* in grey, form *C* molecule 1 is in red, form C molecule 2 in purple and form *D* in orange. H atoms are omitted for clarity.

form *C* are similar, adopting the same torsions about the ethyl but differing in the twist of the phenyl group. The two molecules in form *E* (see Fig. S1) have distinct conformations from those found in forms *A*–*D*, with one molecule having the central phenyl ring rotated by approximately 120° compared with all of the other experimental conformations. Forms *B* and *C* have a similar gross packing, but deviate due to the two different conformations of the molecules in the asymmetric unit of form *C*. Forms *A* and *D* are also related in terms of their packing, featuring similar layers or sheets of molecules as seen in Fig. 3, again, differing only due to the different conformations of the end phenyl group. Given their close resemblance, interconversion of forms *A* and *D*, and forms *B* and *C*, respectively, might be expected to be facile but conversion of *A* or *D* to *B* or *C* might be much slower. Disorder might also be expected, with small energy barriers between some of the conformations.

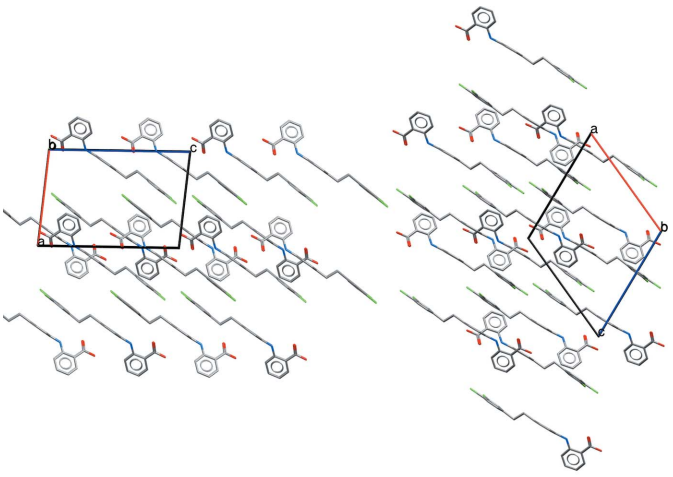
The three $Z' = 1$ forms of (XXIII) were the main targets for this molecule, with 14 attempted predictions and three submissions re-ranking structures. Form *A* was generated four times in the top 100 structures, form *B* ten times and form *D* three times, with two methods (Day *et al.*; Neumann, Leusen, Kendrick) generating all three structures. In some cases the heavy-atom positions of the polymorphs were predicted, but not the correct ordering of the protons of the carboxylic acid dimer. These predictions are not counted in the totals above, as the proton environments are likely to be very different and distinguishable, but are denoted in parenthesis in Table 3.

The ranking of the experimental structures is more varied than for (XXII), with only a few of the predictions ranking the experimental structures as being one of the ten most stable structures, with form *A* having the best rank of 23 (Day *et al.*). A number of submissions predicted form *B* to be the most stable of the three $Z' = 1$ polymorphs, with a highest rank of 1 (Price *et al.*). In all of the experimentally observed conformations the molecule is extended. However, some of the low-energy predicted crystal structures have more compact conformations, with the terminal phenyl ring bending back towards the other end of the molecule. Such conformations could be favoured *in vacuo*, but not necessarily in solution or the solid state (Thompson & Day, 2014). Conformation and packing are the main differences between many of the predicted structures of (XXIII), as the CO_2H dimer motif is found in the majority of low-energy structures.

As for (XXII), second lists and re-ranking submissions shed some light on the sensitivity of the results and methods. Price *et al.* predicted form *D* to be ranked 85th based on lattice energies from distributed multipoles and the FIT intermolecular potential. Re-ranking by Tkatchenko *et al.* placed the experimental structure as 14th in terms of lattice energy. Both submissions employed Helmholtz free energies (calculated at 300 K) in their second lists, which also significantly changed the polymorph rankings, and in the case of Tkatchenko *et al.* changed the relative ordering of the *B* and *D* polymorphs, improving the rank of *D* to second. Shifting through different levels of theory, from minimal basis-set Hartree–Fock theory to DFT (Brandenburg & Grimme, 2014), also altered Brandenburg & Grimme's ranking of form *B* from number 11 to number 1.

Four attempts were made at predicting the $Z' = 2$ polymorphs. Form *C* was predicted by one method (Neumann, Kendrick and Leusen), ranking at number six in a list of both $Z' = 1$ and 2 structures. The second $Z' = 2$ polymorph, form *E*, was not predicted by any submission. The potential disorder in the experimental structure might point to this being difficult to predict, but post-test analysis results suggest that most ranking methods have a valid local minimum corresponding to the experimental structure of form *E*, which means the structure should have been predictable with these methods.

Following the disclosure of the structures after the submission deadline, the experimental structures have been optimized and ranked using a number of different methods. The resulting calculated relative stabilities of the five poly-



(*a*) (*b*)

Figure 3
Crystal structures of (*a*) form *A* and (*b*) form *D* of (XXIII), showing the similar layers found in the two structures. H atoms are omitted for clarity.

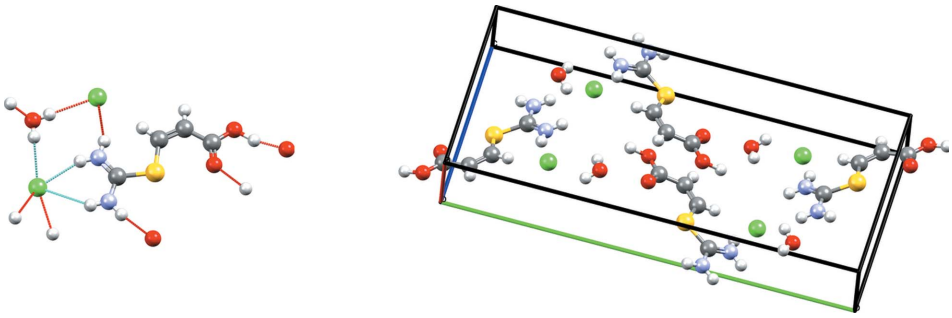


Figure 4
Experimental crystal structure of (XXIV) showing both the hydrogen bonds of the asymmetric unit and the unit cell; C atoms are in grey, H in white, O in red, N in blue, S in yellow and Cl in green.

morphs are presented in Table S12. Of the experimental structures, forms *B* and *C* are most often found to be the lowest-energy polymorph, although they are not generally found as the global minimum. This contrasts with the experimental stabilities from the slurrying data, where form *A* is most stable at 257 K and form *D* at 293 K. Directly comparing their rank or position on the energy landscape of each submission is difficult, as some methods may generate more or fewer local minima than others. This is demonstrated by the combined $Z' = 1$ and 2 list of Neumann, Kendrick and Leusen, where some of the additional $Z' = 2$ structures are lower in energy than some of the $Z' = 1$ structures, making the ranks of the latter worse. However, post-test analysis does suggest that some of the more recent vdW-inclusive DFT methods (*e.g.* TPSS-D3 and PBE+MBD) would have ranked the experimental structures better, perhaps within the top 10–15 putative structures, if applied to a larger set of initial crystal structures or combined with different search methods.

4.3. Target (XXIV)

Target (XXIV) is a chloride salt hydrate of (*Z*)-3-((diaminomethyl)thio)acrylic acid [$(C_4H_8N_2O^2)^+Cl^-\cdot H_2O$], which was crystallized in the monoclinic $P2_1/c$ space group from a 1 *M* HCl solution, with the structure determined at 240 K (Foxman, 2016). The experimental crystal structure is shown in Fig. 4. Graph-set analysis (Etter *et al.*, 1990) yields over 25 distinct hydrogen-bond types. The Cl^- ions are six coordinate, with four short contacts and two longer ones, forming separate $C_2^1(4)$ hydrogen-bond chains with thiouronium groups of the acid and water molecules. An $R_2^2(16)$ ring motif is also formed between carbonyl O atoms and the thiouronium groups of the acid molecules. As the molecule has a relatively flat conformation, the combination of the two motif types is to form interlocking layers or strands of acid molecules.

Of the eight full submissions for this target system, only the method of Neumann, Kendrick and Leusen generated the known experimental structure, ranking it as the second most stable structure with the PBE functional plus the Neumann–Perrin dispersion correction. Other structures in this and other submissions contain a large variety of different hydrogen-bonding patterns. The experimental hydrogen-bonding set is found in a few predictions, while some of the individual motifs (in particular, the $C_2^1(4)$ $Cl^-\cdots$water$\cdots Cl^-$ chains) are found in a number of structures generated by other methods.

As there are three components in the asymmetric unit, this is one of the most challenging target systems in the series of blind tests to date. This is both in terms of generating the complex hydrogen-bond patterns of the crystal structure and the demands of correctly ranking the strength of such interactions. Dealing with charged species, modelling charge penetration (Stone, 2013), capturing the coordination preferences of the Cl^- ion, and modelling polarization within the crystal are all serious challenges for empirical potentials. A number of submissions reported significant re-ordering of their predicted structures based on the type of Cl potential employed, and the dielectric constant used to model the effect of polarization on the electrostatic interactions in the crystal structures. Post-test analysis has borne this out, with some methods ranking the experimental structure more than 20 kJ mol^{-1} above the global minimum. Standard density-functional approximations can also struggle to deal with charged systems and charge transfer adequately due to self-interaction errors (Cohen *et al.*, 2008, 2012), but in the case of (XXIV), DFT provides a good basis for fitting a bespoke potential and ranking the predicted structures.

4.4. Target (XXV)

(XXV) is a multi-component system consisting of 3,5-dinitrobenzoic acid ($C_7H_4N_2O_6$) and 2,8-dimethyl-6*H*,12*H*-5,11-methanodibenzo[*b*,*f*][1,5]diazocine ($C_{17}H_{18}N_2$), also known as Tröger's base. The N atoms of Tröger's base are unable to invert and therefore the molecule is chiral, but the structure was crystallized from a methanol solution that contained both

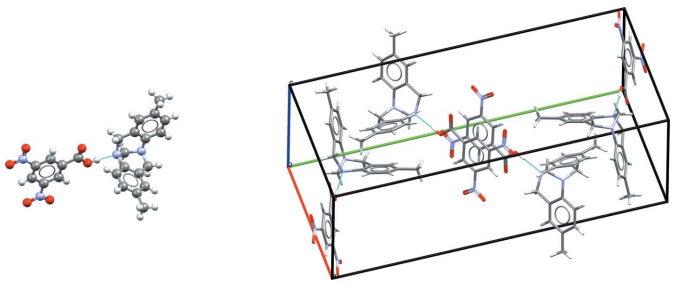


Figure 5
Experimental crystal structure of (XXV) at 300 K, showing the asymmetric unit and the unit cell; C atoms are in grey, H in white, O in red and N in blue. The proton is shown as originally refined at 300 K, attached to the carboxylic acid. Close analysis of the data and further data collected at 100 K suggest that a disordered structure with the H atom occupying two sites is more representative.

enantiomers. X-ray diffraction data were initially collected at 300 K (Wheeler & Breen, 2016*a*). The two components crystallize in the monoclinic $P2_1/c$ space group, with the asymmetric unit and unit cell shown in Fig. 5. Both molecules in the structure adopt their expected conformation, with only a slight tilting of the NO_2 groups of the acid. The position of the H atom between the two co-formers was determined from a Fourier difference map, which shows that the proton is mostly located on the O atom, forming a co-crystal. Experimental data collected at 100 K after the blind test had concluded, show more clearly that the system is disordered with a two-site refinement suggesting the proton occupancy on the O atom is 0.58 (3) and that on the N atom is 0.42 (3) (Wheeler & Breen, 2016*b*). More variable-temperature studies and neutron diffraction may resolve whether the proton disorder is a dynamic, temperature-related effect. In a few experimental structures of Tröger's base derivatives, the N atoms appear to be clearly protonated, forming salts rather than co-crystals (see, for example, CSD refcodes: LEMBEL, CUNQAE), while neutral hydrogen bonds are observed in other structures such as PECDIM and PIPXAP.

In total, 14 attempted predictions were made for (XXV), with five groups generating the experimental structure and two re-ranking submissions also ranking the experimental structure within their list of 100 structures. All of these predicted a co-crystal, with no isostructural salt being found in any submissions. Once generated, (XXV) has generally been ranked as one of the most stable structures in the predicted landscape, with three predictions (van Eijck; Pantelides, Adjiman *et al.*; Price *et al.*) ranking it as being the most stable structure, and the worst rank being sixth. The re-ranking submissions of Brandenburg & Grimme, and Tkatchenko *et al.* ranked it as being the second-most or most stable structure, respectively.

The proton position in (XXV) is a significant challenge both for theory and experiment. As (XXV) was stated to be a co-crystal in the blind-test announcement, it is expected and understandable that no method explored the proton position explicitly, and for a number of methods the protonation state is fixed on the basis of the information given and cannot vary during the CSP calculation. Had the disorder been known in advance, it is likely that many methods would have been adapted as well, perhaps employing multiple searches with both neutral and charged co-formers and the potential parameters or 'typing' used for the N and O atoms would have been varied or explored, all of which could affect the results of the prediction (Mohamed *et al.*, 2011). Three methods (Facelli *et al.*; Neumann, Kendrick and Leusen; Zhu, Oganov, Masunov) did predict a non-isostructural salt form as being the most stable form for (XXV), although the latter two submissions do rank the experimental form as being one of the most stable structures. The prediction of a salt form for (XXV) is possible due to their use of DFT in the final ranking stage, which allows for proton migration and transfer to occur, although only if there is no barrier for this with the DFA used. Many of the other methods that use DFAs also predicted salt structures somewhere in their submitted lists.

While the disorder in (XXV) was an unexpected complication, it highlights the ongoing challenges of modelling proton positions and disorder. Salts and co-crystals are often considered distinct types of solid forms, but (XXV) also demonstrates the fine line between the two and the challenges of predicting or even characterizing them.

4.5. Target (XXVI)

N,*N*′-([1,1′-Binaphthalene]-2,2′-diyl)bis(2-chlorobenzamide) ($C_{34}H_{22}C_{12}N_2O_2$) was crystallized from a 1:1 mixture of hexane and dichloromethane in the triclinic $P\bar{1}$ space group, with data collected at room temperature (Wheeler & Hopkins, 2016). This crystal structure was the original target for this molecule and is referred to as form 1. Polymorph screening (Sharp *et al.*, 2016) found that form 1 undergoes a phase transition to another polymorph at around 428 K. This high-temperature polymorph is known as form 11 and has been characterized using high-resolution powder diffraction, with structure solution on-going (Sharp *et al.*, 2016). The polymorph screen also found nine solvates of (XXVI) (known as forms 2–10).

Compounds containing the 1,1′-binaphthalene fragment can feature axial chirality, however, no chiral precursors were used in the synthesis of (XXVI). While the category for this target stated that the experimental crystal structure was $Z' \leq 2$, the experimental structure for form 1 is $Z' = 1$, with one molecule in the asymmetric unit. In the crystal structure, shown in Fig. 6, the two molecules in the unit cell form an $R_2^2(18)$ dimer. There is also a close contact within the molecule between the Cl and an amide hydrogen on one of the two amide groups in the molecule. One of the two amide O atoms in the molecule is unsatisfied in terms of hydrogen bonds. As noted by a number of groups, the bulky binaphthalene and phenyl groups may well cause frustration in the molecular conformation, leading to difficulty in forming a more extensive intermolecular hydrogen-bond network, although intramolecular NH$\cdots$O hydrogen bonds might be observed. Comparing the experimental intramolecular geometry to CSD-derived angle and torsion distributions (using *Mogul*; Bruno *et al.*, 2004) suggests

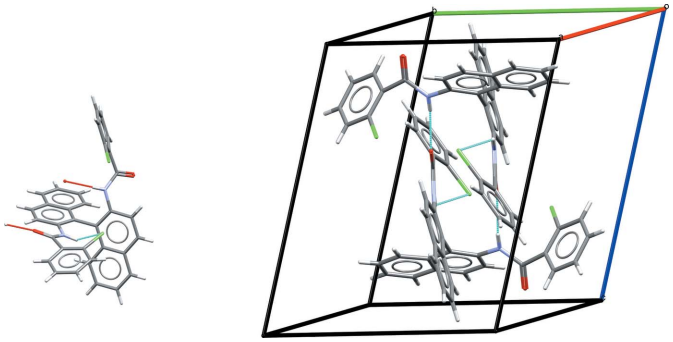


Figure 6
Experimental crystal structure of (XXVI), showing the molecular conformation and the unit cell, with hydrogen bonds shown by blue lines; C atoms are in grey, H in white, O in red, Cl in green and N in blue.

that the angle and torsions between the amide group and phenyl ring that are involved in both hydrogen bonds are unusual compared with expected CSD values.

There were 12 attempted predictions for molecule (XXVI), five of which explicitly considered the possibility of the experimental structure being $Z' = 2$. Three methods (Elking & Fusti-Molnar, Neumann, Kendrick and Leusen, and Price *et al.*) generated the experimental structure of form 1. All three submissions ranked form 1 as being the most stable polymorph in at least one of their two lists. For one submission (Elking & Fusti-Molnar), form 1 was ranked as number eight by an empirical potential, with DFT (PBE+XDM) improving the ranking to be number one in the second list. A comparison of the experimental structure of form 1 with the correction predictions is given in Table S8.

In many of the submissions, high-ranking structures (*e.g.* within the ten highest ranked predictions) do not feature intermolecular hydrogen bonds and conversely in some cases low packing coefficients are reported. This reflects the difficulty the molecule has in forming stable close-packed structures and intermolecular hydrogen bonds simultaneously and perhaps tallies with the preponderance of solvates in the experimental solid-form screen. For a number of methods, the failure to generate the form 1 structure can be attributed to difficulties in generating the experimental conformation due to its distorted nature. This posed a significant difficulty for searches employing rigid conformations, but even with flexibility permitted some methods would have needed more exhaustive searches to generate the correct conformation.

4.6. Computational resources

As in previous blind tests, participants were asked to include a brief summary of the computational resources and hardware used to carry out their predictions. Directly comparing these data is difficult not only due to the different CPUs used but also the wide range of architectures employed, ranging from standard desktop PCs to massively parallel machines at national supercomputing facilities. As a result the data have not been normalized. A summary of each submission's usage is provided in Table S9, with more details available in each submission's supporting-information document.

In general, the resources employed for predictions have increased significantly since the last blind test, with 13 submissions employing more than 100 000 CPU hours, compared to four in the fifth blind test. This is partly due to the increased use of more sophisticated ranking and refinement methods (such as DFT, tailor-made force fields and flexible multipoles) and partly due to more detailed and demanding searches of the conformational and structural landscapes of the targets, increasing the number of putative structures. A number of the full submissions that targeted all five systems employed over 500 000 CPU hours. For a single target, 100 000 CPU hours would amount to approximately 16 d elapsed time on a 256-core machine, representing a substantial investment of computational resources and time. Nevertheless, the increased importance and potential of computational modelling in general means that such computational resources are more widely available in both academia and industry, and further advancements and optimization in algorithms and software might well yield significant reductions in computational costs.

However, as in previous blind tests, there is a significant disparity in the amount of computational resources employed in obtaining a successful prediction. For (XXII), a number of successful predictions employed 10 000–30 000 CPU hours, while a few submissions predicted the known experimental structure with less than 200 CPU hours, using comparatively simple empirical potentials and, at most, rigid multipole electrostatics. Conversely, a number of full DFT/*ab initio* submissions for (XXII) failed to predict the experimental structure, despite using orders of magnitude more computational resources. A few methods generated some of the experimental structures of (XXIII) and (XXV) with a fraction of the CPU resources of other approaches and in some cases comparable ranking. This disparity suggests that there remains considerable scope to improve our understanding of where simple potentials are sufficient for some or all of the CSP calculation, where instead bespoke potentials and *ab initio* information and calculations must be used, and where optimizations and improvements in algorithms are possible.

As a final point, it is worth noting that as computational resources become more widely available and cheaper, the personnel cost of the methods becomes more important. This too likely varies significantly between the different methods and approaches to the problem. Whereas ranking is the most time-consuming process from a computational perspective, conformational analysis and interpretation of the CSP results are likely the most demanding parts of the calculation in terms of human resources.

4.7. Performance and progress of crystal structure prediction methods

The performance and 'success' of a CSP calculation is naturally first assessed in terms of whether experimental structures are generated by the calculation and where they are placed on the putative crystal-structure landscape. Generation relies on the experimental structure corresponding to a local minimum of the fitness function (or potential-energy surface) used. All the experimental structures in the sixth blind test, apart from the potentially disordered form *E* of (XXIII), were generated by one or more methods and submissions, with one method (Neumann, Kendrick and Leusen) generating all of them [apart from (XXIII) *E*].

While all of the structures have been generated, their ranking and placement on the predicted landscapes is more variable. (XXII), form *B* of (XXIII), (XXV) and (XXVI) were ranked as the lowest-energy, most-stable putative structure by a few methods but not consistently by a single method. This inconsistency may be explained, in part, by the possibility that some higher-ranked predicted structures might correspond to undiscovered experimental forms of (XXII), (XXIV) and

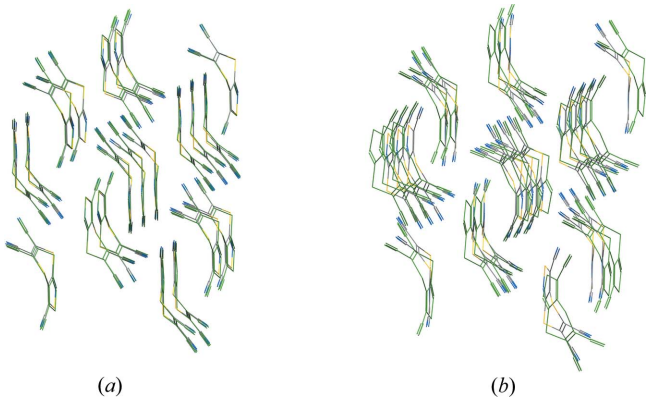
(XXV), which have not been subject to extensive solid-form screening.

The extent to which experimental structures have been reproduced in terms of the crystal structure is also variable. One measure of this is the RMSD between clusters from the experimental and predicted crystal structures, with example structure overlays for (XXII) shown in Fig. 7 (see Tables S1–S8 and §2.3 for more values and details, respectively). The values for this blind test are comparable to those in the previous one, although some are relatively large at ~ 0.8 Å. The RMSD value is often a combination of deviations in the gross packing and conformation, and therefore expected values may vary depending on the conformational flexibility of a molecule and the degree to which flexibility was permitted in the CSP calculation. In general, the smallest RMSD values are found for methods using DFAs for the final optimization and ranking step. However, it is worth remembering that experimental structures feature thermal-expansion effects, whereas the majority of the CSP methods are predicting 0 K 'equilibrium' geometries. MD simulations, which have been used by two submissions (Podeszwa *et al.* and Tuckerman, Szalewicz *et al.*), should capture these effects and provide better comparison with experiment. Such simulations require the temperature of the diffraction experiment as input though, which was not disclosed to participants. For (XXII), MD simulations at 300 K gave an $RMSD_{20}$ of 0.187 Å (Tuckerman, Szalewicz *et al.*), but a post-test MD simulation at the experimental temperature of 150 K, gives a value of 0.140 Å, which is smaller than any of the RMSD values for the submitted structures. This demonstrates the significant contribution of thermal and zero-point motion to RMSDs. Although zero-point motion would not be expected to influence ranking and RMSD values in molecules such as target (XXII), which contains all heavy atoms, in general, this is a factor that needs to be carefully considered.

To understand how the field has progressed and developed we can compare the sixth blind test with the previous fifth one. In that test the targets were generated and ranked within the top 100 structures between three and five times with typically 10–15 submissions (Bardwell *et al.*, 2011), leading to around 24 out of 68 predictions generating the experimental structure, although it should be noted that the criteria in the fifth blind test considered only the top-three predicted structures as a success and not all submissions provided extended lists of structures. In the present blind test, 36 predictions out of 70 (for $Z' = 1$ structures) generated the experimental structure. Some systems have been generated by a number of methods, *e.g.* 10 of 14 submissions generating or ranking (XXIII) form *B*, while only one method predicted the experimental structure of (XXIV) and none predicted (XXIII) *E*.

However, a key difference and development is the nature of the target molecules, which represent a significantly increased challenge. (XXIV) is the first three-component and salt–hydrate system, with both salts and hydrates having proven difficult individually in the previous blind test (Bardwell *et al.*, 2011). (XXVI) is the largest molecule attempted in a blind test to date, while the polymorphic nature of (XXIII), its intramolecular flexibility and two $Z' = 2$ forms makes it a serious challenge and test for methods as well, and (XXII) cannot be considered a strictly rigid molecule.

In this sense, the current blind test shows the advancement in the capabilities of CSP methods in the five years since the last test, and the broadening of their applicability to new types of solid forms and more complex molecules. While many challenges remain, as will be discussed below, the wide range of methods, many of them applied for the first time in this blind test, does bode well for the CSP in the future. There is a wealth of information in the submissions that points to new and continuing developments, as post-test analysis has already begun to show. Another important aspect of the development of CSP methods is the establishment of more well defined protocols and 'best practice' guidelines for performing the calculations, which will be further developed in light of the results of this blind test.



(*a*) (*b*)

Figure 7
Two example overlays of the experimental crystal structure of (XXII) with predicted structures of (*a*) Tkatchenko *et al.* with an RMSD of 0.166 Å, and (*b*) Obata & Goto with an RMSD of 0.808 Å. The predicted structures are shown in green for clarity. With the smaller RMSD in (*a*) the two structures are difficult to distinguish visually, while for the larger RMSD in (*b*) the predicted and experimental molecules are clearly offset.

4.8. Challenges in CSP methods

The sixth blind test highlights the continuing development of CSP methods but also the challenges they face. The first of these is in the initial generation of the experimental crystal structure. In many cases where methods failed this can be traced back to issues in generating the experimental conformation, either due to the search using rigid conformations significantly different from those in the experimentally observed forms or not considering a wide enough search space in flexible CSP calculations, which was seen in particular for (XXVI). In other cases, assumptions or limits placed on the search space or possible intermolecular interactions prevented the search from finding the observed crystal structure, or the search was simply not exhaustive enough. Experimental structures were initially generated by some search algorithms but not ranked highly by the intermediate optimization and ranking methods, and therefore not brought forward to the final stages where these missing structures could have ranked

highly. Encouragingly, post-test analysis has suggested a number of adjustments and refinements to different methods that should limit or prevent these issues in future.

The final, definitive ranking of the predicted structures remains a long-standing issue for CSP methods. The majority of methods based their final rankings on differences in static, 0 K lattice energies. DFT is emerging as a leading method for calculating these differences, being used by 12 CSP methods in this blind test. However, a number of benchmark studies of density-functional approximations and models of vdW interactions (Otero-de-la-Roza & Johnson, 2012; Reilly & Tkatchenko, 2013; Moellmann & Grimme, 2014) suggest accuracies of 3–4 kJ mol^{-1} for absolute lattice energies, while one of the most sophisticated quantum-chemical calculations of the lattice energy of benzene is accurate to only 1 kJ mol^{-1} (Yang *et al.*, 2014). Given the small energy differences needed to resolve some polymorphs (Price, 2009), accuracies of lattice-energy differences therefore may still involve fortuitous cancellation of errors, which is not assured with so many different types of interactions, conformations and packing arrangements possible. Post-test analysis of the (XXIII) polymorphs (Table S12) highlights the differences between ranking methods with a range of different relative orderings and absolute differences.

Many of the benchmark DFT studies have pointed towards ways of improving accuracy and transferability, including many-body vdW interactions (Risthaus & Grimme, 2013; Tkatchenko *et al.*, 2012) and the use of hybrid and *meta*-GGA density functionals (Reilly & Tkatchenko, 2013; Moellmann & Grimme, 2014). Affordable periodic quantum-chemical calculations are also emerging (Wen & Beran, 2011; Bygrave *et al.*, 2012; Del Ben *et al.*, 2012), and are already providing insights into polymorphism (Wen & Beran, 2012*a*,*b*; Bygrave *et al.*, 2012). The cost of *ab initio* calculations is a related issue for ranking, with less-intensive intermediate ranking methods still important for making CSP calculations tractable. The decline in the use of generic empirical potentials points to the need for better potentials to be developed or wider use of bespoke potentials based on first-principles methods, such as DFT (*e.g.* Neumann *et al.*, 2008; Grimme, 2014) or SAPT(DFT) (Misquitta *et al.*, 2005). Such intermediate methods may lead to more confidence in selecting the final set of structures for optimization and ranking with more expensive methods.

After considering static lattice energies, it is important to remember the contributions of vibrations, disorder and, if it is an experimental variable, pressure to the free-energy differences of crystal structures. Vibrational contributions can be readily estimated in the harmonic limit using lattice dynamics (Born & Huang, 1954; Dove, 1993), and have been used as part of a number of methods in this blind test and shown to affect rankings of a number of systems and the ordering of the polymorphs of (XXIII). However, such calculations neglect the contributions of anharmonic vibrations and thermal expansion, the role of which in polymorph free-energy differences is not well understood. Wider use of anharmonic lattice dynamics (Monserrat *et al.*, 2013) and MD simulations may shed more light on this. Configurational disorder can also be modelled, for example using ensemble approaches (Habgood *et al.*, 2011) or approaches based on Monte Carlo and substitution methods (Neumann *et al.*, 2015). However, the cost of all of these calculations is substantial, often more than an order of magnitude more than the initial geometry optimization (see the supporting-information documents of a number of submissions), making a fully consistent estimate of thermodynamic ordering very computationally demanding and challenging. Given the small energy differences observed between some low-energy structures in this blind test, it may become more important to include these contributions in future.

Beyond thermodynamics, there remains the fundamental role of kinetics in determining the experimentally observed or accessible solid forms (Threlfall, 2003; Blagden & Davey, 2003; Price, 2013). Some thermodynamically stable solid forms may be slow to nucleate, for example, due to the required molecular conformation being unstable in the crystallization solution, while metastable solid forms favoured by the fastest pathway to crystallization may be slow to revert to other forms. The similarities between some of the forms of (XXIII) and significant differences between others suggests that the balance between kinetics and thermodynamics might well be important for (XXIII). Only one method in the present blind test explicitly considered kinetics (using kinetic Monte Carlo simulations to determine critical-nucleus sizes), and no submission took account of the crystallization conditions supplied. There have been many advances in the modelling of nucleation (Anwar & Zahn, 2011) and crystal growth (Piana *et al.*, 2005; Salvalaglio *et al.*, 2012), but again these are involved and computationally demanding simulations, mostly limited, to date, to considering model systems, with relatively generic empirical potentials.

While direct modelling of kinetics is not routine, some CSP methods do involve considering differences between predicted structures, with the aim of rationalizing whether they would amount to distinct solid forms that would be expected to crystallize separately (Price, 2013, 2014). Structural informatics based on experimental crystal structures, such as hydrogen-bond propensities (Galek *et al.*, 2007, 2009), could also be used to assess the experimental likelihood of features in predicted structures. Approaches such as these may provide a bridge between the thermodynamic ranking produced by CSP calculations and the more demanding investigations of how kinetics affect the final solid form(s) of a molecule.

4.9. Beyond predicting 'the' crystal structure

While significant challenges remain for routine and definitive prediction of the stable solid forms of organic molecules, this is not always the true aim of performing CSP calculations, which are emerging as a general tool to complement experimental studies of organic solid forms. On a fundamental level, CSP calculations represent one of the most demanding challenges of the reliability of empirical potentials and first-prin-

ciples methods. Their role in providing information for solving or confirming crystal structures from powder X-ray diffraction data is now well established, and they can also aid alternative structure-characterization methods, such as NMR or electron diffraction (Baias *et al.*, 2013; Eddleston *et al.*, 2013).

CSP calculations also have a significant role in understanding the potential solid forms of a molecule. This has been demonstrated by a number of studies combining CSP calculations with experimental solid-form screening, as has already been noted in §1, and the sixth blind test further illustrates this. For (XXV), some of the submitted lists show large gaps in terms of energy between the lowest-energy structure and other putative structures. For other systems, the results show a range of structures close to the global minimum, which is more indicative of potential polymorphism. The experimental form of (XXII) was predicted by 12 out of 21 submissions, but a number of the other structures predicted as being low in energy were found in multiple submissions. While the exact predictions of the experimental structures are not always correct, these observations might help guide where more experiments, *e.g.* solid-form screening, are more likely to be needed. Indeed, it is worth remembering that the practical use of CSP calculations is unlikely to be 'blind', with either structures known experimentally or the difficulty of crystallization having been established. A CSP calculation that predicts many possible putative structures competitive with an experimentally observed form, as seen for all of the submissions for (XXIII), would suggest more experimental studies as being advisable.

Beyond guiding experiment directly, the landscapes or sets of putative crystal structures can inform on the general behaviour of a target molecule. For a number of submissions, low-energy predictions that do not match the experimental structures are nevertheless closely related to them, with a number of the unsuccessful submissions for (XXII) predicting structures that matched the experimental form with 14 out of 20 molecules. Such structures might well have similar properties to the observed solid form. In other cases, the submissions show how CSP enables one to explore the general ability of a molecule to pack with itself. A number of submissions for (XXVI) show the distorted nature of the molecular conformation and the difficulty the molecule has in forming extended hydrogen-bonding networks. Low packing coefficients are also reported, correlating well with the experimental observation of nine solvates in solid-form screening.

In the context of these wider applications of CSP methods, the 'success' of a CSP calculation can only be measured in terms of its specific goals and aims, which will rarely mean a completely blind prediction. These types of applications of CSP methods will require not only developments in the methods themselves but clear protocols for analysing the putative structures generated, as well as a greater understanding of how to turn information on possible or putative structures into new experiments and ultimately new solid forms. This will no doubt be one focus of ongoing research in CSP methods and future blind tests might well reflect this in the choice of target systems and goals.

5. Conclusion

The sixth blind test of organic CSP methods has been the biggest to date, with 21 submissions attempting to predict one or more of the five target systems, and four submissions re-ranking other predictions with different methods. The range of methods and approaches show the development of the field, with progress in the treatment of conformational flexibility in molecules, wider use of *ab initio* or *ab initio*-based methods for optimizing and ranking the final structures, as well as more well defined and systematic protocols for performing CSP calculations.

Apart from the potentially disordered form *E* of (XXIII), all of the experimental crystal structures of the five targets were predicted by one or more submissions, with one method based on Monte Carlo parallel tempering for structure generation and final ranking with DFT (Neumann, Kendrick and Leusen) generating all of them. While the rate of success is comparable to the previous blind test, the target systems are significantly more challenging, and include a polymorphic former drug candidate, a three-component chloride salt hydrate and a bulky flexible molecule that is the largest attempted in a blind test to date. In this context, we conclude that state-of-the-art CSP calculations are now applicable to a wider range of solid forms, such as salts and hydrates, as well as larger more flexible molecules.

However, significant challenges remain for routine and reliable CSP calculations. One source of difficulties in generating structures was the conformational flexibility and preferences of the targets. For (XXII), force fields and even some density-functional approximations had difficulty with the hinged nature of the molecule, while searches with rigid conformations had difficulties for (XXIII) and (XXVI). Encouragingly, post-test analysis of the results has already suggested a number of refinements to the CSP workflows used in the submissions.

The definitive ranking of the predicted crystal structures remains difficult and computationally expensive. While the experimental structures of many of the targets were ranked as being the most stable or one of the most stable predicted crystal structures, no method consistently ranked all of the experimental structures, as (XXIII) highlights. Post-test analysis again suggests that state-of-the-art density-functional approximations could improve upon the submitted results and ongoing developments in *ab initio* and DFT methods, algorithms and the use of bespoke force fields bode well. As ranking based on lattice energies improves, considering additional contributions such as entropy will be more important, with this blind test also seeing an increase in the number of submissions ranking structures based on Helmholtz free energies.

Overall, the results of this blind test have demonstrated the increased maturity of CSP methods. They have also illustrated the role for CSP calculations to guide and complement our

understanding and experimental studies of organic solid forms. This is likely to be an important focus for the application and development of CSP methods moving forward.

Acknowledgements

The organisers and participants are very grateful to the crystallographers who supplied the candidate structures: Dr Peter Horton (XXII), Dr Brian Samas (XXIII), Professor Bruce Foxman (XXIV) and Professor Kraig Wheeler [(XXV) and (XXVI)]. We are also grateful to Dr Emma Sharp and colleagues at Johnson Matthey (Pharmorphix) for the polymorph screening of (XXVI), as well as numerous colleagues at the CCDC for assistance in organizing the blind test.

Submission 2: We acknowledge Dr Oliver Korb for numerous useful discussions.

Submission 3: The Day group acknowledge the use of the IRIDIS High Performance Computing Facility, and associated support services at the University of Southampton, in the completion of this work. We acknowledge funding from the EPSRC (grants EP/J01110X/1 and EP/K018132/1) and the European Research Council under the European Union's Seventh Framework Programme (FP/2007-2013)/ERC through grant agreements No. 307358 (ERC-stG-2012-ANGLE) and No. 321156 (ERC-AG-PE5-ROBOT).

Submission 4: I am grateful to Mikhail Kuzminskii for calculations of molecular structures using the *GAUSSIAN*98 program in the Institute of Organic Chemistry RAS. The Russian Foundation for Basic Research is acknowledged for financial support (14-03-01091).

Submission 5: Toine Schreurs provided computer facilities and assistance. I am grateful to Matthew Habgood at AWE company for providing a travel grant.

Submission 6: We would like to acknowledge support of this work by GlaxoSmithKline, Merck, and Vertex.

Submission 7: The research was financially supported by the VIDI Research Program 700.10.427, which is financed by The Netherlands Organization for Scientific Research (NWO), and the European Research Council (ERC-2010-StG, grant agreement n. 259510-KISMOL). We acknowledge the support of the Foundation for Fundamental Research on Matter (FOM). Supercomputer facilities were provided by the National Computing Facilities Foundation (NCF).

Submission 8: Computer resources were provided by the Center for High Performance Computing at the University of Utah and the Extreme Science and Engineering Discovery Environment (XSEDE), supported by NSF grant number ACI-1053575. MBF and GIP acknowledge support from the University of Buenos Aires and the Argentinian Research Council.

Submission 9: We thank Dr Bouke van Eijck for his valuable advice on our predicted structure of (XXV). We thank the promotion office for TUT programs on advanced simulation engineering (ADSIM), the leading program for training brain information architects (BRAIN), and the information and media center (IMC) at Toyohashi University of Technology for the use of the TUT supercomputer systems and application software. We also thank the ACCMS at Kyoto University for the use of their supercomputer. In addition, we wish to thank financial support from Conflex Corp. and Ministry of Education, Culture, Sports, Science and Technology.

Submission 12: We thank Leslie Leiserowitz from the Weizmann Institute of Science and Geoffrey Hutchinson from the University of Pittsburgh for helpful discussions. We thank Adam Scovel at the Argonne Leadership Computing Facility (ALCF) for technical support. Work at Tulane University was funded by the Louisiana Board of Regents Award # LEQSF(2014-17)-RD-A-10 'Toward Crystal Engineering from First Principles', by the NSF award # EPS-1003897 'The Louisiana Alliance for Simulation-Guided Materials Applications (LA-SiGMA)', and by the Tulane Committee on Research Summer Fellowship. Work at the Technical University of Munich was supported by the Solar Technologies Go Hybrid initiative of the State of Bavaria, Germany. Computer time was provided by the Argonne Leadership Computing Facility (ALCF), which is supported by the Office of Science of the US Department of Energy under contract DE-AC02-06CH11357.

Submission 13: This work would not have been possible without funding from the College of Engineering at Khalifa University and I am grateful for the support of Professor Robert Bennell and Professor Bayan Sharif in facilitating the acquisition of all necessary resources. All of the theoretical data reported in this work were obtained using the High Performance Computing Cluster of Khalifa University and Dr Yacine Addad is acknowledged for providing systems support. Dr Louise S. Price is thanked for her guidance on the use of *DMACRYS* and *NEIGHCRYS* during the course of this research. She is also thanked for useful discussions and numerous email exchanges concerning the blind test. Professor Sarah L. Price is acknowledged for her support and guidance over many years and for providing access to *DMACRYS* and *NEIGHCRYS*.

Submission 15: The work was supported by the United Kingdom's Engineering and Physical Sciences Research Council (EPSRC) (EP/J003840/1, EP/J014958/1) and was made possible through access to computational resources and support from the High Performance Computing Cluster at Imperial College London. We are grateful to Professor Sarah L. Price for supplying the DMACRYS code for use within *CrystalOptimizer*, and to her and her research group for support with DMACRYS and feedback on *CrystalPredictor* and *CrystalOptimizer*.

Submission 16: RJN acknowledges financial support from the Engineering and Physical Sciences Research Council (EPSRC) of the UK [EP/J017639/1]. RJN and CJP acknowledge use of the Archer facilities of the UK's national high-performance computing service (for which access was obtained *via* the UKCP consortium [EP/K014560/1]). CJP also acknowledges a Leadership Fellowship Grant [EP/K013688/1]. BM acknowledges Robinson College,

Cambridge, and the Cambridge Philosophical Society for a Henslow Research Fellowship.

Submission 17: The work at the University of Delaware was supported by the Army Research Office under Grant W911NF-13-1-0387 and by the National Science Foundation Grant CHE-1152899. The work at the University of Silesia was supported by the Polish National Science Centre Grant No. DEC-2012/05/B/ST4/00086.

Submission 18: We would like to thank Constantinos Pantelides, Claire Adjiman and Isaac Sugden of Imperial College for their support of our use of *CrystalPredictor* and *CrystalOptimizer* in this and Submission 19. The CSP work of the group is supported by EPSRC, through grant ESPRC EP/K039229/1, and Eli Lilly. The PhD students support: RKH by a joint UCL Max-Planck Society Magdeburg Impact studentship, REW by a UCL Impact studentship; LI by the Cambridge Crystallographic Data Centre and the M3S Centre for Doctoral Training (EPSRC EP/G036675/1).

Submission 19: The potential generation work at the University of Delaware was supported by the Army Research Office under Grant W911NF-13-1-0387 and by the National Science Foundation Grant CHE-1152899.

Submission 20: The work at New York University was supported, in part, by the US Army Research Laboratory and the US Army Research Office under contract/grant number W911NF-13-1-0387 (MET and LV) and, in part, by the Materials Research Science and Engineering Center (MRSEC) program of the National Science Foundation under Award Number DMR-1420073 (MET and ES). The work at the University of Delaware was supported by the US Army Research Laboratory and the US Army Research Office under contract/grant number W911NF-13-1-0387 and by the National Science Foundation Grant CHE-1152899.

Submission 21: We thank the National Science Foundation (DMR-1231586), the Government of Russian Federation (Grant No. 14.A12.31.0003), the Foreign Talents Introduction and Academic Exchange Program (No. B08040) and the Russian Science Foundation, project No. 14-43-00052, base organization Photochemistry Center of the Russian Academy of Sciences. Calculations were performed on the Rurik supercomputer at Moscow Institute of Physics and Technology.

Submission 22: The computational results presented have been achieved in part using the Vienna Scientific Cluster (VSC).

Submission 24: The potential generation work at the University of Delaware was supported by the Army Research Office under Grant W911NF-13-1-0387 and by the National Science Foundation Grant CHE-1152899.

Submission 25: JH and AT acknowledge support from the Deutsche Forschungsgemeinschaft under the program DFG-SPP 1807. H-YK, RAD and RC acknowledge support from the Department of Energy (DOE) under Grant No. DE-SC0008626. This research used resources of the Argonne Leadership Computing Facility at Argonne National Laboratory, which is supported by the Office of Science of the US Department of Energy under Contract No. DE-AC02-06CH11357. This research used resources of the National Energy Research Scientific Computing Center, which is supported by the Office of Science of the US Department of Energy under Contract No. DE-AC02-05CH11231. Additional computational resources were provided by the Terascale Infrastructure for Groundbreaking Research in Science and Engineering (TIGRESS) High Performance Computing Center and Visualization Laboratory at Princeton University.

References

Adamo, C. & Barone, V. (1999). *J. Chem. Phys.* **110**, 6158–6170.
Ambrosetti, A., Reilly, A. M., DiStasio, R. A. Jr & Tkatchenko, A. (2014). *J. Chem. Phys.* **140**, 18A508.
Anwar, J. & Zahn, D. (2011). *Angew. Chem. Int. Ed.* **50**, 1996–2013.
Augelli-Szafran, C. *et al.* (2002). WO Patent App. PCT/US2000/015,071.
Axilrod, B. M. & Teller, E. (1943). *J. Chem. Phys.* **11**, 299–300.
Baias, M., Dumez, J.-N., Svensson, P. H., Schantz, S., Day, G. M. & Emsley, L. (2013). *J. Am. Chem. Soc.* **135**, 17501–17507.
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. (2011). *Acta Cryst.* B**67**, 535–551.
Becke, A. D. (1988). *Phys. Rev. A*, **38**, 3098–3100.
Becke, A. D. & Johnson, E. R. (2007). *J. Chem. Phys.* **127**, 154108.
Bhardwaj, R. M., Price, L. S., Price, S. L., Reutzel-Edens, S. M., Miller, G. J., Oswald, I. D. H., Johnston, B. F. & Florence, A. J. (2013). *Cryst. Growth Des.* **13**, 1602–1617.
Blagden, N. & Davey, R. J. (2003). *Cryst. Growth Des.* **3**, 873–885.
Boerrigter, S. X. M., Josten, G. P. H., van de Streek, J., Hollander, F. F. A., Los, J., Cuppen, H. M., Bennema, P. & Meekes, H. (2004). *J. Phys. Chem. A*, **108**, 5894–5902.
Born, M. & Huang, K. (1954). *Dynamical Theory of Crystal Lattices.* International Series of Monographs on Physics. Oxford: Clarendon Press.
Brandenburg, J. G. & Grimme, S. (2014). *Top. Curr. Chem.* **345**, 1–23.
Bruno, I. J., Cole, J. C., Kessler, M., Luo, J., Motherwell, W. D. S., Purkis, L. H., Smith, B. R., Taylor, R., Cooper, R. I., Harris, S. E. & Orpen, A. G. (2004). *J. Chem. Inf. Comput. Sci.* **44**, 2133–2144.
Bygrave, P. J., Allan, N. L. & Manby, F. R. (2012). *J. Chem. Phys.* **137**, 164102.
Case, D. H., Campbell, J. E., Bygrave, P. J. & Day, G. M. (2016). *J. Chem. Theory Comput.* **12**, 910–924.
Chisholm, J. A. & Motherwell, S. (2005). *J. Appl. Cryst.* **38**, 228–231.
Cohen, A. J., Mori-Sánchez, P. & Yang, W. (2008). *Science*, **321**, 792–794.
Cohen, A. J., Mori-Sánchez, P. & Yang, W. (2012). *Chem. Rev.* **112**, 289–320.
Cruz-Cabeza, A. J. & Bernstein, J. (2014). *Chem. Rev.* **114**, 2170–2191.
Cruz-Cabeza, A. J., Liebeschuetz, J. W. & Allen, F. H. (2012). *CrystEngComm*, **14**, 6797–6811.
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. (2009). *Acta Cryst.* B**65**, 107–125.
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. (2005). *Acta Cryst.* B**61**, 511–527.
Deij, M. A., ter Horst, J. H., Meekes, H., Jansens, P. & Vlieg, E. (2007). *J. Phys. Chem. B*, **111**, 1523–1530.
Del Ben, M., Hutter, J. & VandeVondele, J. (2012). *J. Chem. Theory Comput.* **8**, 4177–4188.
Della Valle, R. G., Venuti, E., Brillante, A. & Girlando, A. (2008). *J. Phys. Chem. A*, **112**, 6715–6722.
Dion, M., Rydberg, H., Schröder, E., Langreth, D. C. & Lundqvist, B. I. (2004). *Phys. Rev. Lett.* **92**, 246401.
Dove, M. (1993). *Introduction to Lattice Dynamics.* Cambridge Topics in *Mineral Physics and Chemistry*. Cambridge University Press.
Eddleston, M. D., Hejczyk, K. E., Bithell, E. G., Day, G. M. & Jones, W. (2013). *Chem. Eur. J.* **19**, 7874–7882.
Eijck, B. P. van, Mooij, W. T. M. & Kroon, J. (2001*a*). *J. Comput. Chem.* **22**, 805–815.
Eijck, B. P. van, Mooij, W. T. M. & Kroon, J. (2001*b*). *J. Phys. Chem. B*, **105**, 10573–10578.
Etter, M. C., MacDonald, J. C. & Bernstein, J. (1990). *Acta Cryst.* B**46**, 256–262.
Foxman, B. M. (2016). CSD Communication: CCDC 1447530. doi: 10.5517/cc1kl8jy.
Galek, P. T. A., Allen, F. H., Fábián, L. & Feeder, N. (2009). *CrystEngComm*, **11**, 2634–2639.
Galek, P. T. A., Fábián, L., Motherwell, W. D. S., Allen, F. H. & Feeder, N. (2007). *Acta Cryst.* B**63**, 768–782.
Gelder, R. de, Wehrens, R. & Hageman, J. A. (2001). *J. Comput. Chem.* **22**, 273–289.
Goto, H. & Osawa, E. (1989). *J. Am. Chem. Soc.* **111**, 8950–8951.
Goto, H. & Osawa, E. (1993). *J. Chem. Soc. Perkin Trans. 2*, pp. 187–198.
Grimme, S. (2006). *J. Comput. Chem.* **27**, 1787–1799.
Grimme, S. (2014). *J. Chem. Theory Comput.* **10**, 4497–4514.
Grimme, S., Antony, J., Ehrlich, S. & Krieg, H. (2010). *J. Chem. Phys.* **132**, 154104.
Groom, C. R., Bruno, I. J., Lightfoot, M. P. & Ward, S. C. (2016). *Acta Cryst.* B**72**, 171–179.
Groom, C. R. & Reilly, A. M. (2014). *Acta Cryst.* B**70**, 776–777.
Habgood, M., Grau-Crespo, R. & Price, S. L. (2011). *Phys. Chem. Chem. Phys.* **13**, 9590–9600.
Habgood, M., Sugden, I., Kazantsev, A. V., Adjiman, C. S. & Pantelides, C. C. (2015). *J. Chem. Theory Comput.* **11**, 1957–1969.
Hofmann, D. W. M. & Kuleshova, L. (2005). *J. Appl. Cryst.* **38**, 861–866.
Horton, P. N. & Gossel, M. C. (2016). CSD Communication: CCDC 1451239. doi: 10.5517/cc1kq45l.
Ismail, S. Z., Anderton, C. L., Copley, R. C. B., Price, L. S. & Price, S. L. (2013). *Cryst. Growth Des.* **13**, 2396–2406.
Kazantsev, A. V., Karamertzanis, P. G., Adjiman, C. S. & Pantelides, C. C. (2011). *J. Chem. Theory Comput.* **7**, 1998–2016.
Klimeš, J., Bowler, D. R. & Michaelides, A. (2011). *Phys. Rev. B*, **83**, 195131.
Klimeš, J. & Michaelides, A. (2012). *J. Chem. Phys.* **137**, 120901.
Kolossváry, I. & Guida, W. C. (1996). *J. Am. Chem. Soc.* **118**, 5011–5019.
Kronik, L. & Tkatchenko, A. (2014). *Acc. Chem. Res.* **47**, 3208–3216.
Kuleshova, L., Hofmann, D. & Boese, R. (2013). *Chem. Phys. Lett.* **564**, 26–32.
Lee, C., Yang, W. & Parr, R. G. (1988). *Phys. Rev. B*, **37**, 785–789.
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., van Eijck, B. P., Verwer, P. & Williams, D. E. (2000). *Acta Cryst.* B**56**, 697–714.
Lund, A. M., Pagola, G. I., Orendt, A. M., Ferraro, M. B. & Facelli, J. C. (2015). *Chem. Phys. Lett.* **626**, 20–24.
Macrae, C. F., Bruno, I. J., Chisholm, J. A., Edgington, P. R., McCabe, P., Pidcock, E., Rodriguez-Monge, L., Taylor, R., van de Streek, J. & Wood, P. A. (2008). *J. Appl. Cryst.* **41**, 466–470.
Marom, N., DiStasio, R. A. Jr, Atalla, V., Levchenko, S., Reilly, A. M., Chelikowsky, J. R., Leiserowitz, L. & Tkatchenko, A. (2013). *Angew. Chem. Int. Ed.* **52**, 6629–6632.
Misquitta, A. J., Podeszwa, R., Jeziorski, B. & Szalewicz, K. (2005). *J. Chem. Phys.* **123**, 214103.
Moellmann, J. & Grimme, S. (2014). *J. Phys. Chem. C*, **118**, 7615–7621.
Mohamed, S., Tocher, D. A. & Price, S. L. (2011). *Int. J. Pharm.* **418**, 187–198.
Monserrat, B., Drummond, N. D. & Needs, R. J. (2013). *Phys. Rev. B*, **87**, 144302.
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., van Eijck, B. P., Verwer, P. & Williams, D. E. (2002). *Acta Cryst.* B**58**, 647–661.
Neumann, M. A. (2008). *J. Phys. Chem. B*, **112**, 9810–9829.
Neumann, M. A., Leusen, F. J. J. & Kendrick, J. (2008). *Angew. Chem. Int. Ed.* **47**, 2427–2430.
Neumann, M. A. & Perrin, M.-A. (2005). *J. Phys. Chem. B*, **109**, 15531–15541.
Neumann, M. A., van de Streek, J., Fabbiani, F. P. A., Hidber, P. & Grassmann, O. (2015). *Nat. Commun.* **6**, 7793.
Nyman, J. & Day, G. M. (2015). *CrystEngComm*, **17**, 5154–5165.
Obata, S. & Goto, H. (2015). *AIP Conf. Proc.* **1649**, 130–134.
Oganov, A. R. & Valle, M. (2009). *J. Chem. Phys.* **130**, 104504.
Ong, S. P., Richards, W. D., Jain, A., Hautier, G., Kocher, M., Cholia, S., Gunter, D., Chevrier, V. L., Persson, K. A. & Ceder, G. (2013). *Comput. Mater. Sci.* **68**, 314–319.
Otero-de-la-Roza, A. & Johnson, E. R. (2012). *J. Chem. Phys.* **136**, 174109.
Perdew, J. P., Burke, K. & Ernzerhof, M. (1996). *Phys. Rev. Lett.* **77**, 3865–3868.
Piana, S., Reyhani, M. & Gale, J. D. (2005). *Nature*, **438**, 70–73.
Price, S. L. (2009). *Acc. Chem. Res.* **42**, 117–126.
Price, S. L. (2013). *Acta Cryst.* B**69**, 313–328.
Price, S. L. (2014). *Chem. Soc. Rev.* **43**, 2098–2111.
Price, S. L., Leslie, M., Welch, G. W. A., Habgood, M., Price, L. S., Karamertzanis, P. G. & Day, G. M. (2010). *Phys. Chem. Chem. Phys.* **12**, 8478–8490.
Pyzer-Knapp, E. O., Thompson, H. P. G., Schiffmann, F., Jelfs, K. E., Chong, S. Y., Little, M. A., Cooper, A. I. & Day, G. M. (2014). *Chem. Sci.* **5**, 2235–2245.
Reilly, A. M. & Tkatchenko, A. (2013). *J. Chem. Phys.* **139**, 024705.
Reilly, A. M. & Tkatchenko, A. (2014). *Phys. Rev. Lett.* **113**, 055701.
Reilly, A. M. & Tkatchenko, A. (2015). *Chem. Sci.* **6**, 3289–3301.
Risthaus, T. & Grimme, S. (2013). *J. Chem. Theory Comput.* **9**, 1580–1591.
Salvalaglio, M., Vetter, T., Giberti, F., Mazzotti, M. & Parrinello, M. (2012). *J. Am. Chem. Soc.* **134**, 17221–17233.
Samas, B. (2015). Personal communication.
Samas, B. (2016*a*). CSD Communication: CCDC 1447522, doi: 10.5517/cc1kl88p.
Samas, B. (2016*b*). CSD Communication: CCDC 1447523, doi: 10.5517/cc1kl89q.
Samas, B. (2016*c*). CSD Communication: CCDC 1447524, doi: 10.5517/cc1kl8br.
Samas, B. (2016*d*). CSD Communication: CCDC 1447525. DOI: 10.5517/cc1kl8cs.
Samas, B. (2016*e*). CSD Communication: CCDC 1447526, doi: 10.5517/cc1kl8dt.
Sharp, E. *et al.* (2016). In preparation.

Simons, L. J., Caprathe, B. W., Callahan, M., Graham, J. M., Kimura, T., Lai, Y., LeVine, H. III, Lipinski, W., Sakkab, A. T., Tasaki, Y., Walker, L. C., Yasunaga, T., Ye, Y., Zhuang, N. & Augelli-Szafran, C. E. (2009). *Bioorg. Med. Chem. Lett.* **19**, 654–657.
Sobol', I. (1967). *USSR Comput. Math. Math. Phys.* **7**, 86–112.
Spek, A. L. (2009). *Acta Cryst.* D**65**, 148–155.
Spglib (2015). http://atztogo.github.io/spglib/.
Stokes, H. T. & Hatch, D. M. (2005). *J. Appl. Cryst.* **38**, 237–238.
Stone, A. J. (2005). *J. Chem. Theory Comput.* **1**, 1128–1132.
Stone, A. J. (2013). *The Theory of Intermolecular Forces*, 2nd ed. Oxford University Press.
Tao, J., Perdew, J. P., Staroverov, V. N. & Scuseria, G. E. (2003). *Phys. Rev. Lett.* **91**, 146401.
Thompson, H. P. G. & Day, G. M. (2014). *Chem. Sci.* **5**, 3173–3182.
Threlfall, T. (2003). *Org. Process Res. Dev.* **7**, 1017–1027.
Tkatchenko, A., DiStasio, R. A. Jr, Car, R. & Scheffler, M. (2012). *Phys. Rev. Lett.* **108**, 236402.
Tkatchenko, A. & Scheffler, M. (2009). *Phys. Rev. Lett.* **102**, 073005.
Verwer, P. & Leusen, F. J. J. (1998). *Computer Simulation to Predict Possible Crystal Polymorphs*, Vol. 12, *Reviews in Computational Chemistry*, pp. 327–365. New York: John Wiley and Sons, Inc.
Wang, Y., Lv, J., Zhu, L. & Ma, Y. (2012). *Comput. Phys. Commun.* **183**, 2063–2070.
Wen, S. & Beran, G. J. O. (2011). *J. Chem. Theory Comput.* **7**, 3733–3742.
Wen, S. & Beran, G. J. O. (2012*a*). *Cryst. Growth Des.* **12**, 2169–2172.
Wen, S. & Beran, G. J. O. (2012*b*). *J. Chem. Theory Comput.* **8**, 2698–2705.
Wheeler, K. A. & Breen, M. E. (2016*a*). CSD Communication: CCDC 1447527, doi: 10.5517/cc1kl8fv.
Wheeler, K. A. & Breen, M. E. (2016*b*). CSD Communication: CCDC 1447528, doi: 10.5517/cc1kl8gw.
Wheeler, K. A. & Hopkins, G. W. (2016). CSD Communication: CCDC 1447529, doi: 10.5517/cc1kl8hx.
Yang, J., Hu, W., Usvyat, D., Matthews, D., Schütz, M. & Chan, G. K.-L. (2014). *Science*, **345**, 640–643.
Yu, T.-Q. & Tuckerman, M. E. (2011). *Phys. Rev. Lett.* **107**, 015701.
Zhu, Q., Oganov, A. R., Glass, C. W. & Stokes, H. T. (2012). *Acta Cryst.* B**68**, 215–226.

GAtor
Supplementary Information

Farren Curtis[1,2], Xiayue Li[1,5], Christoph Schober[6], Katherine Cosburn[1,7], Sanjaya Lohani[1], Francesca Vacarro[1,8], Harald Oberhofer[6], Karsten Reuter[6], Saswata Bhattacharya[4], Álvaro Vázquez-Mayagoitia[5], Luca M. Ghiringhellli[4], and Noa Marom[*1,3].

[1]*Department of Physics and Engineering Physics, Tulane University, New Orleans, Louisiana 70118, USA.* [2]*Department of Physics, Carnegie Mellon University, Pittsburgh, PA 15213, USA.* [3]*Department of Materials Science and Engineering and Department of Physics, Carnegie Mellon University, Pittsburgh, PA 15213, USA.* [4]*Fritz-Haber-Institut der Max-Planck-Gesellschaft, Faradayweg 4-6, 14195, Berlin, Germany.* [5]*Argonne Leadership Computing Facility, Argonne National Laboratory, Lemont, Illinois 60439, USA.* [6]*Chair for Theoretical Chemistry and Catalysis Research Center, Technische Universität München, Lichtenbergstr. 4, D-85747 Garching, Germany.* [7]*Department of Physics, University of Toronto, Toronto, M5S 1A7, Canada.* [8]*Department of Chemistry, Loyola University, New Orleans, Louisiana 70118, USA.*

1 Overview of Methodology

We developed a first-principles genetic algorithm (GA) for molecular crystal structure prediction, GAtor, which finds the most energetically stable crystal structures for rigid molecules, and applied it to target XXII. GAtor uses principles from evolutionary theory such as survival of the fittest, crossover, and mutation that are implemented as operators acting on individual molecules and/or lattice vectors of the fittest crystal structures selected for mating. Structural relaxations and energy evaluations are performed using dispersion-inclusive density functional theory (DFT). For this purpose, GAtor interfaces with the all-electron electronic structure package FHI-aims [1]. Massive parallelization is achieved by eliminating the concept of GA generations and running several replicas in parallel that read and write to a common pool of structures [2, 3, 4]. Crucial to the success of GAtor is the generation and ranking of an initial pool of structures that maintains diversity and explores the most physically-relevant and promising regions of the potential energy landscape. Below we provide brief descriptions of our initial pool generation as well as the implementation of the genetic algorithm itself. We also explain the methodology we used for post-processing, refinement, and re-ranking of the final structures produced by the genetic algorithm. It should be noted that our approach is a fully quantum mechanical first-principles approach that does not use force fields at any point.

Single Molecule Structure

Since target XXII can adopt different conformations (i.e. the six membered ring can be bent along the S-S axis, such that the CN groups may lie above or below the plane of the 5 membered ring) we analyzed the effect of the bending angle on the energy of the single molecule. First, we performed single-point calculations on a range of angles from 20° to 160° using the Perdew-Burke-Ernzerhof generalized gradient approximation (PBE)[5, 6] with the Tkatchenko-Scheffler (TS)

*Corresponding author: nmarom@andrew.cmu.edu

pairwise dispersion-correction, PBE+TS [7], which employs an inexpensive pairwise approach to account for the van der Waals (vdW) contribution to the energy. We obtained a symmetric double-well potential and fully relaxed the two stable conformations to obtain final bending angles of 155.8° and 204.2°, respectively. We used the rigid, single molecule geometries at these equilibirum angles for the initial pool generation.

Initial Pool Generation

Since target XXII possessed two enantiomers, we generated separate chiral and racemic initial pools to serve as inputs for our genetic algorithm. In total we generated four different pools— each having 2 or 4 molecules per unit cell that contained one or both conformers. For each of these pools, 50,000 structures were generated both in random unit cells (with no enforced symmetry between the molecules) as well as in the most likely space groups. The random, symmetric structures were generated in the $P2_1$ and $P2$ space groups for the Z=2 chiral pool, $P\bar{1}$, Pc, and Pm for the Z=2 racemic pool, $P2_12_12_1$, $P2_12_12$, and $C2$ for the Z=4 chiral pool, and $P2_1/c$, $Pca2_1$, and $Pna2_1$ for the Z=4 racemic pool. The volume range we used for generating the structures was 160-300 $Å^3$/molecule.

The four initial pools were independently ranked in energy using an implementation of the Harris approximation integrated with FHI-aims. Within the Harris approximation, the total density of a system is constructed by a superposition of fragment densities [8]. In this scenario, the DFT total energy can be evaluated for the Harris density without performing a self-consistent cycle, allowing almost instantaneous energy evaluations.

The Harris density of a molecular crystal is constructed by replicating, translating, and rotating a single molecule's density, which is calculated only once. The numerical atom-centered orbital (NAO) basis functions of FHI-aims are based on real-valued linear combinations of spherical harmonics [1]. Since the spherical harmonics are fixed with respect to the *xyz*-coordinate system, rotation of a molecule produces a new linear combination of basis functions. Modified Wigner matrices [9] are employed to obtain the rotated coefficients of each basis function.

A binding energy curve computed with PBE+TS and PBE+TS@Harris for a molecule similar to target XXII, tetracyano-1,4-dithiin, is shown in Fig.1. The curve was calculated along the direction of the closest C$\cdots$N contacts in the crystal [10] which are similar to the close contacts of target XXII. When the molecules are non-interacting at large distances the Harris approximation and the fully self-consistent method converge to the same result. As the molecules come closer together, the Harris approximation fails to account for the change in density due to the electrostatic interactions between the molecules and polarization effects. This produces a weaker binding energy than the fully self-consistent result. The difference between the self-consistent and Harris densities for the tetracyano-1,4-dithiin dimer is also shown in Fig. 1. The red regions around the N atoms indicate areas where the Harris density overestimates the self-consistent density while the blue regions around the S atoms indicate where the Harris approximation underestimates the self-consistent density. The self-consistent density is concentrated closer to the molecular framework because of Coulomb repulsion between nitrogen lone-pairs.

The Harris approximation allows GAtor to perform fast screening of initial structures using an unbiased first-principles approach without resorting to force fields, which can be difficult to parameterize for non-standard molecules. We used PBE+TS for performing the Harris approximation for target XXII. The parameters of the TS correction (i.e. the C_6 coefficients and the van der Waals radii) are calculated on the fly based on the DFT (or Harris) density. This makes the TS method more accurate than semi-empirical pairwise methods [11].

After the initial structures in each of the four pools were ranked with the Harris approximation, local geometry optimization was performed for the best 6%, using PBE+TS with *lower-level*

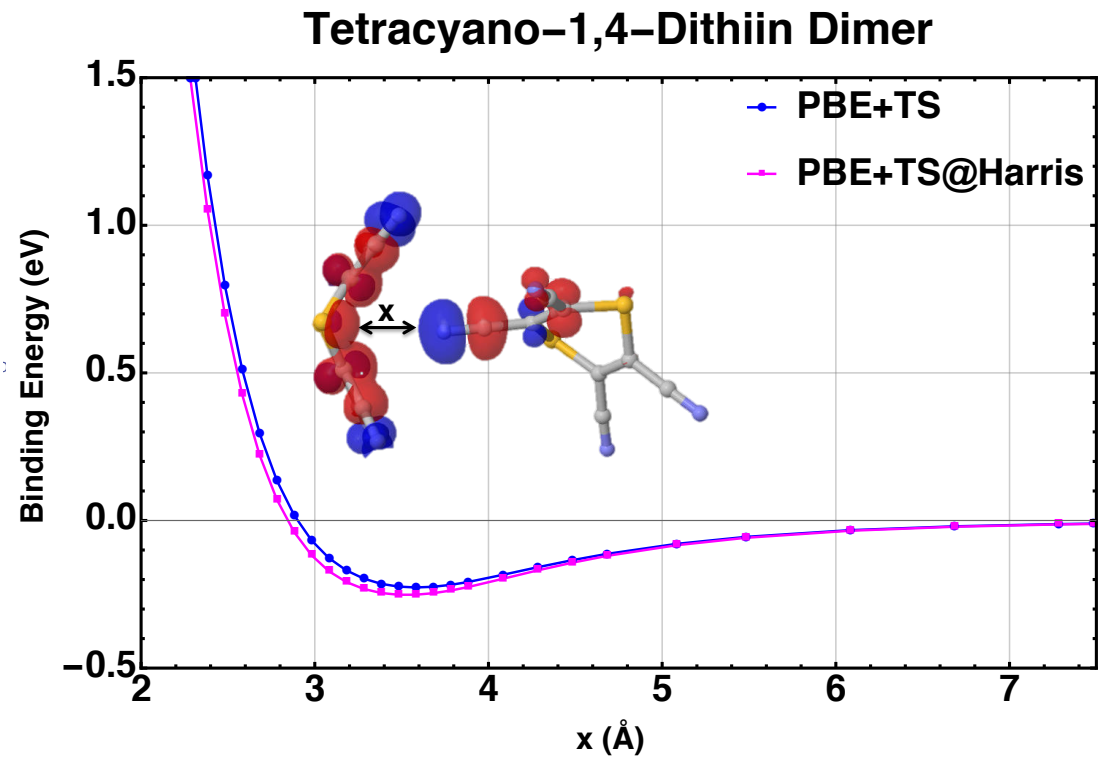


Figure 1: A binding energy curve for tetracyano-1,4-dithiin computed with PBE+TS and PBE+TS@Harris. An isosurface of the density difference between the self-consistent and Harris densities is shown for the equilibrium distance of $x = 3.48$ Å. The red regions indicate areas of a positive density difference while the blue areas indicate regions with a negetive density difference.

numerical settings which correspond to the light/tier1 settings of FHI-aims [1]. For the best 50% of this subset, full unit cell relaxations were performed using PBE+TS and *lower-level* numerical settings. During relaxation no constraints were imposed on the unit cell symmetry or the structure of the molecule (i.e. the central bending angle was allowed to vary for different structures). The fully relaxed structures served as the initial pool for the genetic algorithm.

The GAtor Genetic Algorithm

Similar to the genetic algorithms reported in [2, 3, 4, 12], the fitness f_i of each structure depends on its normalized relative energy given by:

$$f_i = \frac{\epsilon_i}{\sum_i \epsilon_i} \tag{1}$$

where ϵ_i is the relative energy of the i^{th} structure and is given by:

$$\epsilon_i = \frac{E_{max} - E_i}{E_{max} - E_{min}}. \tag{2}$$

In Eq. 2, E_{max} is the current maximum energy in the pool and E_{min} is the minimum. Hence, the fitness for each individual is dynamically updated with each new addition to the common pool. Using a "roulette-wheel" selection criterion [13], structures with higher fitness values have a higher probability of selection. A small fitness reversal probability similar to [3] allows for the occasional selection of an unfit structure to avoid biasing the search towards pre-converging to local minima.

Crossover randomly combines the lattice vectors of each parent structure and randomly selects molecules from one or both parents. After crossover, child structures have a 15% chance of un-

dergoing mutation, which applies random displacements or rotations to the molecules in the child structure or random symmetric or asymmetric strains to the unit cell.

After the child structure has been formed, its single-point energy is computed using PBE+TS with *lower-level settings*. If the single-point energy of the trial structure is outside of a user-specified range from the current global minimum, it is immediately rejected. The structure is also rejected if it is found to be a duplicate of any other structure in the common pool. If the structure is found to be unique then it is passed on to full unit cell relaxation with *lower-level* settings and added to the common pool. GAtor stops when it can no longer find new low-energy structures. The top 10% of the most stable structures found by the GA are then selected for post-processing.

Post-Processing

After combining the top structures found in each of our searches, approximately 500 structures were re-relaxed and re-ranked using PBE+TS and *higher-level* numerical settings, which correspond to the tight/tier 2 settings of FHI-aims [1]. We then performed single-point energy evaluations using PBE with the many-body dispersion (MBD)[14, 15] method for 200 of the best structures as ranked by PBE+TS. The MBD method accounts for long-range electrostatic screening effects and for non-pairwise-additive many-body contributions to the dispersion energy. It has been shown to accurately rank the stability of molecular crystal polymorphs in cases where the pairwise TS approach is not sufficiently accurate [16, 17].

We also performed single-point energy evaluations for 150 of the best structures as ranked by PBE+MBD using the hybrid functional PBE0 [18, 19] with the MBD correction. The inclusion of 25% exact exchange in PBE0 mitigates the self-interaction error, leading to a more accurate description of electron densities and multipoles [17, 20, 21]. For some molecular crystals, such as glycine and oxalic acid, the correct polymorph ranking is reproduced only when using PBE0+MBD [16]. We therefore consider the ranking of PBE0+MBD to be the most reliable of the methods used here.

2 Results and Analysis

Our blind test submission included one list of the top 100 structures as ranked by PBE+TS and another as ranked by PBE+MBD. Since we performed single-point calculations on more than 100 structures from the GA, the two lists did not consist of a simple re-ranking of the exact same structures. The PBE0+MBD calculations were not completed in time for the submission deadline and the full PBE0+MBD list was appended after the submission.

The final top 10 structures as ranked by PBE+MBD and re-ranked by PBE+TS and PBE0+MBD are shown in Table 1 along with the experimental structure which was not found in our searches. The experimental structure would have been ranked in the top three by all three methods and as #1 by PBE0+MBD. Further analysis of the effect of the choice of DFT functional and dispersion method on the ranking of structures is provided below.

Target XXII crystallized in $P2_1/n$, a nonstandard spacegroup used for orthogonal representations of oblique $P2_1/c$ unit cells. We did not explicitly generate structures in $P2_1/n$ because the structures generated in $P2_1/c$ for the initial pool were constrained to have angles between 60 to 120 degrees. Although the GA in principle still could have found the target by various strain and/or rotation mutations, it was biased for selecting, crossing over, and propagating the traits (including the orientation of the structural motifs) of the best structures in the initial pool. Furthermore, our post-analysis revealed that the duplicate check tolerance throughout the GA was set too tight. This allowed some duplicate structures into the pool, leading to an artificial increase in the representation of orthogonal cells. Overall, generating intial pool structures explicitly in $P2_1/n$ would have greatly increased the likelihood of finding the experimental structure.

Name	PBE+MBD		PBE+TS		PBE0+MBD		Z	Space Group	a	b	c	α	β	γ
	Rank	ΔE(eV)	Rank	ΔE(eV)	Rank	ΔE(eV)								
7471226271	1	0.000	1	0.000	2	0.007	4	$Pna2_1$	13.4	10.2	7.1	90	90	90
9f774c9e27	2	0.005	2	0.018	1	0.000	4	$P2_1/c$	14.4	10.3	6.7	90	90	94
dab6897b90	3	0.021	3	0.028	3	0.033	4	$P2_12_12_1$	10.2	7.0	13.7	90	90	90
52cdef12ff	4	0.037	4	0.032	30	0.055	4	$P\bar{1}$	14.0	7.1	10.3	90	90	70
42a9600b47	5	0.038	79	0.077	4	0.036	4	$Pna2_1$	20.5	7.4	6.7	90	90	90
197ac7c454	6	0.040	94	0.081	11	0.045	4	$P2_1/c$	9.7	11.3	9.4	90	85	90
f191f2a68b	7	0.040	6	0.044	10	0.044	4	$Pnma$	20.7	7.1	6.6	90	90	90
585d18ed08	8	0.041	9	0.051	18	0.051	2	$P\bar{1}$	9.0	9.8	6.0	110	93	98
71fe1a6200	9	0.043	20	0.059	31	0.056	4	$P2_1$	10.2	7.3	13.5	90	90	83
a206286cd3	10	0.044	67	0.075	6	0.039	2	$P2_1$	10.3	7.5	6.6	90	90	88
Experimental	(3)	0.006	(2)	0.017	(1)	-0.005	4	$P2_1/n$	12.0	6.7	12.6	90	109	90

Table 1: The top 10 structures as ranked by PBE+MBD with the re-ranking of PBE+TS and PBE0+MBD. The experimental structure which was not found in our search is shown with its hypothetical ranking and relative energy to the respective global minimum of the submitted structures.

We did, however, generate several structures with similar binding motifs to the blind test target. Structure 9f774c9e27, in space group $P2_1/c$, is compared to the experimental structure of target XXII in Fig. 2. These structures are stabilized by similar intramolecular C$\cdots$N interactions.

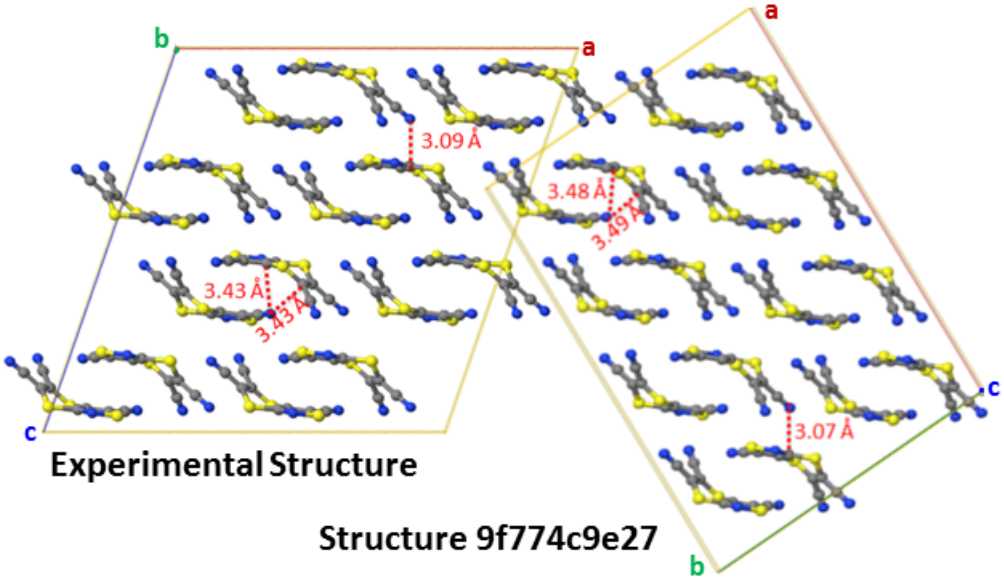


Figure 2: Motif comparisons between target XXII and our PBE+TS and PBE+MBD #2 structure. Distances of the closest intermolecular C$\cdots$N interactions are shown.

The energy differences between target XXII and 9f774c9e27 are extremely small, ranging from 1 meV to 5 meV depending on the method, as shown in Table 1. Furthermore, several other groups submitted this structure within their top 10 structures. Another structure, 7471226271, which was also in the top 2 for all ranking methods, was listed among the top structures of several other groups as well. Since structures 9f774c9e27, 7471226271, and the experimental structure are all extremely close in energy and were found by several other groups, we believe these three structures may be polymorphs.

The effect of the choice of DFT functional and dispersion method on the potential energy landscape and the ranking of structures is illustrated in Figs. 3 and 4. Graphs of the volume per

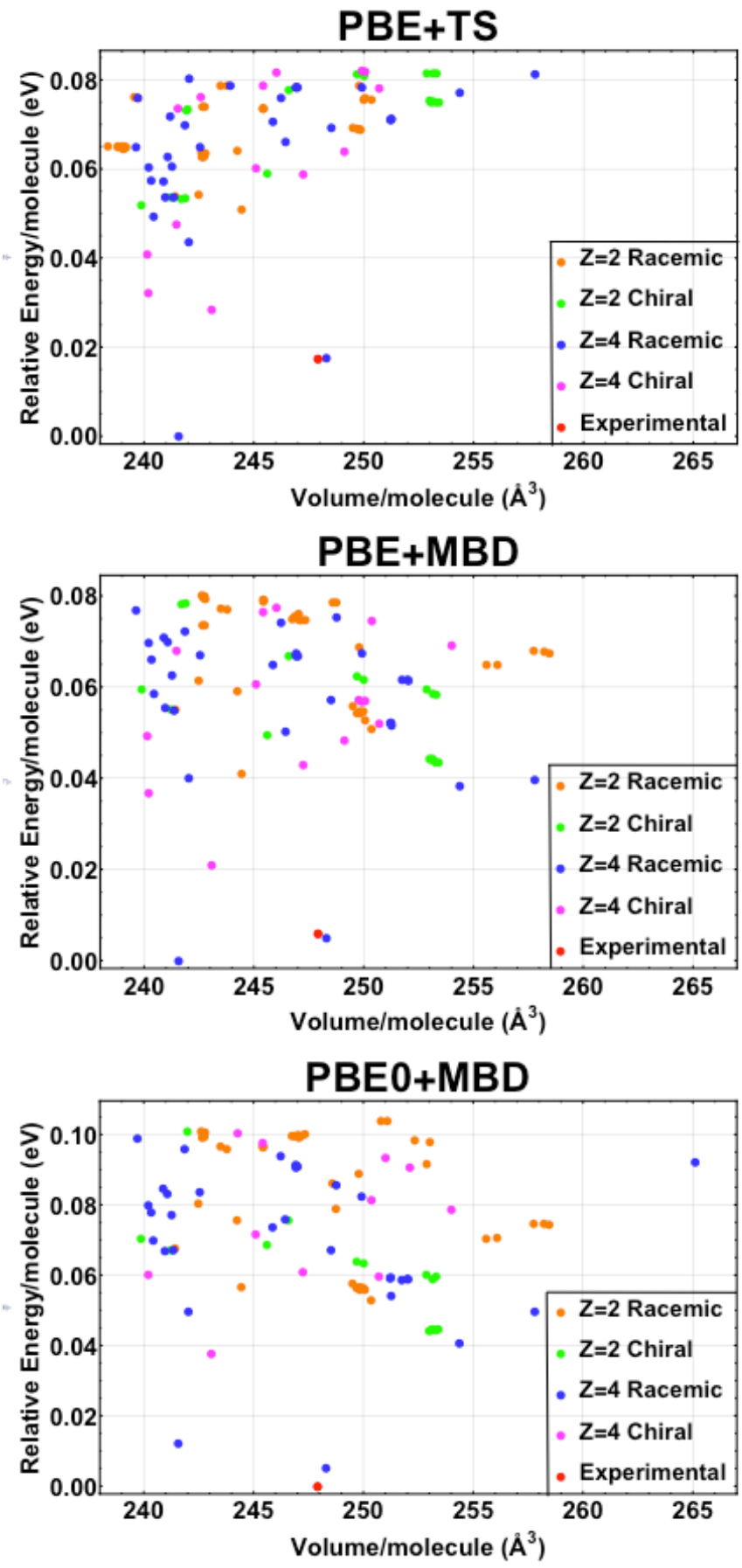


Figure 3: Volume per molecule versus energy plots for the top 100 structures as ranked by PBE+TS, PBE+MBD, and PBE0+MBD. The experimental structure is shown for reference.

molecule versus the energy per molecule for the best 100 structures from PBE+TS, PBE+MBD, and PBE0+MBD are shown in Fig. 3 along with the experimental structure. All of the top structures are extremely close in energy and fall within an interval of about 0.1 eV. The distribution of structures changes significantly depending on the method used. The TS method, which tends to overbind, favors structures with smaller specific volumes than the MBD method. The PBE0 functional increases the energy differences between structures and further stabilizes structures with lower densities as compared to PBE.

Fig. 4 shows the ranking of the top five PBE0+MBD structures for each method used. There is significant rearrangement between methods for most of the structures except for the top 4. Some structures, such as 42a9600b47 and d037fff743, have higher relative energies with PBE+TS but are stabilized dramatically by PBE+MBD and PBE0+MBD. The two best structures as ranked by PBE+TS (9f774c9e27 and 7471226271) become even closer in energy when computed with PBE+MBD but swap rankings with PBE0+MBD. The experimental structure is stabilized by both MBD and PBE0 and would have been ranked as number one with PBE0+MBD.

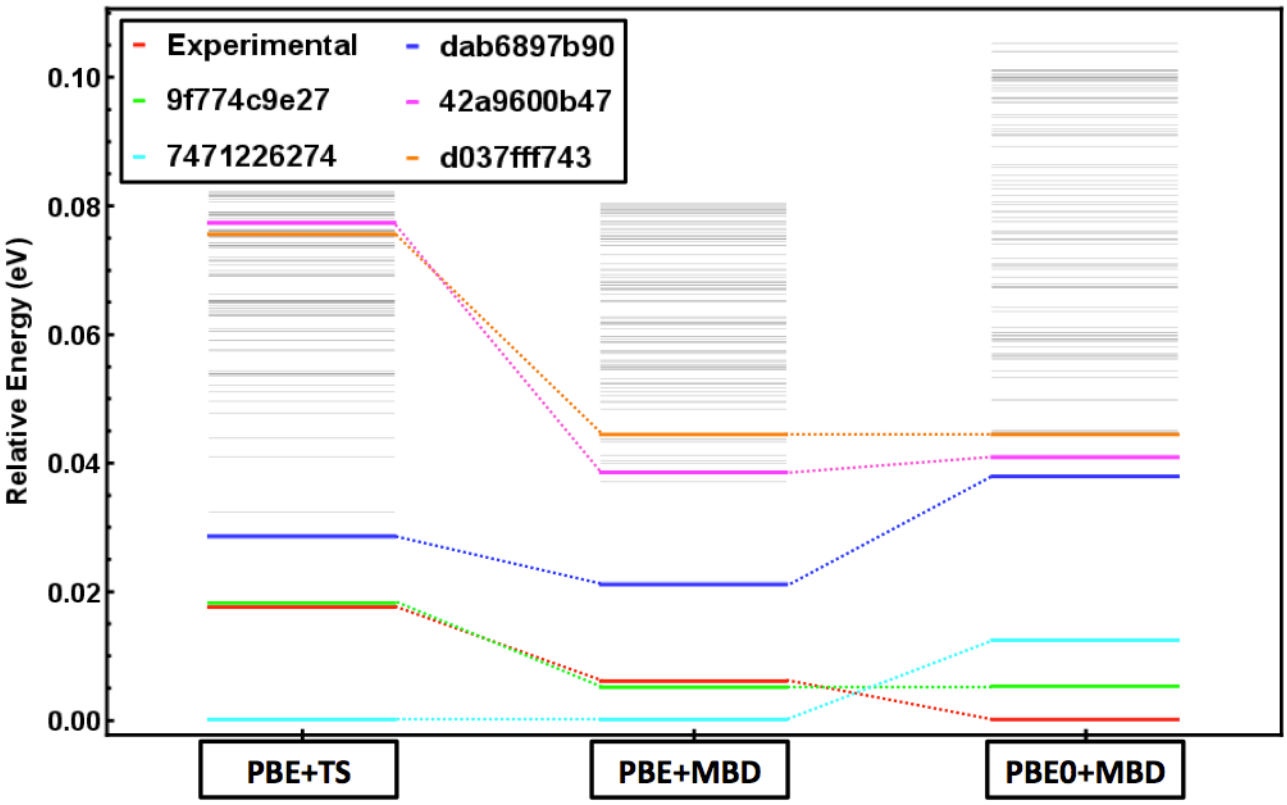


Figure 4: The top 5 PBE0+MBD structures along with the experimental structure as ranked by the different methods. All other structures in the top 100 are shown in gray.

Computational Time

We used approximately 30M CPU hours. Most calculations were done on Mira at the Argonne Leadership Computing Facility (ALCF) which is an IBM Blue Gene/Q, with 16-core 1.6 GHz PowerPC processors. Some additional calculations were performed on Tulane University's Intel Xeon E5-2680 cluster, Cypress, which has dual 10-core 2.8 GHz processors. Most of our CPU time was spent doing full unit cell relaxations with a fully quantum mechanical first-principles approach for approximately 10,000 structures within the GA itself. To the best of our knowledge, we were

the only group in the blind test to use an entirely DFT-based approach which also contributed to a much higher computational cost than if we had used force fields or other semi-empirical methods.

Acknowledgments

We thank Leslie Leiserowitz from the Weizmann Institute of Science and Geoffrey Hutchinson from the University of Pittsburgh for helpful discussions. We thank Adam Scovel at the Argonne Leadership Computing Facility (ALCF) for technical support. Work at Tulane University was funded by the Louisiana Board of Regents Award # LEQSF(2014-17)-RD-A-10 "Toward Crystal Engineering from First Principles", by the NSF award #EPS-1003897 "The Louisiana Alliance for Simulation-Guided Materials Applications (LA-SiGMA)", and by the Tulane Committee on Research Summer Fellowship. Work at the Technical University of Munich was supported by the Solar Technologies Go Hybrid initiative of the State of Bavaria, Germany. Computer time was provided by the Argonne Leadership Computing Facility (ALCF), which is supported by the Office of Science of the U.S. Department of Energy under contract DE-AC02-06CH11357.

References

[1] V. Blum, R. Gehrke, F. Hanke, P. Havu, V. Havu, X. Ren, K. Reuter, and M. Scheffler, Comput. Phys. Commun. **180**, 2175 (2009).
[2] S. Bhattacharya, S. V. Levchenko, L. M. Ghiringhelli, and M. Scheffler, Phys. Rev. Lett. **111**, 135501 (2013).
[3] S. Bhattacharya, S. V Levchenko, L. M. Ghiringhelli, and M. Scheffler, New J. Phys. **16**, 123016 (2014).
[4] S. Bhattacharya, B. H. Sonin, C. J. Jumonville, L. M. Ghiringhelli, and N. Marom, Phys. Rev. B **91**, 241115 (2015).
[5] J. P. Perdew, K. Burke, M. Ernzerhof, Phys. Rev. Lett. **77**, 3865 (1996).
[6] J. P. Perdew, K. Burke, M. Ernzerhof, Phys. Rev. Lett. **78**, 1396 (1997).
[7] A. Tkatchenko and M. Scheffler, Phys. Rev. Lett. **102**, 073005 (2009).
[8] K. Berland, E. Londero, E. Schröder, and P. Hyldgaard, Phys. Rev. B. **88**, 045431 (2013).
[9] M. A. Blanco , M. Flórez, M. Bermejo. J. Mol. Struct. **419**, 1927 (1997).
[10] W. A. Dollase. J. Am. Chem. Soc. **87**, 5 (1965).
[11] N. Marom, A. Tkatchenko, M. Rossi, V. V. Gobre, O. Hod, M. Scheffler, and L. Kronik, J. Chem. Theory Comput. **7**, 3944 (2011).
[12] D. C. Lonie and E. Zurek, Comput. Phys. Commun. **182**, 372 (2011).
[13] D. E. Goldberg, *Genetic Algorithms in Search, Optimization and Machine Learning* (Addison-Wesley, MA, 1989).
[14] A. Tkatchenko, R. A. Distasio, R. Car, and M. Scheffler, Phys. Rev. Lett. **108**, 236402 (2012).
[15] A. Ambrosetti, A. M. Reilly, R. A. Distasio, and A. Tkatchenko, J. Chem. Phys. **140**, 18A508 (2014).
[16] N. Marom, R. A. Distasio, V. Atalla, S. Levchenko, A. M. Reilly, J. R. Chelikowsky, L. Leiserowitz, and A. Tkatchenko, Angew. Chemie - Int. Ed. **52**, 6629 (2013).
[17] A. M. Reilly and A. Tkatchenko, J. Phys. Chem. Lett. **4**, 1028 (2013).
[18] J. P. Perdew, M. Ernzerhof, and K. Burke, J. Chem. Phys. **105**, 9982 (1996).
[19] C. Adamo and V. Barone, J. Chem. Phys. **110**, 6158 (1999).
[20] B. Santra, J. Klimeš, D. Alfé, A. Tkatchenko, B. Slater, A. Michaelides, R. Car, and M. Scheffler, Phys. Rev. Lett. **107**, 185701 (2011).
[21] A. M. Reilly and A. Tkatchenko, J. Chem. Phys. **139**, 024705 (2013).

3.2 Published Paper: Effect of packing motifs on the energy ranking and electronic properties of putative crystal structures of tricyano-1, 4-dithiino[c]-isothiazole

In this publication the structures generated for Target XXII in the sixth blind test were combined with the experimental structure. It was found that different dispersion-inclusive DFT methods systematically favored particular packing motifs, an important consideration in the context of hierarchal screening approaches. The electronic properties of a structure within 0.02 eV of the experimental structure are compared with those of the experimental structure. My contributions to this work include running the different DFT local relaxations and total energy evaluations of the putative crystal structures. I also categorized the structures into four main packing motifs and analyzed their sensitivity to the particular total energy method used. I performed Non Covalent Interaction (NCI) [142] calculations for the visualization of the different intermolecular interactions in selected crystal structures. I also wrote the entire manuscript.

STRUCTURAL SCIENCE
CRYSTAL ENGINEERING
MATERIALS

ISSN: 2052-5206
journals.iucr.org/b

Effect of packing motifs on the energy ranking and electronic properties of putative crystal structures of tricyano-1,4-dithiino[*c*]-isothiazole

Farren Curtis, Xiaopeng Wang and Noa Marom

Acta Cryst. (2016). B**72**, 562–570

STRUCTURAL SCIENCE
CRYSTAL ENGINEERING
MATERIALS

ISSN 2052-5206

Effect of packing motifs on the energy ranking and electronic properties of putative crystal structures of tricyano-1,4-dithiino[*c*]-isothiazole

Farren Curtis,[a] Xiaopeng Wang[b] and Noa Marom[b]*

[a]Department of Physics, Carnegie Mellon University, Pittsburgh, PA 15213, USA, and [b]Department of Materials Science and Engineering, Carnegie Mellon University, Pittsburgh, PA 15213, USA. *Correspondence e-mail: nmarom@andrew.cmu.edu

Received 14 March 2016
Accepted 7 June 2016

Edited by G. M. Day, University of Southampton, England

Keywords: crystal structure prediction; packing motifs; energy ranking; polymorphism; organic semiconductors; electronic structure; density functional theory.

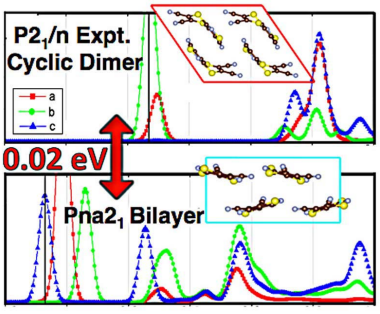
Supporting information: this article has supporting information at journals.iucr.org/b

We present an analysis of putative structures of tricyano-1,4-dithiino[*c*]-isothiazole (TCS3), generated within the sixth crystal structure prediction blind test. Typical packing motifs are identified and characterized in terms of distinct patterns of close contacts and regions of electrostatic and dispersion interactions. We find that different dispersion-inclusive density functional theory (DFT) methods systematically favor specific packing motifs, which may affect the outcome of crystal structure prediction efforts. The effect of crystal packing on the electronic and optical properties of TCS3 is investigated using many-body perturbation theory within the *GW* approximation and the Bethe–Salpeter equation (BSE). We find that a structure with $Pna2_1$ symmetry and a bilayer packing motif exhibits intermolecular bonding patterns reminiscent of π–π stacking and has markedly different electronic and optical properties than the experimentally observed $P2_1/n$ structure with a cyclic dimer motif, including a narrower band gap, enhanced band dispersion and broader optical absorption. The $Pna2_1$ bilayer structure is close in energy to the observed structure and may be feasible to grow.

1. Introduction

The ability of a molecule to crystallize in several different forms, or polymorphs, is of central importance to a wide variety of pharmaceutical and technological applications. Polymorphism has long been a point of interest for the pharmaceutical industry because many drugs are marketed as molecular crystals of the pharmaceutically active ingredient. Since subtle differences in molecular packing can lead to vastly different chemical and physical properties, pharmaceutical companies must be able to predict and control which form is being produced in order to deliver consistent products (Hilfiker, 2006; Sun, 2009). Furthermore, exploring different polymorphs can lead to the discovery of novel solid forms and the design of new pharmaceuticals (Price, 2014).

More recently, polymorphism has also received considerable attention for its role in organic electronics. Numerous small molecule organic semiconductors have been synthesized and characterized, showing promising charge transfer and optoelectronic properties, as reviewed in Yassar (2014). Single crystals of organic semiconductors, as opposed to polycrystalline films, are of particular interest due to their long-range molecular order, absence of grain boundaries and minimal concentration of charge traps (Jiang *et al.*, 2010). Different organic semiconductor polymorphs may display markedly different band structures, optoelectronic properties, electronic couplings and electron–phonon couplings that can drastically modify their performance in organic field effect



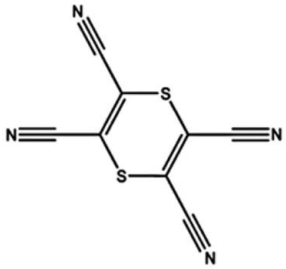
transistors and organic photovoltaics (Tseng *et al.*, 2008; Pfattner *et al.*, 2010; Wang *et al.*, 2013; Li *et al.*, 2015).

Organic field effect transistors require high carrier mobility. This is often achieved in organic semiconductors that possess π-conjugation and crystallize such that there is strong wavefunction overlap between neighboring molecules (Yassar, 2014). For example, rubrene has been one of the most widely studied organic semiconductors due to its excellent charge-transport properties. Single crystals of orthorhombic rubrene have demonstrated charge mobilities up to $20\ cm^2\ V^{-1}\ s^{-1}$ (Sundar *et al.*, 2004; Hulea *et al.*, 2006); Hasegawa & Takeya, 2009) which far surpasses amorphous silicon. The record-breaking mobility of the orthorhombic form is attributed to the herringbone packing motif which allows significant overlap of the π-conjugated tetracene backbone (da Silva Filho *et al.*, 2005). The monoclinic and triclinic forms, however, have different packing motifs and show much lower mobilities (Hathwar *et al.*, 2015). Even organic semiconductors with visually similar packing motifs may exhibit markedly different transport characteristics. Such is the case in a recent study of two π-stacked solution and vapor grown polymorphs of 6,13-dichloropentacene crystals (Hatcher *et al.*, 2015).

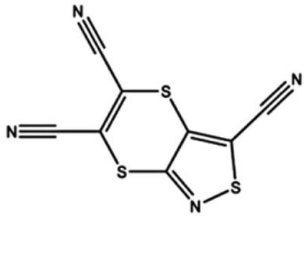
Singlet fission in molecular organic semiconductors has been the focus of many recent studies since it could allow organic photovoltaic devices to surpass the traditional Shockley–Queisser limit of solar cell conversion efficiency (Shockley & Queisser, 1961). Singlet fission is a spin-allowed process where one singlet exciton is converted into two triplet excitons, which are dipole-forbidden states that would not couple to light otherwise (Teichen & Eaves, 2015). This allows the extraction of two carriers per absorbed photon. Different polymorphs of organic semiconductors display different rates of singlet fission (Dillon *et al.*, 2013). The herringbone packing motif can actually suppress exciton fission, while slip-stacked arrangements favor delocalization and provide efficient exciton fission (Kolata *et al.*, 2014). For rubrene, it has been suggested that the triclinic and monoclinic forms may exhibit more efficient singlet fission than the orthorhombic form (Wang *et al.*, 2016).

Although there have been many advances in organic semiconductor research and design, there is still much to be understood about the role of intermolecular interactions in solid-state packing, how molecular packing affects electronic properties, and how to experimentally stabilize thermodynamically feasible alternative forms with the desired electronic properties. Experimental techniques such as tailor-made additives (Lahav & Leiserowitz, 2015), epitaxial templating (Salzmann *et al.*, 2012) and solution shearing (Giri *et al.*, 2011; Diao *et al.*, 2013) show promise for molecular packing motifs in crystal structures that are not predicted as the most thermodynamically stable. Crystal structure prediction efforts, combined with highly accurate electronic structure calculations may also provide insight (Beran, 2016).

Crystal structure prediction blind tests, organized by the Cambridge Crystallographic Data Centre (CCDC), are held every few years to assess the advances and remaining challenges of organic crystal structure prediction (Lommerse *et al.*, 2000; Motherwell *et al.*, 2002; Day *et al.*, 2005, 2009; Bardwell *et al.*, 2011). The small rigid target for the sixth blind test [Target (XXII)] was tricyano-1,4-dithiino[*c*]-isothiazole (TCS3; Reilly *et al.*, 2016). TCS3 crystallizes as bright yellow needles and belongs to a class of thiacyanocarbon compounds which only contain carbon, nitrogen, sulfur and a plurality of cyano groups (Simmons *et al.*, 1962). They are very electron-deficient and react readily with electron-rich and neutral molecules (Webster, 2002). Furthermore, thiacyanocarbons like TCS3 offer a unique opportunity to study intermolecular interactions other than highly directional hydrogen bonds (Dollase, 1965).



tetracyano-1,4-dithiin (TCS2)



tricyano-1,4-dithiino[c]-isothiazole (TCS3)

Here, putative structures of TCS3, generated within the sixth crystal structure prediction blind test, are used as a test case. We identify typical low-energy packing motifs and analyze their effect on the energy ranking by dispersion-inclusive density functional theory (DFT) methods. We further investigate the nature of the intermolecular interactions that generate these packing motifs. Finally, we relate specific packing motifs to trends in the electronic and optical properties. We find that a putative $Pna2_1$ structure with a bilayer packing motif has a narrower band gap, greater band dispersion and broader optical absorption. This structure is close in energy to the experimentally observed structure and may be possible to grow.

2. Methodology

2.1. Ranking of putative TSC3 crystal structures

Within the sixth crystal structure prediction blind test approximately 5000 putative crystal structures of TCS3 were produced by the first-principles DFT-based genetic algorithm (GA), GAtor (for more details, see the supporting information of submission 12 in Reilly *et al.*, 2016). These crystal structures were relaxed within GAtor using the all-electron electronic structure package FHI-aims (Blum *et al.*, 2009) employing the Perdew–Burke–Ernzerhof generalized gradient approximation (PBE; Perdew *et al.*, 1996, 1997) coupled to the Tkatchenko–Scheffler (TS) pairwise dispersion correction, PBE+TS (Tkatchenko & Scheffler, 2009), with *lower-level* numerical settings, which correspond to the light/tier1 settings of FHI-aims (Blum *et al.*, 2009). The top 500 structures generated by the GA were re-relaxed and re-ranked using PBE+TS with *higher-level* numerical settings, which correspond to the tight/tier2 settings of FHI-aims. Single-point

energy evaluations were performed using PBE with the many-body dispersion method (MBD) (Tkatchenko *et al.*, 2012; Ambrosetti *et al.*, 2014) for 200 of the best structures as ranked by PBE+TS. The MBD method accounts for long-range electrostatic screening effects and for non-pairwise-additive many-body contributions to the dispersion energy. It has been shown to accurately rank the energetic stability of molecular crystal polymorphs in cases where the pairwise TS approach is not sufficiently accurate (Marom *et al.*, 2013; Reilly & Tkatchenko, 2013*a*). Finally, single-point energy evaluations were performed for 150 of the best structures as ranked by PBE+MBD using the hybrid functional PBE0 (Perdew *et al.*, 1996; Adamo & Barone, 1999) with the MBD correction. The inclusion of 25% exact exchange in PBE0 mitigates the self-interaction error, leading to a more accurate description of electron densities and multipoles (Santra *et al.*, 2013; Reilly & Tkatchenko, 2013*a*,*b*). For some molecular crystals, such as glycine and oxalic acid, the correct polymorph ranking is reproduced only when using PBE0+MBD (Marom *et al.*, 2013). We therefore consider the ranking by PBE0+MBD to be the most reliable of the methods used here. All energy evaluations were performed using a $3 \times 3 \times 3$ k-point grid.

2.2. Analysis of intermolecular interactions and packing motifs

The effect of non-covalent interactions on molecular crystal packing is often described in terms of close pairwise distances between atoms shorter than the sum of their van der Waals (vdW) radii, but this criterion can be unreliable (Klein, 2006; Schiemenz, 2007). In addition to close-contact analysis, more sophisticated algorithms can be used for analyzing non-covalent interactions based on the electronic and kinetic energy densities of the system. One such approach, known as the non-covalent interaction index (NCI) (Johnson *et al.*, 2010; Otero-de-la-Roza *et al.*, 2012), identifies the stabilizing and destabilizing non-covalent interactions *via* the change in the dimensionless reduced density gradient (RDG) in regions between interacting atoms. Close contact and NCI analyses were performed for the putative crystal structures of TCS3 to highlight the differences between common packing motifs. Periodic NCI calculations were performed with the *Critic*2 program (Otero-de-la-Roza *et al.*, 2009, 2014) using a cutoff of RDG = 0.3. A color scale ranging from blue to green to red ($-0.02 < \rho < 0.02$) is used for plotting isosurfaces, signifying stabilizing, intermediate and destabilizing overlap regions, respectively.

2.3. Electronic properties

Band structures of the top 36 structures, within an energy window of 70 meV of the experimental structure, as ranked by PBE0+MBD, were calculated using PBE with the tight/tier2 settings of FHI-aims. Although PBE is known to severely underestimate band gaps, it still provides valuable qualitative information on the relationship between different packing motifs and their electronic properties, which can help identify the most promising candidate structures for further analysis. The band structures were computed along the high-symmetry paths suggested in Setyawan & Curtarolo (2010). The quasi-particle band structures and optical properties of the experimental $P2_1/n$ cyclic dimer structure and the most stable $Pna2_1$ bilayer structure were calculated using the GW approximation and the Bethe–Salpeter equation (BSE) with the BerkeleyGW code (Deslippe *et al.*, 2012). First, PBE eigenvectors and eigenvalues were generated with Quantum Espresso (Giannozzi *et al.*, 2009), using Troullier–Martins norm-conserving pseudopotentials (Troullier & Martins, 1991). The pseudopotentials were generated with FHI98PP (Fuchs & Scheffler, 1999) considering 4, 5 and 6 valence electrons for carbon, nitrogen and sulfur, respectively. The PBE calculation was performed with a $4 \times 4 \times 2$ k-point grid and a kinetic energy

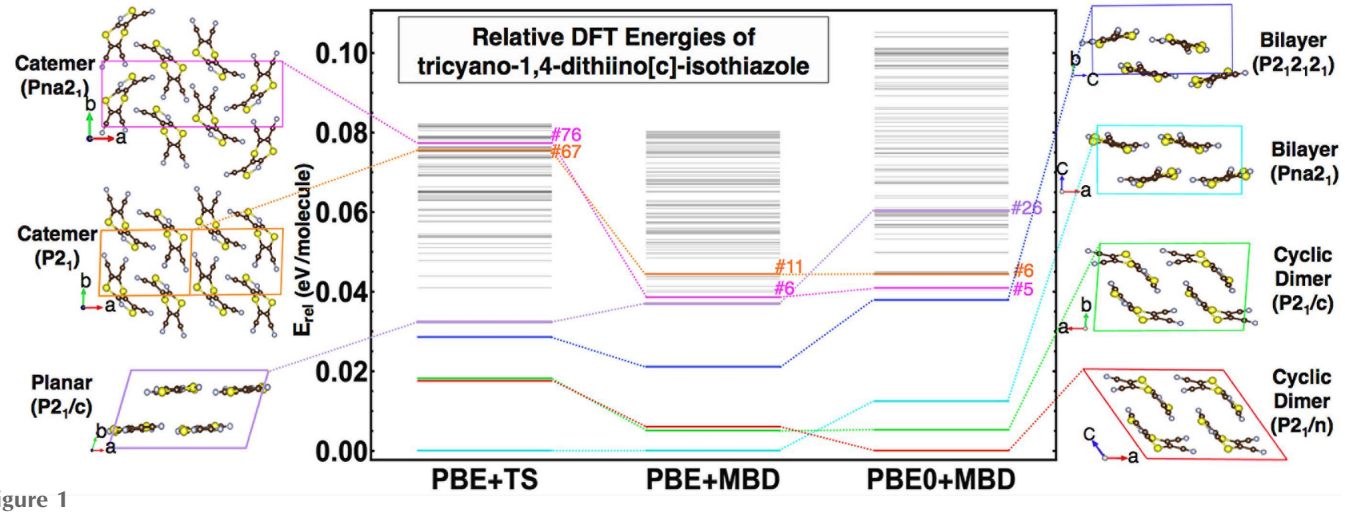


Figure 1
The relative energies of putative crystal structures of tricyano-1,4-dithiino[*c*]-isothiazole as obtained with different dispersion-inclusive DFT methods. The ranking of higher-lying structures is indicated by numbers. The experimental structure crystallizes in $P2_1/n$, but is displayed in the alternative $P2_1/c$ representation (shown in red) to highlight structural similarities to the $P2_1/c$ structure (shown in green). C, N and S atoms are plotted in brown, light blue and yellow, respectively.

cutoff of 60 Ry. Results based on these parameters agree well with the tight/tier2 numerical basis set of FHI-aims (Blum *et al.*, 2009) with a difference of less than 20 meV in the band gap. Next, non-self-consistent G_0W_0 was used to compute the quasiparticle GW band structures. The dielectric function and self-energy operator were constructed by summing over 1058 unoccupied bands for both the $Pna2_1$ bilayer and the $P2_1/n$ cyclic dimer structures. The static remainder correction (Deslippe *et al.*, 2013) was applied to accelerate convergence with respect to the number of unoccupied states. Energy cutoffs of 8 Ry and 60 Ry were adopted for the screened and bare Coulomb potentials, respectively. Lastly, the optical excitation properties of the two polymorphs were obtained by solving the Bethe–Salpeter equation (Rohlfing & Louie, 2000) within the Tamm–Dancoff approximation (Deslippe *et al.*, 2012). For the BSE calculation, a denser k-point grid of $8 \times 8 \times 4$ was used. 20 valence bands and 20 conduction bands were considered for both structures. In these calculations the polarization of light was directed along the three crystal axes of each polymorph. The parameters for the GW/BSE calculations are similar to those used in Sharifzadeh *et al.* (2013) and Samsonidze *et al.* (2014). The exciton wavefunctions were visualized for the lowest energy singlet and triplet excitations of the $Pna2_1$ bilayer and the $P2_1/n$ cyclic dimer structures by fixing the hole position, about 1 Å above a S atom, where there is a high hole probability (Sharifzadeh *et al.*, 2012, 2013), using an $8 \times 8 \times 4$ supercell.

3. Results and discussion

3.1. Relation between DFT energy ranking and packing motifs of TCS3

Fig. 1 shows the ranking of the top 100 putative crystal structures of TCS3, obtained with different exchange–correlation functionals and dispersion methods. A full account of the coordinates and energies of these structures is provided in the supporting information of submission 12 in Reilly *et al.* (2016). The most stable structures and representative structures whose ranking are strongly method dependent are highlighted in color. The coordinates of these structures are provided in the supporting information of this article. The structures are illustrated and classified according to packing motifs, which are labelled cyclic dimer, catemer (borrowing the terminology used for packing motifs of carboxylic acids), bilayer and planar (in the sense that the molecules adopt a planar conformation).[1]

The two cyclic dimer structures are predicted to be nearly degenerate by PBE+TS and PBE+MBD. Only PBE0+MBD ranks the experimentally observed $P2_1/n$ structure as the most stable. This is consistent with the superior performance of PBE0+MBD for other polymorphic systems (Marom *et al.*, 2013; Reilly & Tkatchenko, 2013*b*, 2013*a*; Santra *et al.*, 2013). The $P2_1$ and $Pna2_1$ catemer structures are systematically destabilized by PBE+TS, which ranks them as #67 and #76, respectively. The inclusion of many-body dispersion interactions stabilizes these structures, which are ranked as #11 and #6 by PBE+MBD and as #6 and #5 by PBE0+MBD. This indicates that a pairwise description of the non-covalent interactions in the catemer-like packing motifs is not sufficiently accurate, consistent with the findings of Marom *et al.* (2013) for carboxylic acids. The layered planar and bilayer structures are overstabilized by PBE+TS and PBE+MBD with respect to the catemer and cyclic dimer motifs. The PBE0 functional systematically destabilizes these layered motifs. In particular the planar structure is ranked as #5 by PBE+TS and PBE+MBD, but as #26 by PBE0+MBD. Our findings are consistent with the overstabilization of the layered α and β polymorphs of glycine by PBE, compared to the helical γ form (Marom *et al.*, 2013). This may be a consequence of the self-interaction error in PBE, the spurious Coulomb repulsion of an electron from itself (Perdew & Zunger, 1981), which favors the delocalized electron densities in layered structures. Our results are also consistent with the findings of Beran (2016) that error cancellation and small biases in electronic structure methods make it difficult to obtain reliable relative rankings of molecular crystal polymorphs with very different packing motifs.

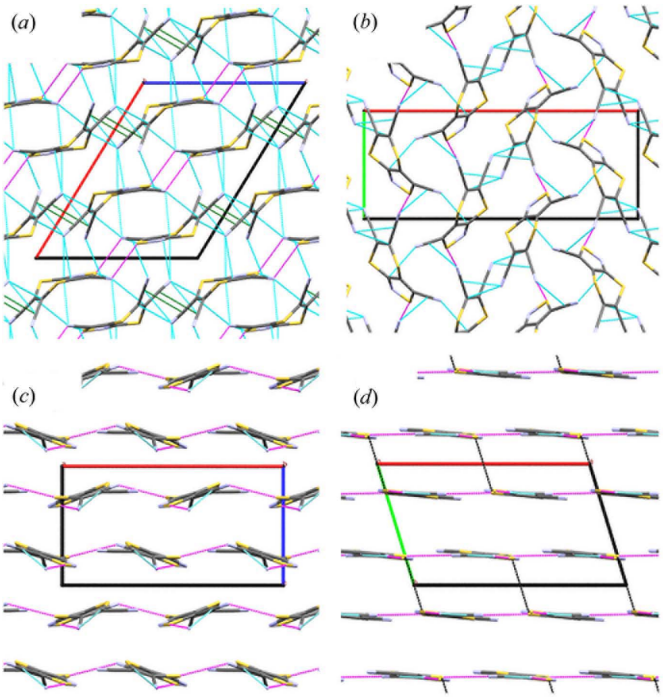


Figure 2
Close C···N (cyan), C···C (green), S···N (magenta) and S···S (black) contacts for (*a*) $P2_1/n$ cyclic dimer, (*b*) $Pna2_1$ catemer, (*c*) $Pna2_1$ bilayer and (*d*) $P2_1/c$ planar structures of TCS3. C, N and S atoms are plotted in grey, light blue and yellow, respectively

[1] We note that the packing motifs of some structures are similar in terms of visual appearance and intermolecular interactions, however, they are not strictly identical. For example, the $Pna2_1$ bilayer structure contains two mirror-related molecules, whereas the two $P2_12_12_1$ structures of opposite handedness contain only one of the two mirror images. This possibly contributes to the packing efficiency and stability of the $Pna2_1$ layered structure. Similarly, the $Pna2_1$ catemer structure may be stabilized compared with the $P2_1$ catemer structure due to the mirror-related molecules.

Since the putative structures considered here are all within 100 meV, it should be noted that both zero-point and finite-temperature vibrational contributions may also affect their relative energy ranking (Reilly & Tkatchenko, 2014; Nyman & Day, 2015; Beran, 2016). Submission 25 of Reilly *et al.* (2016) re-ranked the final structures of Submission 18 using PBE+TS and PBE+MBD with PBE+TS vibrational free energies at 300 K. Upon inclusion of the vibrational free energy contribution the experimental structure was ranked as #1, the $P2_1/c$ cyclic dimer as #2 and the $Pna2_1$ bilayer as #3, with a maximum difference of 20 meV. These results are consistent with the PBE0+MBD relative energies found here.

3.2. The nature of intermolecular interactions in TCS3

Fig. 2 shows the network of close contacts formed by the different packing motifs of TCS3 for representative structures. The *a*, *b* and *c* crystal axes are colored in red, green and blue, respectively. C···N intermolecular contacts shorter than 3.25 Å and S···N contacts shorter than 3.35 Å are shown in cyan and magenta, respectively. C···C intermolecular contacts shorter than 3.4 Å and S···S contacts shorter than 3.6 Å are shown in green and black, respectively. In the experimental $P2_1/n$ cyclic dimer motif (panel a), two symmetric S···N close contacts are observed between the S in the isothiazole ring and the N in the additional CN group between molecules paired together along the *a* direction, while networks of C···N contacts are found between molecules in the *c* direction and also along the stacking direction. Additionally, the experimental $P2_1/n$ cyclic dimer contains close C···C contacts, which are 3.36 Å each and displayed in green. For the $Pna2_1$ catemer motif (panel b), networks of C···N intermolecular contacts are formed along the direction of the catemer chains, while the S in the thiazole ring and the N in the adjacent CN align with posterior N atoms of molecules in the *b* direction. This forms a triangular network of S···N and C···N contacts. In the layered $Pna2_1$ structure (panel c), the same S···N and C···N triangular network is formed in the plane of the molecules, directed out of the page, while S···N contacts are formed between the S on the six-membered dithiin ring and the N in the additional CN group of a neighboring molecule directed along the *a* direction. No close contacts are found along the stacking direction. In the planar $P2_1/c$ structures (panel d), the S···N and C···N triangular network is directed in the *c* direction and intra-layer S···N contacts are found in the *a* direction. These contacts are also present in the bilayer motif. Additionally, the planar structure contains S···S close contacts, displayed in black, which are found between alternating layers in the stack. These have a distance of 3.45 Å. The planar structure has the smallest S···N distance of 2.9 Å (in the plane of the molecules), the layered structures show longer S···N distances around 3.0 Å, followed by the catemer motifs with 3.1–3.2 Å, and the cyclic dimer motifs with approximately 3.3 Å. For additional details see Table S1 in the supporting information.

Fig. 3 shows NCI isosurfaces for representative structures. NCI analysis can reveal complex intermolecular interactions derived from the electron density and reduced-density gradient (RDG). Regions of localized intermolecular interactions manifest as well defined spheroidal shapes while extended regions of stabilization are represented by more delocalized, less directional surfaces. Since TCS3 comprises only C, N and S atoms, it does not form highly directional hydrogen bonds or the π–π interactions typical of aromatic hydrocarbons. Rather, the C, N and S atoms form complex intermolecular interactions with the five- and six-membered rings.

In the cyclic dimer motifs (panel a), complex delocalized interactions are found in the C···N and C···C network of contacts between molecules in the *c* direction (boxed in cyan). Significant overlap, represented as green ribbons, is found in the region between the dimer pairs (boxed in magenta). This may be attributed to the combined effects of interactions between the thiazole ring, the additional CN group and the dithiin ring containing C and S. This sheet-like feature is reminscent of a surface one would find for π–π interactions in a benzene dimer (Johnson *et al.*, 2010). In the $Pna2_1$ catemer (panel b), there are blue stabilizing regions between N and C contacts (boxed in cyan). In this region there are only two C···N contacts, as opposed to the cyclic dimer case where four similar C···N contacts and two C···C contacts are observed.

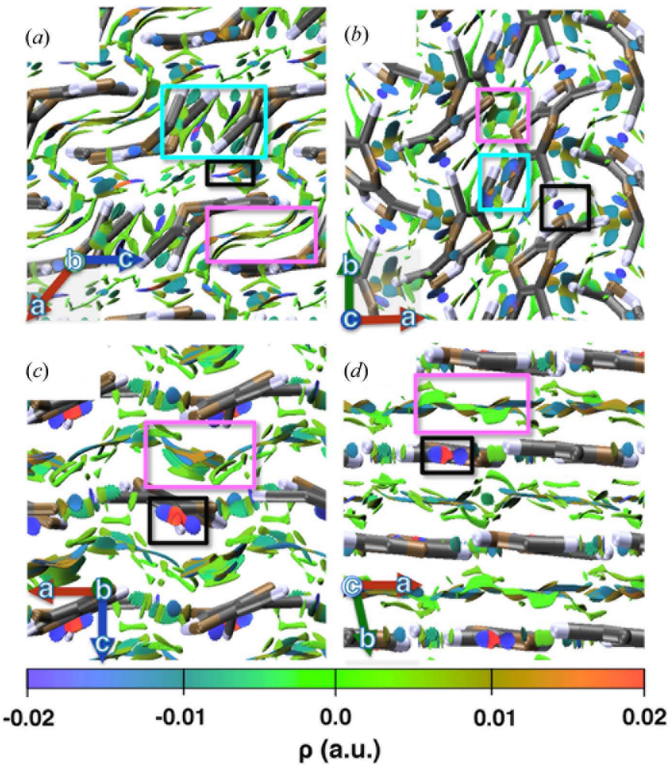


Figure 3
NCI plots for (*a*) $P2_1/n$ cyclic dimer, (*b*) $Pna2_1$ catemer, (*c*) $Pna2_1$ bilayer and (*d*) $P2_1/c$ planar structures of TCS3. NCI isosurfaces were obtained with RDG = 0.6 a.u. The color scale, which ranges from blue to green to red ($-0.02 < \rho < 0.02$ a.u.), signifies stabilizing, intermediate and destabilizing overlap regions, respectively. Cyan, magenta and black boxes highlight different regions of each motif containing C···N, π–π like and S···N interactions, respectively.

Additionally, there are π–π-like interactions between oppositely faced six-membered dithiin rings, as indicated by a diffuse blue/green isosurface region (boxed in magenta). The directional S···N interactions in the *b* direction are localized between the N in the posterior end of one molecule, and the S in the thiazole ring of a neighboring molecule (boxed in black).

In the $Pna2_1$ bilayer motif (panel c), there are large regions of delocalized interactions between the thiazole ring, the six-membered dithiin rings and the additional CN group along the *a* and *b* directions (boxed in magenta). These significant interactions are not identified by close-contact analysis, and only revealed by NCI. The surfaces between layers along the *b* direction resemble those of extended π–π interactions. Additionally, there are strongly stabilizing directional S···N and C···N interactions in the plane of the layers in the *b*-direction (coming out of the page) between the S and N of the thiazole ring and the N of a neighboring molecule, as also indicated by close-contact analysis, boxed in black. The planar structure (panel d), exhibits similar features to the bilayer motif. There are extended regions of π–π-like interactions between the thiazole and dithiin rings (boxed in magenta) between the planes of the molecules in the *c* direction, and there are directional in-plane S, N and C interactions in the *c* direction, also observed in the bilayer. Significant extended regions are observed for the alternating layers that contain close S···S contacts. It should be noted that a related molecule, tetracyano-1,4-dithiin (TCS2), which only contains the dithiin ring, has been reported to crystallize in the dimer and catemer motifs (Simmons *et al.*, 1962; Dollase, 1965). It is possible that the additional thiazole ring in TCS3 stabilizes the planar and layered packing motifs that would not be favorable for TCS2. Overall, NCI analysis provides a more complete picture and a deeper understanding of the interplay of intermolecular interactions that give rise to the different packing motifs of TCS3.

3.3. Effect of crystal packing on the electronic and optical properties of TCS3

The electronic and optical properties of TCS3 have not yet been characterized experimentally. The PBE band structures of the top 36 PBE0+MBD structures, displayed in Fig. S1 of the supporting information, reveal the variety of electronic properties that can be obtained from the same organic semiconductor by modifying its crystal packing. Layered structures generally exhibit smaller band gaps and greater band dispersion than structures with dimer and catemer packing motifs. Within the 36 structures investigated, the PBE band gaps of the structures with planar, bilayer, dimer and catemer motifs are 1.40, 1.39, 1.64 and 1.74 eV, respectively. A cyclic dimer motif which differs from that of the experimental structure is shown in panel (i) of Fig. S1, where the two dimer pairs in the unit cell are more orthogonal in their mutual orientation than seen in the experimental structure. This structure has a smaller-than-average PBE gap of 1.57 eV and particularly flat bands in comparison with the experimental structure and other structures with cyclic dimer motifs in the set. The most stable planar structure [panel (h) of Fig. S1] is within 50 meV of the experimental structure. It possesses the smallest gap in the set (1.33 eV) and shows significant band dispersion. The most stable $Pna2_1$ bilayer structure has a much smaller PBE gap [panel (b) of Fig. S1] than the experimentally observed structure (1.40 *versus* 1.76 eV, respectively) and is within 20 meV of the experimental structure as ranked by PBE0+MBD. A recent study (Nyman & Day, 2015) of 508 polymorphic organic molecules showed that over half of the polymorphic pairs had energy differences smaller than 20 meV. Therefore, the $Pna2_1$ bilayer structure may be possible to grow and we focus on it for further analysis.

Fig. 4 shows G_0W_0@PBE quasiparticle band structures of the experimentally observed $P2_1/n$ structure and the $Pna2_1$ bilayer structure. The bilayer structure has a fundamental gap of 3.80 eV, smaller by 0.35 eV than the fundamental gap of the cyclic dimer structure, 4.15 eV. While the bands of the cyclic dimer structure are nearly flat in most of the first Brillouin zone, the bilayer structure has significant band dispersion, in particular, near the top of the valence band. The band dispersion of the layered structure is reminiscent of π-stacked organic semiconductors (Cudazzo *et al.*, 2015; Zhu *et al.*, 2014; Fonari *et al.*, 2014) and may contribute to a smaller carrier effective mass and better transport properties.

Fig. 5 shows the BSE absorption spectra of the experimental structure and the $Pna2_1$ bilayer structure, for light polarized along the three crystal axes. The energy range shown is, for the most part, below the fundamental gaps of the two structures, corresponding to bound excitons. The optical gap of the layered structure, 2.93 eV, is smaller by 0.34 eV than the optical gap of the cyclic dimer structure, 3.27 eV. Both structures exhibit strong absorption of light polarized along the *a* and *b* directions and considerably weaker absorption along the *c* direction. For the cyclic dimer structure the *a* and *b* direc-

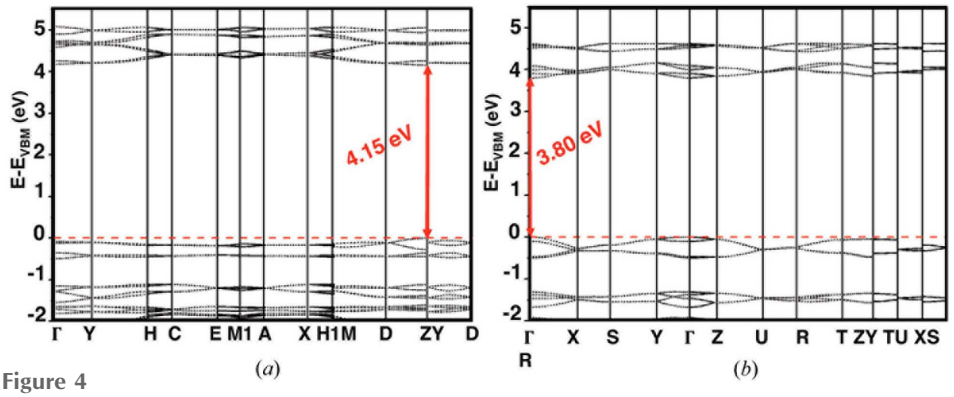


Figure 4
The G_0W_0@PBE quasiparticle band structures of (*a*) the experimental $P2_1/n$ cyclic dimer structure (represented in $P2_1/a$) and (*b*) The $Pna2_1$ bilayer structure.

tions correspond to the direction in which dimer pairs are stacked and the direction in which dimer pairs show symmetric C···N and C···C interactions, respectively. For the bilayer structure, the *a* and *b* directions correspond to the in-plane C···N and S···N interactions and the direction of stacking, respectively. For both structures, light polarized along the stacking directions yields the strongest absorption peak. In addition, for the layered structure the first absorption peak shows a significant splitting between the three crystal axes. The cyclic dimer structure has two relatively narrow absorption bands centered around 3.25 and 3.85 eV, whereas the bilayer structure has a much broader absorption spectrum with several prominent features, consistent with its greater band dispersion.

Excitons in molecular crystals are often described as having a Frenkel character, where the electron and hole charges are localized on the same molecule, or a charge transfer character, where the electron and hole charges are localized on different molecules, or a combination thereof (Cudazzo *et al.*, 2013, 2015). The excitonic properties of organic semiconductors are related to both molecular structure and crystal packing (Cudazzo *et al.*, 2012, 2013, 2015; Sharifzadeh *et al.*, 2012, 2013, 2015; Li *et al.*, 2014; Sai *et al.*, 2008; Hummer *et al.*, 2004, 2005; Hummer & Ambrosch-Draxl, 2005; Ambrosch-Draxl *et al.*, 2009). Here we isolate the effect of crystal packing for TCS3. Fig. 6 shows the wavefunctions of the lowest energy singlet and triplet excitons for the experimentally observed $P2_1/n$ cyclic dimer structure and the $Pna2_1$ layered structure, represented as the electronic charge distribution with respect to a hole, located near an S atom (marked in red). In general, the excitons in both structures are distributed over many unit cells. The singlet exciton of the layered structure (Fig. 6*b*) has significant electron density surrounding the hole site. It may therefore be described as having a combined Frenkel and charge-transfer character. The singlet exciton of the cyclic dimer structure (Fig. 6*a*) and the triplet excitons of both structures (Figs. 6*c*, *d*) have little charge distribution in the vicinity of the hole and may therefore be described as more charge transfer like. The lowest triplet excitation energies are 2.45 eV for the cyclic dimer structure and 2.04 eV for the layered structure, making singlet fission energetically unfavorable in both. However, the strong dependence of the excitonic properties of TCS3 on crystal packing hints at a possible way of tuning the crystal structure of chemically similar singlet fission chromophores (Busby, Xia, Low *et al.*, 2015; Busby, Xia, Wu *et al.*, 2015) to achieve improved efficiency.

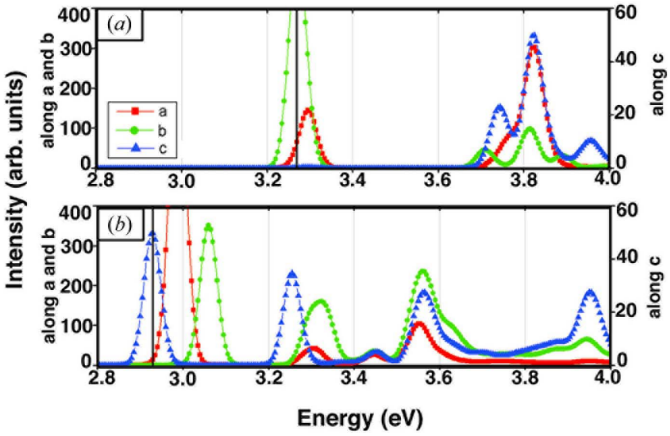


Figure 5
The G_0W_0/BSE absorption spectra of (*a*) the experimental $P2_1/n$ cyclic dimer structure (represented in $P2_1/a$) and (*b*) the $Pna2_1$ bilayer structure for light polarized along their respective crystal axes. The optical gap is marked with a black vertical line. Gaussian broadening with a width of 0.02 eV was used for the plots.

4. Conclusion

We have analyzed putative structures of TCS3, generated within the sixth crystal structure prediction blind test, and identified typical packing motifs, characterized by distinct patterns of close contacts and regions of electrostatic and dispersion interactions. We find that different dispersion-inclusive DFT methods systematically favor particular packing motifs. Structures with catemer-like chain motifs are destabilized with respect to cyclic dimer structures by the TS pairwise dispersion method and stabilized by the inclusion of many-body dispersion interactions and long-range screening effects in the MBD method. Structures with layered motifs are overstabilized by the semi-local PBE functional, compared with the hybrid PBE0 functional, possibly due to the self-interaction error. Only with PBE0+MBD is the experimentally observed $P2_1/n$ cyclic dimer structure found to be the

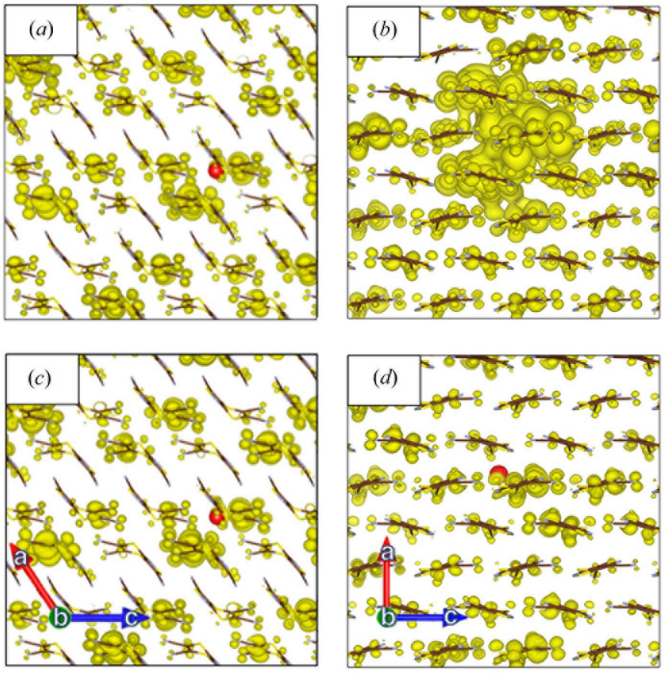


Figure 6
Exciton wavefunctions represented by the electronic charge distribution with respect to a hole position, marked by a red sphere, for the lowest energy singlet exciton (panels *a*, *b*) and triplet exciton (panels *c*, *d*) of the $P2_1/n$ cyclic dimer structure (left) and the $Pna2_1$ Bilayer structure (right). An isosurface value of 200 e Bohr^{-3} was used for plotting the surfaces.

most stable. The preference of specific packing motifs by different total energy methods may substantially alter the results of crystal structure prediction efforts. In particular, erroneous destabilization may lead to loss of important structures in hierarchical screening approaches. We therefore recommend careful validation and analysis of the sensitivity of the energy ranking to packing motifs as a best practice for crystal structure prediction.

Further analysis of the effect of crystal packing on the electronic and optical properties of TCS3 focused in particular on comparing the experimentally observed $P2_1/n$ cyclic dimer structure to a closely ranked $Pna2_1$ bilayer structure. The layered structure exhibits delocalized intermolecular bonding patterns reminiscent of π–π stacking, which do not exist in *e.g.* TCS2, and emerge due to the additional five-membered thiazole ring and strong in-plane S$\cdots$N and C$\cdots$N interactions. The layered structure possesses markedly different electronic and optical properties from the cyclic dimer structure, including a narrower band gap, enhanced band dispersion and broader optical absorption. This demonstrates that the electronic properties of organic semiconductors depend strongly on crystal packing and may thus be tuned to achieve improved device performance.

The $Pna2_1$ bilayer structure is close in energy to the $P2_1/n$ cyclic dimer structure (and even predicted to be more stable by some methods). Therefore, it may be feasible to grow it, *e.g.* by using tailor-made additives (Lahav & Leiserowitz, 2015). If the $Pna2_1$ structure can crystallize from the same solvent as the $P2_1/n$ structure, the solution may contain pre-critical nuclei of both polymorphs. A tailor-made additive could then be designed to inhibit the growth of the centrosymmetric $P2_1/n$ crystal by adsorbing onto opposite ends of the crystal. At the same time, the additive would only adsorb at one polar end of the $Pna2_1$ polymorph, allowing it to grow. Thus, crystal forms of TCS3 with different electronic properties may be synthesized.

Acknowledgements

Computer time was provided by the Argonne Leadership Computing Facility (ALCF), which is supported by the Office of Science of the US Department of Energy under contract DE-AC02-06CH11357, and by the National Energy Research Scientific Computing Center (NERSC), which is supported by the Office of Science of the US Department of Energy under Contract No. DE-AC02-05CH11231. The authors thank Álvaro Vázquez-Mayagoitia from ALCF and Leslie Leiserowitz from the Weizmann Institute of Science for helpful discussions.

References

Adamo, C. & Barone, V. (1999). *J. Chem. Phys.* **110**, 6158–6170.
Ambrosch-Draxl, C., Nabok, D., Puschnig, P. & Meisenbichler, C. (2009). *New J. Phys.* **11**, 125010.
Ambrosetti, A., Reilly, A. M., DiStasio, R. A. & Tkatchenko, A. (2014). *J. Chem. Phys.* **140**, 18A508.
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. (2011). *Acta Cryst.* B**67**, 535–551.
Beran, G. J. (2016). *Chem. Rev.* **116**, 5567–5613.
Blum, V., Gehrke, R., Hanke, F., Havu, P., Havu, V., Ren, X., Reuter, K. & Scheffler, M. (2009). *Comput. Phys. Commun.* **180**, 2175–2196.
Busby, E., Xia, J., Low, J. Z., Wu, Q., Hoy, J., Campos, L. M. & Sfeir, M. Y. (2015). *J. Phys. Chem. B*, **119**, 7644–7650.
Busby, E., Xia, J., Wu, Q., Low, J. Z., Song, R., Miller, J. R., Zhu, X.-Y., Campos, L. M. & Sfeir, M. Y. (2015). *Nat. Mater.* **14**, 426–433.
Cudazzo, P., Gatti, M. & Rubio, A. (2012). *Phys. Rev. B*, **86**, 195307.
Cudazzo, P., Gatti, M., Rubio, A. & Sottile, F. (2013). *Phys. Rev. B*, **88**, 195152.
Cudazzo, P., Sottile, F., Rubio, A. & Gatti, M. (2015). *J. Phys. Condens. Matter*, **27**, 113204.
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. (2009). *Acta Cryst.* B**65**, 107–125.
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. (2005). *Acta Cryst.* B**61**, 511–527.
Deslippe, J., Samsonidze, G., Jain, M., Cohen, M. L. & Louie, S. G. (2013). *Phys. Rev. B*, **87**, 165124.
Deslippe, J., Samsonidze, G., Strubbe, D. A., Jain, M., Cohen, M. L. & Louie, S. G. (2012). *Comput. Phys. Commun.* **183**, 1269–1289.
Diao, Y., Tee, B. C.-K., Giri, G., Xu, J., Kim, D. H., Becerril, H. A., Stoltenberg, R. M., Lee, T. H., Xue, G., Mannsfeld, S. C. B. & Bao, Z. (2013). *Nat. Mater.* **12**, 665–671.
Dillon, R. J., Piland, G. B. & Bardeen, C. J. (2013). *J. Am. Chem. Soc.* **135**, 17278–17281.
Dollase, W. A. (1965). *J. Am. Chem. Soc.* **87**, 979–982.
Fonari, A., Sutton, C., Brédas, J.-L. & Coropceanu, V. (2014). *Phys. Rev. B*, **90**, 165205.
Fuchs, M. & Scheffler, M. (1999). *Comput. Phys. Commun.* **119**, 67–98.
Giannozzi, P., Baroni, S., Bonini, N., Calandra, M., Car, R., Cavazzoni, C., Ceresoli, D., Chiarotti, G. L., Cococcioni, M., Dabo, I., Dal Corso, A., de Gironcoli, S., Fabris, S., Fratesi, G., Gebauer, R., Gerstmann, U., Gougoussis, C., Kokalj, A., Lazzeri, M., Martin-Samos, L., Marzari, N., Mauri, F., Mazzarello, R., Paolini, S., Pasquarello, A., Paulatto, L., Sbraccia, C., Scandolo, S., Sclauzero, G., Seitsonen, A. P., Smogunov, A., Umari, P. & Wentzcovitch, R. M. (2009). *J. Phys. Condens. Matter*, **21**, 395502.
Giri, G., Verploegen, E., Mannsfeld, S. C. B., Atahan-Evrenk, S., Kim, D. H., Lee, S. Y., Becerril, H. A., Aspuru-Guzik, A., Toney, M. F. & Bao, Z. (2011). *Nature*, **480**, 504–508.
Hasegawa, T. & Takeya, J. (2009). *Sci. Technol. Adv. Mater.* **10**, 024314.
Hatcher, P. V., Reibenspies, J. H., Haddon, R. C., Li, D., Lopez, N. & Chi, X. (2015). *CrystEngComm*, **17**, 4172–4178.

Hathwar, V. R., Sist, M., Jørgensen, M. R. V., Mamakhel, A. H., Wang, X., Hoffmann, C. M., Sugimoto, K., Overgaard, J. & Iversen, B. B. (2015). *IUCrJ*, **2**, 563–574.
Hilfiker, R. (2006). Editor. *Polymorphism: In the Pharmaceutical Industry*. New York: Wiley-VCH Verlag.
Hulea, I. N., Fratini, S., Xie, H., Mulder, C. L., Iossad, N. N., Rastelli, G., Ciuchi, S. & Morpurgo, A. F. (2006). *Nat. Mater.* **5**, 982–986.
Hummer, K. & Ambrosch-Draxl, C. (2005). *Phys. Rev. B*, **71**, 081202.
Hummer, K., Ambrosch-Draxl, C., Bussi, G., Ruini, A., Caldas, M., Molinari, E., Laskowski, R. & Christensen, N. (2005). *Phys. Status Solidi. (b)*, **242**, 1754–1758.
Hummer, K., Puschnig, P. & Ambrosch-Draxl, C. (2004). *Phys. Rev. Lett.* **92**, 147402.
Jiang, L., Dong, H. & Hu, W. (2010). *J. Mater. Chem.* **20**, 4994–5007.
Johnson, E. R., Keinan, S., Mori-Sánchez, P., Contreras-García, J., Cohen, A. J. & Yang, W. (2010). *J. Am. Chem. Soc.* **132**, 6498–6506.
Klein, R. A. (2006). *Chem. Phys. Lett.* **425**, 128–133.
Kolata, K., Breuer, T., Witte, G. & Chatterjee, S. (2014). *ACS Nano*, **8**, 7377–7383.
Lahav, M. & Leiserowitz, L. (2015). *Phys. Scr.* **90**, 118003.
Li, Y., Ji, D., Liu, J., Yao, Y., Fu, X., Zhu, W., Xu, C., Dong, H., Li, J. & Hu, W. (2015). *Sci. Rep.* **5**, 13195.
Li, L.-H., Kontsevoi, O. Y. & Freeman, A. J. (2014). *Phys. Rev. B*, **90**, 195203.
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., van Eijck, B. P., Verwer, P. & Williams, D. E. (2000). *Acta Cryst.* B**56**, 697–714.
Marom, N., DiStasio, R. A., Atalla, V., Levchenko, S., Reilly, A. M., Chelikowsky, J. R., Leiserowitz, L. & Tkatchenko, A. (2013). *Angew. Chem. Int. Ed.* **52**, 6629–6632.
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., van Eijck, B. P., Verwer, P. & Williams, D. E. (2002). *Acta Cryst.* B**58**, 647–661.
Nyman, J. & Day, G. M. (2015). *CrystEngComm*, **17**, 5154–5165.
Otero-de-la-Roza, A., Blanco, M., Pendás, A. M. & Luaña, V. (2009). *Comput. Phys. Commun.* **180**, 157–166.
Otero-de-la-Roza, A., Johnson, E. R. & Contreras-García, J. (2012). *Phys. Chem. Chem. Phys.* **14**, 12165–12172.
Otero-de-la-Roza, A., Johnson, E. R. & Luaña, V. (2014). *Comput. Phys. Commun.* **185**, 1007–1018.
Perdew, J. P., Burke, K. & Ernzerhof, M. (1996). *Phys. Rev. Lett.* **77**, 3865–3868.
Perdew, J. P., Burke, K. & Ernzerhof, M. (1997). *Phys. Rev. Lett.* **78**, 1396.
Perdew, J. P., Ernzerhof, M. & Burke, K. (1996). *J. Chem. Phys.* **105**, 9982–9985.
Perdew, J. P. & Zunger, A. (1981). *Phys. Rev. B*, **23**, 5048–5079.
Pfattner, R., Mas-Torrent, M., Bilotti, I., Brillante, A., Milita, S., Liscio, F., Biscarini, F., Marszalek, T., Ulanski, J., Nosal, A., Gazicki-Lipman, M., Leufgen, M., Schmidt, G., Molenkamp, L. W., Laukhin, V., Veciana, J. & Rovira, C. (2010). *Adv. Mater.* **22**, 4198–4203.
Price, S. L. (2014). *Chem. Soc. Rev.* **43**, 2098–2111.
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. l., 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., Vázquez-Mayagoitia, Á., Vogt, L., Wang, Y., Watson, R. E., de Wijs, G. A., Yang, J., Zhu, Q. & Groom, C. R. (2016). *Acta Cryst.* B**72**, 439–459.
Reilly, A. M. & Tkatchenko, A. (2013*a*). *J. Phys. Chem. Lett.* **4**, 1028–1033.
Reilly, A. M. & Tkatchenko, A. (2013*b*). *J. Chem. Phys.* **139**, 024705.
Reilly, A. M. & Tkatchenko, A. (2014). *Phys. Rev. Lett.* **113**, 055701.
Rohlfing, M. & Louie, S. G. (2000). *Phys. Rev. B*, **62**, 4927–4944.
Sai, N., Tiago, M. L., Chelikowsky, J. R. & Reboredo, F. A. (2008). *Phys. Rev. B*, **77**, 161306.
Salzmann, I., Moser, A., Oehzelt, M., Breuer, T., Feng, X., Juang, Z.-Y., Nabok, D., Della Valle, R. G., Duhm, S., Heimel, G., Brillante, A., Venuti, E., Bilotti, I., Christodoulou, C., Frisch, J., Puschnig, P., Draxl, C., Witte, G., Müllen, K. & Koch, N. (2012). *ACS Nano*, **6**, 10874–10883.
Samsonidze, G., Ribeiro, F. J., Cohen, M. L. & Louie, S. G. (2014). *Phys. Rev. B*, **90**, 035123.
Santra, B., Klimeš, J., Tkatchenko, A., Alfè, D., Slater, B., Michaelides, A., Car, R. & Scheffler, M. (2013). *J. Chem. Phys.* **139**, 154702.
Schiemenz, G. P. (2007). *Z. Naturforsch. B Chem. Sci.* **62**, 235–243.
Setyawan, W. & Curtarolo, S. (2010). *Comput. Mater. Sci.* **49**, 299–312.
Sharifzadeh, S., Biller, A., Kronik, L. & Neaton, J. B. (2012). *Phys. Rev. B*, **85**, 125307.
Sharifzadeh, S., Darancet, P., Kronik, L. & Neaton, J. B. (2013). *J. Phys. Chem. Lett.* **4**, 2197–2201.
Sharifzadeh, S., Wong, C. Y., Wu, H., Cotts, B. L., Kronik, L., Ginsberg, N. S. & Neaton, J. B. (2015). *Adv. Funct. Mater.* **25**, 2038–2046.
Shockley, W. & Queisser, H. J. (1961). *J. Appl. Phys.* **32**, 510–519.
Silva Filho, D. A. da, Kim, E.-G. & Brédas, J.-L. (2005). *Adv. Mater.* **17**, 1072–1076.
Simmons, H. E., Vest, R. D., Blomstrom, D. C., Roland, J. R. & Cairns, T. L. (1962). *J. Am. Chem. Soc.* **84**, 4746–4756.
Sun, C. C. (2009). *J. Pharm. Sci.* **98**, 1671–1687.
Sundar, V. C., Zaumseil, J., Podzorov, V., Menard, E., Willett, R. L., Someya, T., Gershenson, M. E. & Rogers, J. A. (2004). *Science*, **303**, 1644–1646.
Teichen, P. E. & Eaves, J. D. (2015). *J. Chem. Phys.* **143**, 044118.
Tkatchenko, A., DiStasio, R. A., Car, R. & Scheffler, M. (2012). *Phys. Rev. Lett.* **108**, 236402.
Tkatchenko, A. & Scheffler, M. (2009). *Phys. Rev. Lett.* **102**, 073005.
Troullier, N. & Martins, J. L. (1991). *Phys. Rev. B*, **43**, 1993–2006.
Tseng, R., Chan, R., Tung, V. & Yang, Y. (2008). *Adv. Mater.* **20**, 435–438.
Wang, X., Garcia, T., Monaco, S., Schatschneider, B. & Marom, N. (2016). *CrystEngComm*, doi: 10.1039/C6CE00873A.
Wang, M., Li, J., Zhao, G., Wu, Q., Huang, Y., Hu, W., Gao, X., Li, H. & Zhu, D. (2013). *Adv. Mater.* **25**, 2229–2233.
Webster, O. W. (2002). *J. Polym. Sci. A Polym. Chem.* **40**, 210–221.
Yassar, A. (2014). *Polym. Sci. Ser. C*, **56**, 4–19.
Zhu, L., Yi, Y., Fonari, A., Corbin, N. S., Coropceanu, V. & Brédas, J.-L. (2014). *J. Phys. Chem. C*, **118**, 14150–14156.

3.3 Accepted Manuscript: Genarris: Random Generation of Molecular Crystal Structures and Fast Screening with a Harris Approximation

This recently accepted manuscript details the methodology and applications of a general purpose molecular crystal structure generation package called Genarris, written in Python. Genarris is used to generate the initial pool for the GAtor genetic algorithm package (see Section 3.4). My contributions to this work include supervising the algorithm design strategy from the initial to the final stages of development. I formulated our approach for including unsupervised learning into the structure generation workflows in order to increase the diversity of the structures generated. I contributed to the writing of the GA methodology, Harris Approximation, and GA analysis sections of the manuscript.

Genarris: Random Generation of Molecular Crystal Structures and Fast Screening with a Harris Approximation

Xiayue Li[1,2], Farren S. Curtis[3], Timothy Rose,[1] Christoph Schober[4], Alvaro Vazquez-Mayagoitia[5], Karsten Reuter[4], Harald Oberhofer[4], Noa Marom[1,3,6*]

[1]*Department of Materials Science and Engineering, Carnegie Mellon University, Pittsburgh, PA 15213, USA*
[2]*Google Inc., Mountain View, CA 94030, USA.*
[3]*Department of Physics, Carnegie Mellon University, Pittsburgh, PA 15213, USA*
[4]*Chair for Theoretical Chemistry and Catalysis Research Center, Technische Universiät München, Lichtenbergstr. 4, D-85747 Garching, Germany*
[5]*Argonne Leadership Computing Facility, Argonne National Lab, Lemont, IL 60439, USA.*
[6]*Department of Chemistry, Carnegie Mellon University, Pittsburgh, PA 15213, USA*
**Email: nmarom@andrew.cmu.edu*

Abstract

We present Genarris, a Python package that performs configuration space screening for molecular crystals of rigid molecules by random sampling with physical constraints. For fast energy evaluations Genarris employs a Harris approximation, whereby the total density of a molecular crystal is constructed via superposition of single molecule densities. Dispersion-inclusive density functional theory (DFT) is then used for the Harris density without performing a self-consistency cycle. Genarris uses machine learning for clustering, based on a relative coordinate descriptor (RCD) developed specifically for molecular crystals, which is shown to be robust in identifying packing motif similarity. In addition to random structure generation, Genarris offers three workflows based on different sequences of successive clustering and selection steps: the "Rigorous" workflow is an exhaustive exploration of the potential energy landscape, the "Energy" workflow produces a set of low energy structures, and the "Diverse" workflow produces a maximally diverse set of structures. The latter is recommended for generating initial populations for genetic algorithms. Here, the implementation of Genarris is reported and its application is demonstrated for three test cases.

1. Introduction

Understanding the solid-state behavior of molecules may inform the design of crystal forms with desired properties for target applications. Traditionally a prime interest of the pharmaceutical industry, molecular crystals also have applications in diverse areas such as solar cells,[1] organic light emitting diodes (OLEDs),[2] and porous materials for gas storage and catalysis.[3,4] Molecular crystals often display polymorphism, the ability of a molecule to crystallize in more than one structure.[5–7] Polymorphs of pharmaceuticals may exhibit significantly different physical and chemical properties such as stability, solubility, and processability.[5,8,9] For organic semiconductors, different polymorphs may display different band structures, optoelectronic properties, and electron–phonon couplings.[10–15]

Crystal structure prediction (CSP) is a grand challenge for the computational condensed

matter community because it requires screening a large number of candidate crystal structures with high accuracy.[16–20] Sampling the configuration space for a given molecule is enormously complex, as one must consider a range of all possible space groups, lattice parameters, values of Z (the number of asymmetric units related by symmetry in the unit cell) and Z' (the number of molecules in the asymmetric unit), molecular orientations, and conformations. Furthermore, weak van der Waals interactions in molecular crystals lead to many local minima that are extremely close in energy, requiring energy resolution of a few meV for accurate ranking of polymorphs.[6,21–25] The progress of the field has been has been periodically assessed by CSP blind tests, organized by the Cambridge Crystallographic Data Centre (CCDC).[26–31] Over the course of six blind tests, spanning nearly two decades, several best practices have emerged for the generation and ranking of molecular crystal structures.

For ranking of putative structures, hierarchical screening approaches are often used, where successive steps employ increasingly accurate energy methods for smaller subsets of structures. Generic force fields have consistently been demonstrated to produce poor results in crystal structure prediction.[29–31] Tailor-made, system-specific force fields parameterized based on *ab initio* calculations have proven more reliable. Dispersion-inclusive density functional theory (DFT) has become the de facto standard for the final ranking of structures.[31] The many-body dispersion (MBD) method, in particular when combined with hybrid DFT functionals, has been shown to be highly accurate.[23,31–35] Fully *ab initio* calculations, however, are too computationally expensive for fast initial screening of a large number of structures. Parameterization or machine learning of tailor-made system specific interatomic potentials may also require a significant number of first principles calculations.

The Harris approximation (HA)[36,37] is a transferable first principles approach with a moderate computational cost that offers a compromise between the efficiency of empirical force fields and the accuracy of *ab initio* DFT calculations. Within the HA, the total density of a system is constructed by superposition of its fragment densities. The DFT total energy is then calculated for the Harris density without performing a self-consistent cycle.[36–38] The HA has been shown to perform well for weakly interacting molecular dimers, where there is no electron density overlap and no significant polarization.[38] To the best of our knowledge, here the HA is used for molecular crystal configuration space screening for the first time.

Random sampling of the configuration space is widely used in the structure generation process.[29,30,39–41] While some of the early pioneers of CSP used purely random or grid searches,[39,40,42] quasi-random sampling using low-discrepancy Sobol sequences provides a more uniform coverage.[31,43–45] Random sampling is often constrained by symmetry, stoichiometry, knowledge of the chemical system, and experimental data.[20,31] Random sampling frequently precedes or is incorporated into more advanced search algorithms,[31] such as genetic algorithms (GAs),[46–48] swarm algorithms, and Bayesian optimization. Several CSP methods rely on random structural modifications, including simulated annealing,[49,50] parallel tempering,[51] and basin hopping.[52–54] Random sampling is often combined with clustering methods to monitor the sampling convergence, as in the conformation family Monte Carlo method[55] and other quasi-random sampling techniques.[42,56]

Recently, data driven approaches, such as machine learning (ML) algorithms have been increasingly employed in computational chemistry and materials science in conjunction with first principles simulations,[57,58] in various capacities, including predicting a material's structure[59–61] and properties,[62–74] generating interatomic potentials[75–81] and DFT functionals,[82] improved sampling,[83–85] revealing structure-property correlations,[86–88] and finding predictive descriptors.[89–

[92] We expect ML to be featured heavily in the next CSP blind test. In particular, best practices for configuration space screening may benefit from using ML to perform (dis)similarity analysis while effectively capturing the similarity and diversity of crystal packing motifs. To this end, one widely used descriptor is the radial distribution function (RDF).[40,56,93] Other descriptors are based on a series of interatomic distances representing specific close intermolecular contacts.[41,56] Both of these descriptors are based on atomic positions. To capture the packing motifs of molecular crystals, we introduce a new relative coordinate descriptor (RCD), based on the relative positions and orientations of neighboring molecules.

Genarris is a Python package that currently performs configuration space screening for crystals of rigid molecules. It is available for download from software.noamarom.com under a BSD3 license. The purpose of Genarris is not necessarily to seek the ultimate convergence of the search (i.e. the global minimum structure), but rather to provide a computationally efficient way of generating a diverse set of reasonable structures that span the potential energy landscape. Genarris was originally developed in order to produce an initial population for the GAtor genetic algorithm package.[48] However, it may be applied more broadly to generate structure sets for any other search algorithm, for fitting system specific interatomic potentials, or for training machine learning algorithms. Genarris generates random structures with physical constraints imposed on symmetry, unit cell parameters, and intermolecular close contacts. The HA is then used for fast energy evaluations. Once a large "raw" pool of random structures has been generated, Genarris offers three standard workflows for further refinement. The "Energy" workflow selects for low energy structures. The "Diverse" workflow favors structural diversity over energetic stability. The "Rigorous" workflow involves hierarchical screening of structures and is essentially a CSP method in and of itself. All workflows incorporate ML using RCD-based clustering. The user may choose the most appropriate workflow, depending on their needs and computational resources. In the following, we report the implementation of Genarris, validate the reliability of the HA and the effectiveness of RCD-based affinity propagation (AP) clustering, and demonstrate the performance of Genarris for configuration space screening of three past CSP blind test targets, shown in Figure 1.

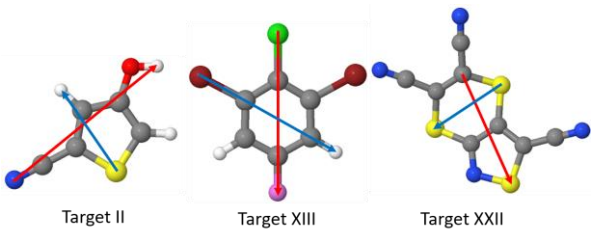


Figure 1: Geometries of the molecules studied here. C atoms are colored in gray, H in white, N a in blue, S in yellow, O in red, Br in dark red, F in pink, and Cl in green. The red and blue arrows indicate the reference axes used to construct the relative coordinate descriptor (RCD), as described in Section 2.3.1.

2. Methods

2.1. Structure Generation

2.1.1. Molecule 3D Coordinates

Genarris takes as input the 3D coordinates of a single molecule. These may be generated by any means. Here, the ChemDraw software is used to obtain an estimate of the molecule's 3D atomic coordinates out of a 2D stick diagram. DFT geometry optimization is then performed using the FHI-aims electronic structure code,[94] with the Perdew-Burke-Ernzerhof (PBE)[95,96] generalized gradient approximation and the Tkatchenko-Scheffler (TS) pairwise dispersion correction.[97] *Higher-level* numerical settings are used, which correspond to the tight/tier 2 settings of FHI-aims.

2.1.2. Unit Cell Generation

Unit cell generation is initialized by obtaining an estimate of the volume of a unit cell with a fixed number of molecules. 10 random structures are generated with a fixed, overestimated volume. Full unit cell relaxation is then performed using PBE+TS with *lower-level* numerical settings, which correspond to the light/tier 1 settings of FHI-aims.[94] The following parameters are used to accelerate the calculation: the k-grid is set to $2 \times 2 \times 2$, the self-consistent accuracy of eigenvalue sum is set to 0.01, and the self-consistent accuracy of forces is not checked. The smallest relaxed volume out of the set of trial structures is taken as an initial volume estimate, denoted hereafter as u.

Genarris uses the standard space group symmetry definitions provided by Bilbao Crystallographic Server.[98] Once the user specifies the number of molecules per cell, Genarris identifies the compatible space groups with matching general Wyckoff position multiplicity. The user may optionally specify which space group(s) to use, or request the use of only chiral space groups. Additionally, special Wyckoff positions may be requested. To generate a structure, Genarris randomly picks one of the compatible or user-defined space groups.

After the space group of the random structure is determined, the lattice vectors are constructed according to the designated Bravais system. The unit cell volume may be fixed, or sampled randomly using a half-normal distribution curve with a user-defined center (i.e. the lower bound), standard deviation, and upper bound. By default, $0.9u$ is chosen as the lower bound, $0.1u$ as the standard deviation, and $1.1u$ as the upper bound. A half-normal distribution curve is used to bias towards the distribution center as a lower bound. Genarris uses this design because the random placement of molecules in a unit cell with smaller volume is more difficult due to constraints imposed by close contacts. Uniform sampling within a user-defined range of unit cell volumes is also implemented in Genarris.

The unit cell orientation is standardized, such that the lattice vectors, $\vec{a} = (a_x, a_y, a_z)^T$, $\vec{b} = (b_x, b_y, b_z)^T$, $\vec{c} = (c_x, c_y, c_z)^T$, form an upper triangular matrix (a_y=a_z=b_z=0). The cell volume, v, is then given by the product of the principal components of each lattice vector: $v = a_x b_y c_z$. The user may control the cell shape by constraining the ratio between each of the principal components and the cube root of v. Genarris constructs the lattice vectors by randomly generating the principal components (whose product is equal to v) within the user-defined range. When the cell angles, α, β and γ, are not constrained by the Bravais system, Genarris randomly generates them from 30 to 150 degrees by default, or in a user defined range. Given the six parameters ($a_x, b_y, c_z, \alpha, \beta, and\ \gamma$), the unit cell is now uniquely-defined. Genarris then solves for the additional cell parameters through equations S1-S6, provided in the supplementary material.

By first ensuring reasonable principal components and then solving for the other cell parameters, Genarris effectively samples cells that are not too compressed in one direction. This leads to a higher success rate in molecule placement for skewed cell configurations, and thus increases the uniformity of sampling. This is especially important for exploring the alternative space group settings not recorded in the standard library currently implemented in Genarris. For example, the space group setting $P2_1/n$ is an alternative setting to the common space group $P2_1/c$. Expressing a structure of space group $P2_1/n$ in the $P2_1/c$ setting requires a matrix transformation of the lattice vectors, which tends to result in very oblique structures. Failure to account for this obliqueness was the reason the experimental structure of target XXII was not found with the preliminary version of our code used in the sixth blind test.[31]

2.1.3. Molecule Placement

Genarris places the molecule in the asymmetric unit by giving it a random orientation and then selecting a random center of mass (COM) position. The random orientation is sampled uniformly by choosing a random rotation axis on a unit sphere (see equations S7-S10 in the supplementary material). The random rotation matrix is then applied to the molecule with its COM fixed at the origin. The COM is then moved to a random position by uniform random sampling between 0 and 1 for each dimension of the fractional coordinates. Once the asymmetric unit is constructed, the chosen space group symmetry is applied to obtain the atomic coordinates of the remaining molecules in the unit cell.

After a structure is randomly generated, a closeness check is performed to avoid unphysical close contacts. Structures that fail the closeness check are rejected. Two types of closeness checks are implemented in Genarris, a COM distance check and an intermolecular atomic distance check. The latter guarantees that no two atoms belonging to different molecules are closer than a user-defined threshold, which may be set as a constant or specific to the atomic species. The user may define a custom radius for each atom type or use the default setting of the van der Waals radii.[99] The parameter s_r is a user-defined fraction of the sum of two atomic radii, such that the distance between the two atoms of different molecules cannot be smaller than $(r_1 + r_2) \times s_r$. The value of s_r should be large enough to avoid unphysical structures (this is particularly important for the reliability of the HA, as discussed below) and small enough to allow for a diversity of crystal packing motifs. Genarris uses a fuzzy s_r setting to increase pool diversity. s_r is randomly selected at each structure generation attempt with a half-normal distribution, defined by an upper bound, standard deviation and a lower bound. The default values used here are 0.9, 0.05 and 0.8, respectively (these choices are motivated by the performance of the HA as shown in Section 4.1 below).

2.2. Fast Screening with the Harris Approximation (HA)

Within the Harris approximation,[36] the total density of a system is constructed by superposition of self-consistent fragment densities (in general, the fragments may be atoms, groups of atoms, or molecules). The DFT total energy may then be evaluated for the Harris density without performing a self-consistent cycle, providing very fast energy evaluations. This has been demonstrated as a reasonable approximation for the treatment dimers of weakly interacting molecules with dispersion-inclusive DFT in the van der Waals regime, where there is no significant density overlap or polarization.[38,100,101] Genarris uses the HA to construct the density of a molecular crystal by replicating, translating, and rotating the self-consistent density of a single molecule, which is calculated only once. This enables fast screening of initial structures using an unbiased first-principles DFT@Harris approach without resorting to force fields, which can be highly inaccurate and difficult to parametrize for atypical molecules.

To this end, we have implemented the Harris approximation in FHI-aims.[102] Others have reported similar implementations for plane-wave[38] and Gaussian[100,101] basis sets. The numeric atom-centered orbital (NAO) basis functions of FHI-aims are based on real valued linear combinations of spherical harmonics.[94] Because the spherical harmonics are fixed with respect to the *xyz*-coordinate system, rotation of a molecule produces a new linear combination of basis functions. Modified Wigner matrices[103] are employed to obtain the rotated coefficients of each basis function (a detailed account is provided in the supplementary material). The present implementation is restricted to Γ-point calculations of crystals of rigid molecules. The HA may be used in conjunction with any DFT functional and dispersion method. Here, PBE+TS@Harris

is used for fast screening purposes. The same method was employed in the preliminary version of Genarris, used within the sixth CSP blind test.

2.3. Structure Clustering

2.3.1. Radial Distribution Function (RDF) and Relative Coordinate Descriptor (RCD)

Recently, there has been significant progress in formulating descriptors of molecular systems for ML purposes, such as the Coulomb matrix and the Bag of Bonds method.[62,64,104,105] Descriptors based on interatomic distances, such as pair correlation functions or distances between specific atoms are still commonly used for molecular crystals.[40,41,56,106] One such descriptor, the radial distribution function (RDF), is implemented in Genarris.[40,106] For this descriptor, the user inputs an element pair (*X*, *Y*). The RDF *G* between *X* and *Y* is defined as:

$$G_{XY}(r) = \frac{\sum_{i,j} \exp\left(-B(r-r_{ij})^2\right)}{N_X}, \quad (1)$$

where *i* and *j* run over *X* and *Y* atoms, and N_X is the number of *X* atoms. The RDF (which is a continuous function) is then sampled at a list of user-defined distance bins to form a vector descriptor. Multiple vectors of different element pairs can be concatenated to form a single RDF descriptor.

In addition to this atomic-level descriptor, we have developed the relative coordinate descriptor (RCD), intended for capturing how the molecules are positioned and oriented with respect to one another. The RCD is constructed by selecting a representative molecule and the *N* molecules with closest COM positions. *N* should be sufficiently large to correctly capture the environment of a molecule in a crystal. The default value is 16. Then, a frame of reference is constructed for each molecule. Two of the axes are vectors pointing from one fixed atom in the molecule to another (defined by user input), orthogonalized and normalized using a Gram-Schmidt procedure. The axes used here for the three targets are shown in Figure 1. The third axis is calculated as the cross product of the two user-defined axes. The relative positions are obtained by calculating the Euclidean distances between the COM positions of each of the surrounding molecules and the representative molecule and expressing them in the basis of the representative's reference frame. The relative orientations are obtained by taking the dot product between each of the three reference axes of a neighboring molecule with those of the representative molecule. The RCD of a crystal is then defined as

$$\vec{R} = \{(\overrightarrow{P^1}, \overrightarrow{Q^1}), \ldots, (\overrightarrow{P^N}, \overrightarrow{Q^N})\}, \quad (2)$$

where $\overrightarrow{P^i}$ and $\overrightarrow{Q^i}$are, respectively, the 3-dimensional relative position and relative orientation of the i^{th} neighboring molecule with respect to the representative.

To compare two RCD vectors of different crystal structures, $\overrightarrow{R_1}$ and $\overrightarrow{R_2}$, an $N \times N$ matrix, $\boldsymbol{D}$, is constructed as

$$D_{i,j} = \left(\frac{|\overrightarrow{P_1^i} - \overrightarrow{P_2^j}|^2}{|\overrightarrow{P_1^i}||\overrightarrow{P_2^j}|}\right) + \frac{k}{3}\left(|\overrightarrow{Q_1^i} - \overrightarrow{Q_2^j}|^2\right), \quad (3)$$

where *k* (by default, 1) is a parameter that enables assigning a different weight to the orientation difference and COM position difference, and 1/3 is a normalization factor. Then, the *M* smallest entries of $\boldsymbol{D}$ are selected, such that no two entries have the same *i* index or the same *j* index (For example, one may select $D_{1,3}$ and $D_{3,2}$, but not both $D_{1,3}$ and $D_{1,4}$). *M* is by default 8. The sum of

the *M* entries serves as a measure of the distance between the two RCD vectors. A distance matrix is constructed for a given pool by calculating the RCD difference for all pairs of structures in the pool, using the above procedure.

2.3.2. Affinity Propagation Clustering

In an initial screening workflow, clustering is useful for classifying an existing sample. For example, in the conformation-family Monte Carlo method,[55] clustering is used to monitor the overall convergence of the search. For our initial screening workflows, clustering helps maintain diversity during the selection process (see section 2.4). Genarris uses the affinity propagation (AP) clustering algorithm. While the more widely used *k*-means clustering calculates coordinate averages as cluster centers,[107] AP clustering identifies a refined set of exemplars from the initial data points.[108] This is useful for selecting representative structures from different clusters. AP clustering does not rely on a user-defined number of clusters; rather, the algorithm determines the number of clusters based on a message passing procedure between data points. The procedure is characterized by a preference value for a message to be passed from one data point to another, which can be manipulated to control the number of clusters. The result of AP clustering is consistent, in the sense that it does not depend on a randomized initialization of centers (as in *k*-means), but begins by considering all points as potential exemplars.[108] AP has also been shown to detect clusters with lower average squared distance to cluster center than *k*-centers, a version of *k*-means that similarly outputs exemplars.[108]

Genarris uses AP clustering as implemented in the scikit-learn package.[109] The input of AP clustering is a distance matrix, generated here from the RCD differences between all the structures in the pool, as explained in Section 2.3.1. AP clustering outputs a cluster number for each structure, and assigns to each cluster an exemplar. By adjusting the preference value, Genarris allows the user to request either a fixed number of clusters, or the number of clusters that reaches a target silhouette score, a number between -1 to 1 that determines how well overall the structures fit into their clusters.[110] Accurate, non-overlapping clustering is characterized by a silhouette score greater than zero. A silhouette score of 0.5 or above indicates strong clustering, meaning that the algorithm identifies actual clusters, rather than arbitrarily dividing a continuous region. Once AP clustering is completed, selection procedures are available to either select the exemplars, or the structures with maximum or minimum properties within a cluster (e.g., the lowest energy), as described in Section 2.4. In Section 3.2 it is demonstrated that AP clustering successfully identifies under-sampled clusters, a desirable behavior for the Diverse workflow of Genarris.

2.4. Structure Selection Workflows

We have developed three standard hierarchical structure selection workflows, shown in Figure 2, whereby increasingly accurate methods are used to screen smaller subsets of structures. The workflows comprise different sequences of successive evaluation, clustering, and filtering steps. These workflows represent typical use cases of Genarris. New structure selection workflows for different purposes may be designed by the user as needed. All workflows of Genarris begin with a raw pool generated with user-defined volume range, space group symmetries, and closeness criteria, as described in Section 2.1. By default, each step of the Diverse and Energy workflows reduces the pool to 10% of its previous size. All three workflows reduce the final population of structures to 1% of the raw pool. These structures may either serve directly as candidates for crystal structure prediction, or as an initial sample for a more advanced algorithm. At the end of each workflow, the final converged pool is fully relaxed, checked for duplicates, and re-ranked.

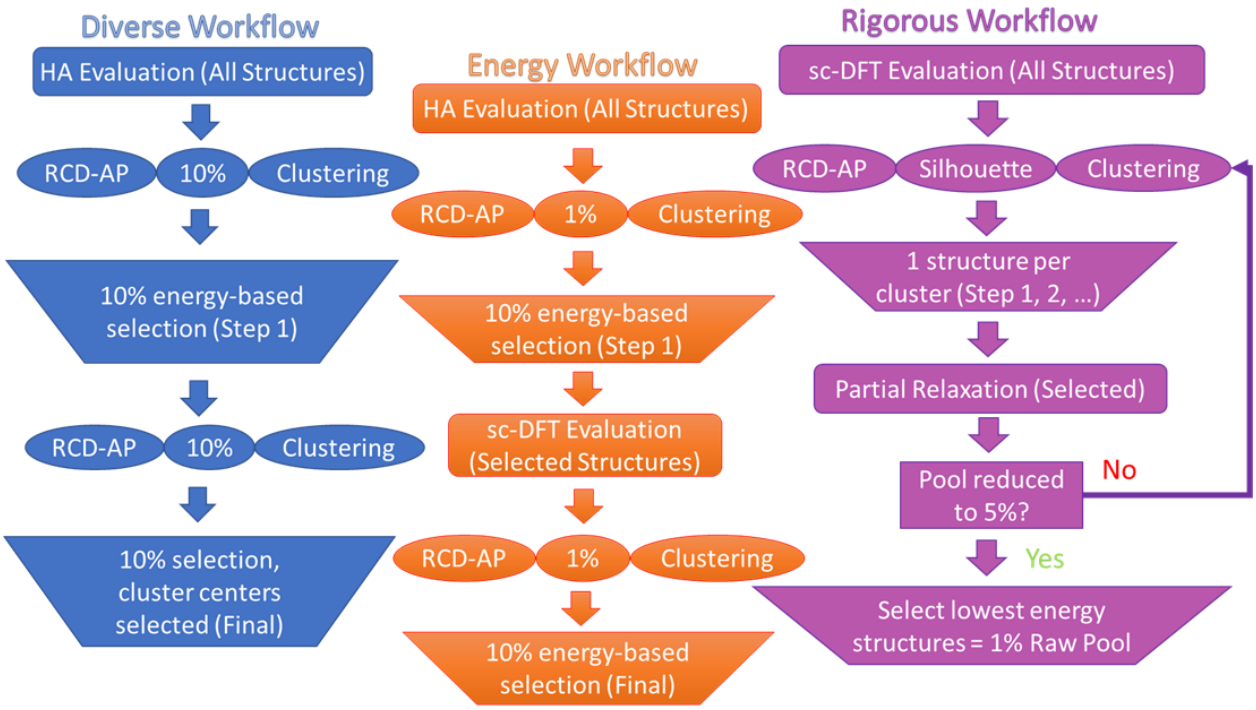


Figure 2: Flow charts of the three screening workflows available in Genarris. RCD-AP clustering indicates AP clustering based on the RCD vector distance matrix. 1%/10% clustering means that the number of clusters is set to 1%/10% of the population. 10% energy-based selection means selecting the 10% of structures with the lowest energy within each cluster. The workflows are presented from left to right by increasing computational cost.

The Diverse workflow is geared towards maximally diverse sampling at a modest computational cost, intended as preparation for an advanced search algorithm. It begins by using the HA to evaluate all the structures in the raw pool. Next, RCD-based AP clustering is performed with the number of clusters set to 10% of the number of structures in the raw pool and the lowest energy structure is selected from each cluster (10% energy-based selection). This ensures the quality of the structures in the pool. Then, RCD-based AP clustering is conducted again with the number of clusters set to 10% of the remaining structures. Lastly, the exemplars chosen by the AP clustering algorithm are selected for the final pool. Because these exemplars represent the center of each cluster, they are expected to be far apart and to provide a maximally diverse sample of the configuration space.

The Energy workflow focuses on targeted sampling of low energy basins of the potential energy surface at a moderate computational cost. It creates fewer clusters than the Diverse workflow in both clustering steps in order to increase intra-cluster energy competition. Employing self-consistent DFT before the final energy-based selection improves the accuracy at the price of a higher computational cost. Like the Diverse workflow, the Energy workflow begins by using the HA to evaluate the energy of all structures in the raw pool. Next, RCD-based AP clustering is performed with the number of clusters set to 1% of the number of structures in the raw pool. The 10 lowest energy structures are selected for single point energy evaluation with FHI-aims, using PBE+TS and minimal numerical settings, where the k-grid is set to $1 \times 1 \times 1$ and the self-consistent accuracy of eigenvalue sum is set to 0.01. Then, RCD-based AP

clustering is conducted with the number of clusters set to 10% of the remaining structures. Lastly, the 10 lowest energy structures in each cluster are selected for the final pool.

The Rigorous workflow is intended for exhaustive sampling of the configuration space and is essentially a standalone crystal structure prediction algorithm, based on hierarchical screening of randomly generated structures with physical constraints. It iteratively refines the pool and reduces its size. Because the Rigorous workflow fully relies on DFT for energy evaluations and structural relaxations, it requires considerable computational resources. The Rigorous workflow begins by performing single point energy evaluations for all the structures in the raw pool using PBE+TS with the *lower-level* numerical settings detailed in Section 2.1.2. RCD-based AP clustering is then performed with the number of clusters adjusted to reach a silhouette score of 0.5. This value corresponds to a midpoint between barely non-overlapping clusters (silhouette score 0) and perfect clustering (silhouette score 1). Empirically, this value can consistently be reached with the number of clusters that provides a reasonable convergence rate (if a score of 0.5 cannot be reached the target score may be adjusted to a lower value). The lowest energy structure from each cluster is selected for full unit cell relaxation using PBE+TS with *lower-level* numerical settings with the number of relaxation steps constrained to 30 by default to reduce the computational cost. Through this partial relaxation, the clusters in the configuration space become more well-defined, such that the RCD-based clustering and selection process more accurately converges to a diverse and low energy post-relaxation pool. The clustering, selection, and relaxation steps are repeated until the pool size is reduced to <5% of the original sample size. At this point, we find that RCD-based clustering begins to fail as the remaining pool becomes too diverse to be reasonably clustered. Therefore, in the final step a purely energy-based selection is performed to reduce the pool size to 1% of the raw pool.

3. Computational Details

Raw pools of 5,000 structures were generated for Target II and Target XIII, and of 10,000 structures for Target XXII (see Figure 1 for molecular structures). The raw pools were constrained to all non-chiral space groups, with Z=4 and Z'=1. These settings correspond to the known experimental structures of the three targets. The initial volume estimates for the three targets were 546, 816, and 988 Å^3, respectively. The lower bound, standard deviation, and upper bound for the half-normal volume sampling (see Section 2.1.2) were respectively, in units of Å^3, (491, 55, 600), (734, 82, 898), and (889, 99, 1098). The lower bound, standard deviation, and upper bound for the half-normal s_r sampling were set to 0.80, 0.05, and 0.90 throughout. COM distance checks were conducted with minimum distances of 4, 4 and 5 Å, respectively. The RCD vectors were generated with 16 closest contacts, with reference axes selected as shown in Figure 1. For the analysis presented in Sec. 4.2.2, the RDF descriptor is calculated using O-N and O-S pairs, with seven 1 Å bins from 2 to 8 Å. For all workflows, the target size of the final pool was set to 1% of the raw pool size (before duplicate screening) i.e., 50, 50, and 100 structures, respectively for Targets II, XIII and XXII. The parameters used for clustering and selection are listed in Table I. For the rigorous workflow, the clustering was performed with a target silhouette score of 0.5 throughout. For the HA used in Diverse and Energy workflow, as well as in the analysis presented in Section 4.1, self-consistent single molecule calculations were performed with PBE+TS light/tier 1 settings, and crystal/dimer HA calculations were conducted with PBE+TS light/tier 1 settings, k-grid of $1 \times 1 \times 1$, and self-consistent iteration limit set to 0.

For each target, the final structures produced using the Random, Diverse, and Energy workflows were used as initial pools for the GAtor genetic algorithm for molecular crystal structure prediction.[48] GAtor starts from an initial population of structures and runs several GA replicas in parallel that perform the core tasks of fitness evaluation, selection, crossover, and mutation while reading from and writing to a dynamically-updated shared population of structures. For each target, the same GA settings were used in order to compare the evolution of the different starting populations. We note that the purpose of these GA runs was not to perform an exhaustive search, for which the recommended best practice is to run GAtor several times with different settings.[48] All local optimizations within GA runs were performed with FHI-aims, using PBE+TS and *lower-level* numerical settings. For Target II, 50% standard crossover and 50% mutation were used with roulette-wheel selection and the energy-based fitness function. The GA was terminated when the common population reached at least 320 structures. For Target XIII, 50% symmetric crossover and 50% mutation were used with roulette-wheel selection and the energy-based fitness function. The GA was terminated when the common population reached at least 1560 structures. For Target XXII, 50% standard crossover and 50% mutation were used with tournament selection and the energy-based fitness function. The GA was terminated when the common population reached at least 650 total structures.

TABLE I. Clustering and selection parameters used here for the Diverse and Energy workflows.

	Target II and XIII		Target XXII		All Targets
Workflow Step	No. of Clusters	No. of Selected Structures	No. of Clusters	No. of Selected Structures	No. of Selected Structures per Cluster
Diverse Step 1	500	500	1000	1000	1
Diverse Step 2	50	50	100	100	1
Energy Step 1	100	500	200	1000	5
Energy Step 2	10	50	20	100	5

4. Results

4.1. Validation of the Harris Approximation

To assess the performance of the HA for chemically diverse species with different types of intermolecular interactions, representative dimers were extracted from the experimental crystal structures of Targets II, XIII, and XXII. The intermolecular distances were varied along the closest O···N, Cl···Cl, and S···N contacts for targets II, XIII, and XXII, respectively. Figure 3 shows binding energy (BE) curves computed with self-consistent PBE+TS (BE_{SCF}) and PBE+TS@Harris (BE_{HA}), as well as the BE error, defined as: $\Delta BE(x) = BE_{HA}(x) - BE_{SCF}(x)$. The Harris density was subtracted from the self-consistent density and the residual is also shown. The HA becomes exact when the molecules are far apart and there is no interaction between them, as indicated by the asymptotic decay of the error to zero. Around the equilibrium distance, $x_{eq.}$, where the intermolecular interactions are still weak, the HA is still found to be sufficiently descriptive: The correct equilibrium distance is obtained and $|\Delta BE(x_{eq.})|$ is fairly small (0.0740, 0.0049, 0.0288 eV for targets II, XIII, and XXII, respectively). As the distance between the molecules decreases and the repulsion between their electron densities becomes significant, the assumption of non-interacting fragment densities breaks down. Because of the non-variational

nature of the HA, $\Delta BE(x)$ is always negative and its magnitude increases asymptotically with decreasing distance. The decent agreement with self-consistent DFT at the equilibrium distance in the BE obtained for all three targets considered here corroborates that the HA is sufficiently quantitatively and (more importantly) qualitatively accurate for fast screening of the initial population of structures. These findings are consistent with earlier reports.[38,100,101,111]

Figure 3 also shows the residual difference between the self-consistent density and the Harris density at the equilibrium distance, $x_{eq.}$. Red (blue) indicates that the self-consistent density is lower (higher) than the Harris density. For Target II, the density difference is concentrated on the O and N atoms of the OH···N close contact, showing that the density difference due to the hydrogen bond is not captured by the HA. The strength of this bond results in a somewhat larger $|\Delta BE(x_{eq.})|$. For Target XIII, the density difference is concentrated on the six-membered ring as well as the Cl and F atoms. In this case, the HA does not capture the change in the density due to the π-π interactions between the aromatic rings and the repulsion between the halogens, which lead to the formation of a dipole with the density shifting from the F side to the Cl side of the molecule. However, the shallow BE curves indicate that these interactions are actually weak in magnitude and thus only a slight $|\Delta BE(x_{eq.})|$ is observed. For Target XXII, the density residuals suggest significant intermolecular dipole-dipole and dipole–induced-dipole interactions due to the highly polarized nitrile groups and intra-ring N atoms resulting in its moderate $|\Delta BE(x_{eq.})|$.

In the following, we further assess the reliability of the HA for energy ranking of randomly generated initial structures. The HA has not been tested in this scenario before. Three initial pools of 2,000 $P2_1/n$ structures were generated for Target XXII, using

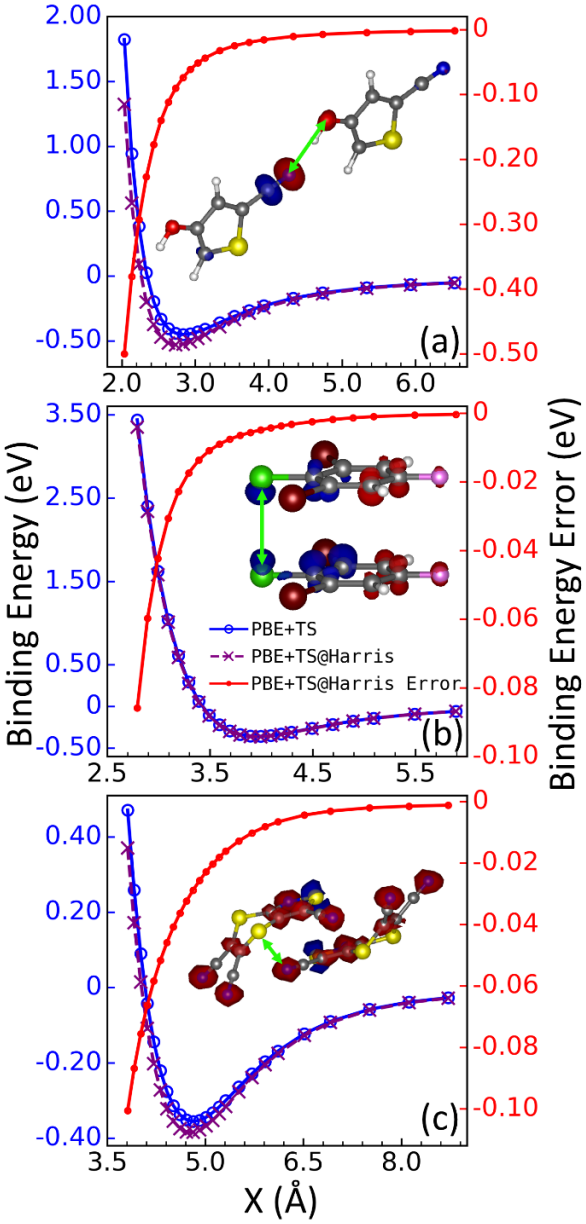


Figure 3: Binding energy curves for dimers of (a) Target II, (b) Target XIII, and (c) Target XXII obtained using PBE+TS@Harris (BE_{HA}) compared to self-consistent PBE+TS (BE_{SCF}), and binding energy error ($BE_{HA}(x)$ - $BE_{SCF}(x)$). The x coordinate corresponds to the intermolecular O···N, Cl···Cl, and S···N distances, indicated by the green arrows. The insets show the density difference between the self-consistent and Harris densities at the equilibrium distance. Red (blue) indicates a negative (positive) density difference.

different closeness criteria with s_r of 0.500, 0.625, and 0.750. Figure 4 compares the performance of PBE+TS@Harris to self-consistent PBE+TS. Panel (a) shows a direct comparison of the BE per molecule and panels (b)-(d) show the ranking based on BE per molecule from low to high. Overall, PBE+TS@Harris shows remarkable agreement with self-consistent PBE+TS for both the BEs and the rankings. The r^2 scores for the BEs are 0.946, 0.960, and 0.994 for s_r=0.500, 0.625, and 0.750, respectively. This is consistent with the above observation for dimers that the accuracy of the HA improves with increasing intermolecular distance, enforced here through a larger s_r value. The r^2 scores for the rankings are 0.976, 0.989, and 0.979 for s_r=0.500, 0.625, and 0.750, respectively. Optimal performance is obtained for s_r=0.625. For s_r=0.500, the performance of the HA is worse due to the presence of more structures with unphysically close intermolecular contacts in the pool. In particular, several of the outliers exhibit unphysically close N···N contacts, which lead to large negative errors in the HA BEs. Three examples are circled in Figure 4 (b) and shown in panel (e). These have N···N distances of 0.164, 0.180 and 0.156 Å. For s_r=0.750 the performance of the HA deteriorates because the structures in the pool are closer in energy than in the s_r=0.500 and s_r=0.625 pools, as shown in panel (a). The accuracy of the HA is insufficient to resolve small energy differences, which leads to more ranking discrepancies.

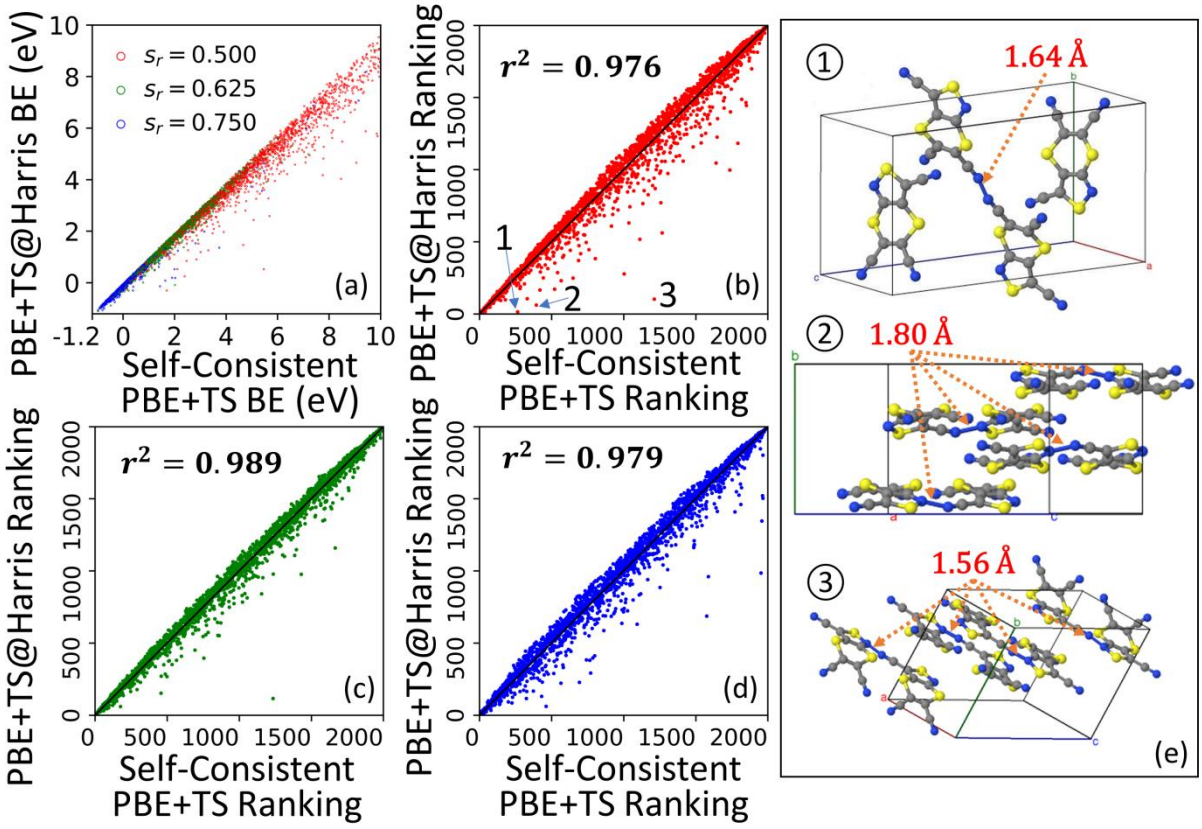


Figure 4: (a) PBE+TS@Harris binding energy vs. self-consistent PBE+TS binding energy; (b)-(d): PBE+TS@Harris ranking vs. self-consistent PBE+TS ranking for sr=0.500, 0.625, and 0.750, respectively. The three significant outliers in (b), labeled as 1, 2 and 3, are illustrated in (e) and their unphysical N···N close contacts are indicated.

4.2. Clustering Analysis

4.2.1. Comparison between *k*-Means and Affinity Propagation clustering

In the workflows of Genarris, AP clustering is used with respect to the RCD, as explained in Section 2.3.2. Here, we illustrate the advantage of AP clustering compared to *k*-means for two dimensional and three dimensional cases, which are easier to visualize than the high dimensional RCD. To highlight the different behavior of the *k*-means and AP clustering algorithms, a set of randomly distributed points were generated within the unit circle. Construction of the data set was initiated from a few anchor points, which simulate low energy basins. Randomly generated points were then accepted or rejected based on their Euclidean distances to one of these anchor points and a random factor. Some of the anchor points had smaller random factors than others, such that fewer points were accepted in their vicinity. The resulting data set is shown in Figure 5, panel (a). The anchor points are shown as larger diamond markers. This dataset is characterized by a large, densely sampled region as well as smaller and separate satellite regions, which simulate narrow disconnected funnels of the potential energy landscape. Ideally, clustering algorithms should assign the satellite regions as distinct clusters. The results of *k*-means and AP, using 15 clusters, are shown in panels (b) and (c), respectively. While *k*-means groups three of the satellite regions together into one cluster, AP successfully identifies them as distinct clusters. This is the behavior desired by Genarris for the purpose of selecting structures from under-sampled regions of the configuration space. The high dimensional configuration space of molecular crystals often has such small clusters that are rarely explored by random sampling, for example because some packing motifs are more difficult to generate. AP clustering can correct this sampling bias by identifying these regions more effectively, provided that an appropriate descriptor is used to resolve structural differences.

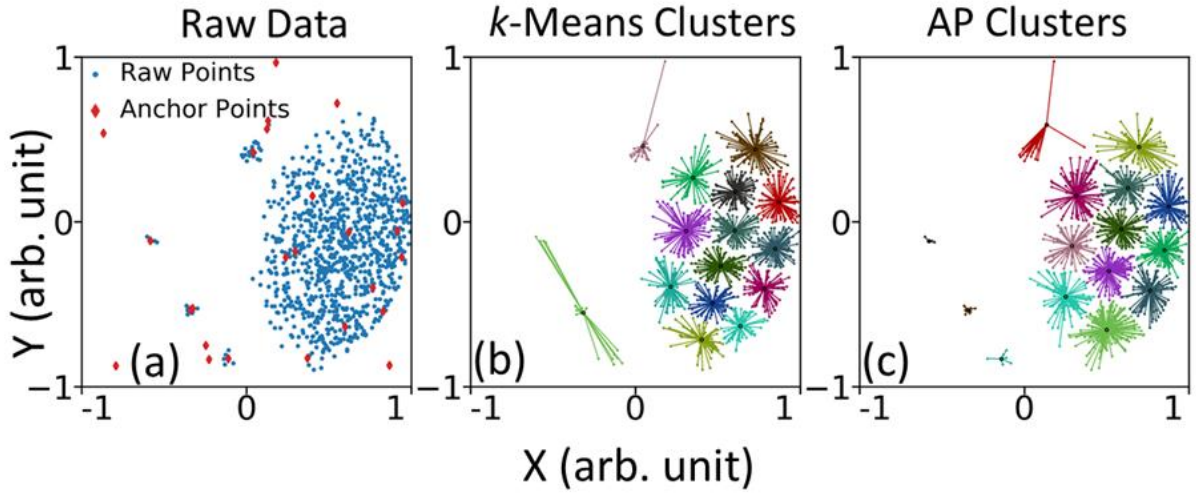


Figure 5: Comparison of the performance of the *k*-means and AP clustering algorithms for randomly generated two-dimensional data with arbitrary units: (a) the raw data with the anchor points colored in red, and 15 clusters as found by (b) *k*-means and (c) AP.

Figure 6 compares the results of *k*-means and AP clustering algorithms in three dimensions, using a descriptor based on lattice parameters for 410 structures of Target XXII generated within the rigorous workflow. The points are grouped into five clusters by the two algorithms. Additionally, each point is colored according to the BE per molecule. The key difference between the two methods is that AP clustering identified a distinct group of structures with a low

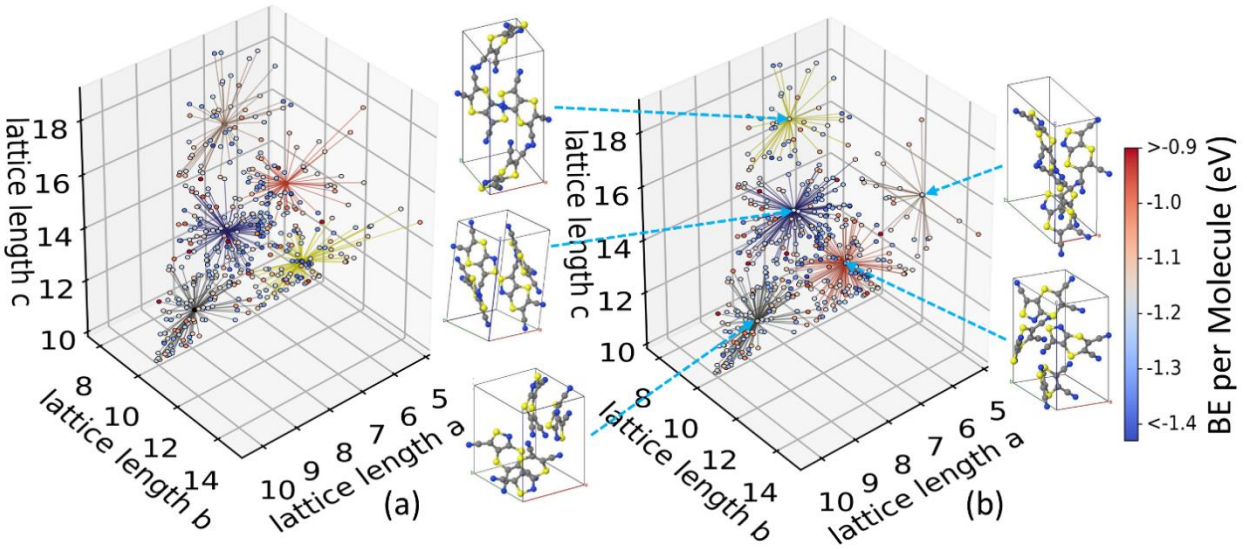


Figure 6: Comparison of the performance of the (a) *k*-means and (b) AP clustering algorithms for 410 structures of Target XXII, clustered into five clusters with respect to a three-dimensional descriptor based on lattice parameters. Each data point is colored according to the structure's BE per molecule. The exemplars found by AP clustering are also shown.

a parameter as a unique cluster. While the majority of the lowest energy structures are concentrated in the center of the graph, the low-*a* cluster contains structures within 0.2 eV from the respective global minimum. Therefore, it should be adequately sampled to ensure overall diversity. By identifying this region as a separate cluster, AP clustering ensures that the structures in this region are better represented in the selected pool.

4.2.2. Comparison between RDF and RCD Descriptors

Figure 7 shows a comparison of clustering based on the RCD to clustering based on an RDF descriptor on the 5,000 random structures in the raw pool of Target II. The RCD and RDF were compared with respect to four performance measures: (a) ability to identify duplicate structures, (b) correlation with space groups, (c) correlation with unit cell volume, and (d) the silhouette score. In order to show that the differences in the clustering performance are due to the descriptor and independent of the clustering method used, both AP and *k*-means were used with the RDF descriptor (*k*-means could not be used with the RCD because its input is a conventional vector descriptor, not a distance matrix).

In the workflows of Genarris full unit cell relaxation is performed only for the final pools of structures. At this point, some structures that are similar but not identical may relax to the same structure and become duplicates. It is desirable for a descriptor to reflect the similarity between such structures, such that they are grouped into the same cluster before relaxation. For Target II, 69 pairs of duplicates were found once the final relaxed pools from the four workflows (Diverse, Energy, Rigorous, and Random) were combined. Panel (a) presents the number of duplicate pairs that were assigned to the same cluster based on their pre-relaxed geometry when the raw pool of 5,000 structures was clustered into 2-10 clusters. As a control, the raw pool was also clustered by randomly assigning a cluster number to each structure. Overall, clustering based on both descriptors significantly increases the predictive grouping of post-relaxation duplicates compared to random assignment. RCD-based clustering had a higher success rate than RDF-based clustering in assigning duplicate pairs to the same cluster. RCD-AP grouped almost all the

duplicate pairs together up to 7 clusters. This helps prevent post-relaxation duplicates by eliminating them earlier in the selection process.

In panels (b)-(d) the raw pool of 5000 Target II structures was clustered into 10, 20, 40, 80, 160 and 320 clusters, based on the RCD and RDF. Panel (b) shows the number of structures whose space group is the same as the mode of its assigned cluster as a function of the number of clusters. RCD-based clustering shows a stronger correlation with space group symmetry than RDF-based clustering, which increases with the number of clusters. This indicates that RCD-based clustering captures packing motifs motifs of molecular crystals, reflected by the space group symmetry, better than RDF-based clustering. Panel (b) shows the average intra-cluster standard deviation of unit cell volume, weighted by the number of structures in each cluster, as a function of the number of clusters. RCD-based clustering has a weaker correlation with the unit cell volume than RDF-based clustering. This trend becomes more pronounced with the number of clusters. This further demonstrates that the RCD is more sensitive to the packing motif, while the RDF is more sensitive to the unit cell volume. As previously described, the silhouette score is a measurement of how well a clustering result identifies unique clusters based on the descriptor vs. clustering a continuous region. Panel (d) shows the silhouette score as a function of the number of clusters. A higher silhouette score indicates better clustering, as explained in Section 2.3.2. RCD-based clustering consistently achieves a significantly higher silhouette score than RDF-based clustering, regardless of the clustering method. Furthermore, the silhouette score for RCD-based clustering generally increases with the number of clusters, while that of RDF-based clustering decreases. This shows that the RCD provides better resolution of clusters in the configuration space. Overall, the RCD provides a superior performance to RDF, as indicated by a higher success rate in identifying duplicate structures, higher sensitivity to packing motifs, and higher silhouette scores.

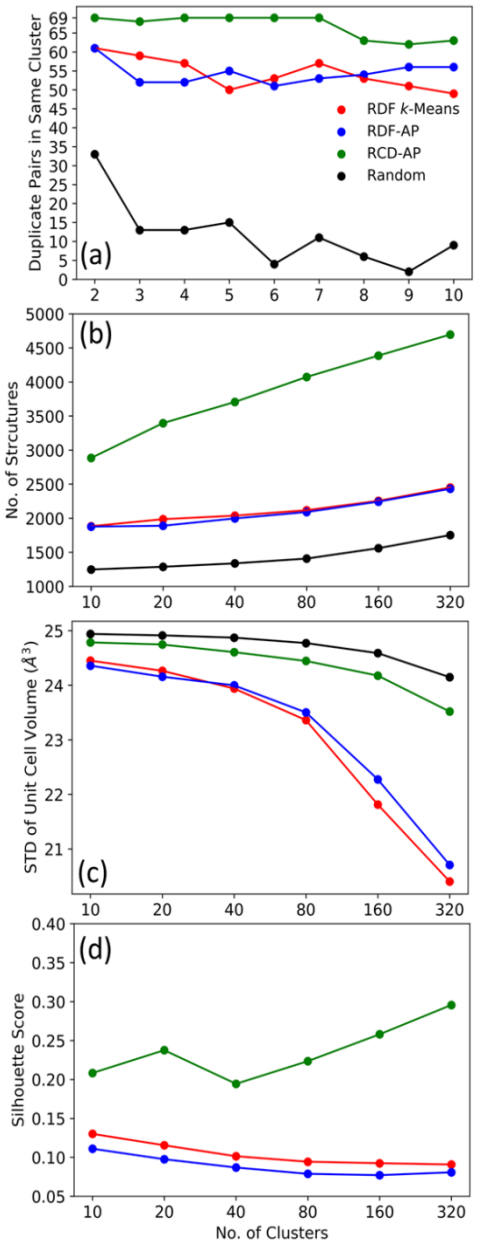


Figure 7: Comparison of RDF-based *k*-means, RDF-based AP, and RCD-based AP clustering with respect to four metrics: (a) ability to identify duplicate structures, (b) correlation with space groups (number of structures whose space group is the same as the mode of their assigned cluster), (c) correlation with unit cell volume (intra-cluster standard deviation of unit cell volume), and (d) the silhouette score.

4.3. Workflow Comparison

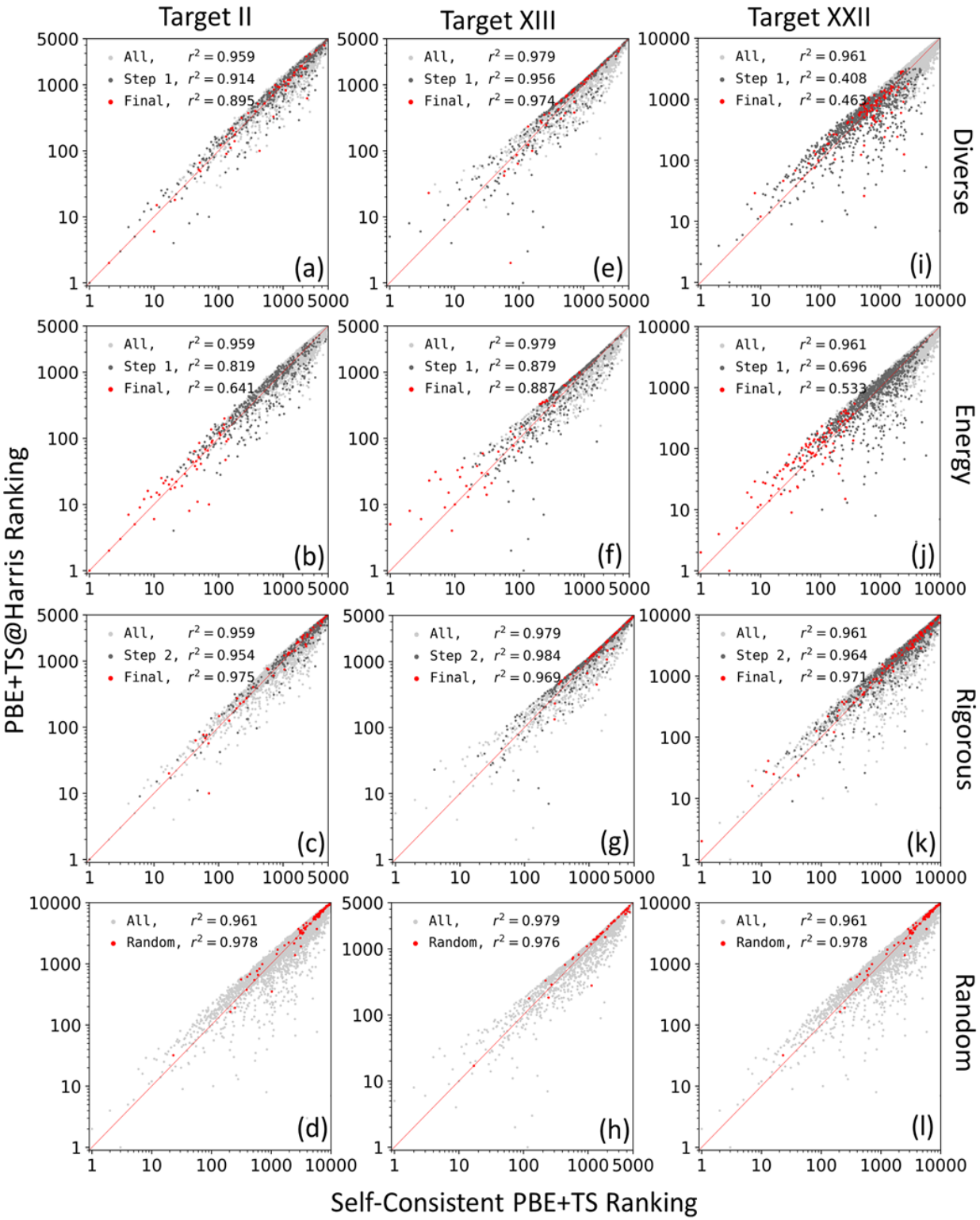


Figure 8: PBE+TS@Harris ranking vs. self-consistent PBE+TS ranking of structures selected in the various steps of the Diverse, Energy, Rigorous, and Random workflows for Target II (panels a, b, c, d), Target XIII (panels e, f, g, h) and Target XXII (panels i, j, k, l). Selections for additional iterations of the Rigorous workflow are omitted for clarity.

Three standard workflows have been developed for Genarris, based on different sequences of successive clustering and filtering steps, as shown in Figure 2. A primary difference among the Diverse, Energy, and Rigorous workflows lies in the selection of structures from the raw pool for further evaluation and optimization. Figure 8 shows the structures selected in different steps of the three standard workflows for Targets II, XIII, and XXII. The Random workflow, used as a control, does not employ any criterion for selection. The selected structures are indicated on a graph of the PBE+TS@Harris ranking vs. the self-consistent PBE+TS ranking, plotted on a log-log scale to provide a higher resolution in the low energy region. For the Diverse and Energy workflows, the structures selected in step 1 (the first 10% selection) are highlighted in dark gray, and the final selected structures (after the second 10% selection) are highlighted in red. For the Rigorous workflow, which involves an iterative selection process, only the structures selected in the second iteration (step 2) and the final structures are highlighted in dark gray and red, respectively (additional iterations are omitted for clarity).

The distributions of structures selected by the different workflows show distinct characteristics. The Energy workflow selects the majority of structures in the lower end of the spectrum for all three targets, as shown in panels (b), (f), and (j) (a few are not selected due to the clustering). Meanwhile, both the Diverse and Rigorous workflows select structure with a broader energy spectrum, with the Rigorous workflow sampling more structures in the higher energy range, as shown in panels (a), (e), (i), (c), (g), and (k). The structures sampled by the Random workflow are scattered across the distribution, with few structures ranked below 100, as shown in panels (d), (h), and (l).

The change of the r^2 score (calculated with self-consistent PBE+TS ranking as the

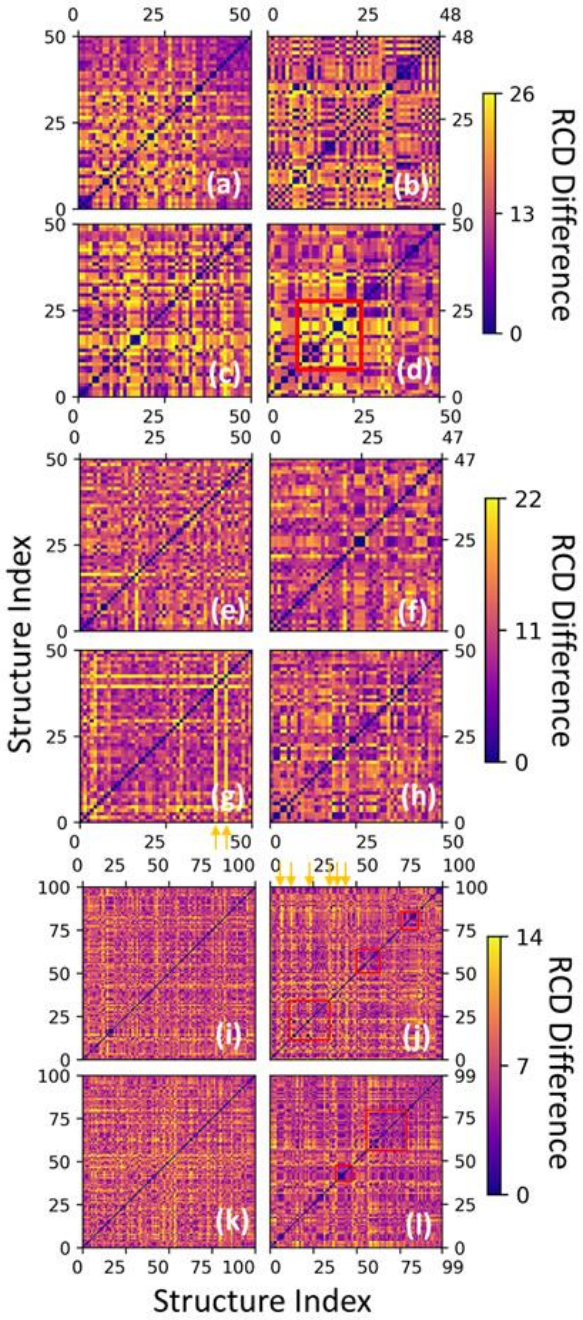


Figure 9: Distance matrices of the Diverse, Energy, Rigorous and Random workflows for Targets II (a, b, c, d), XIII (e, f, g, h) and XXII (i, j, k l). Distances are based on the RCD as described in Section 2.3.1. The red box in (d) indicates concentrated clusters that are far from one another, and those in (j) and (l) indicate oversampled clusters. Orange arrows in (g) and (j) indicate isolated structures in under-sampled regions.

"true" value and PBE+TS@Harris as "prediction") through the different workflow steps also reveals distinct patterns. In the Energy workflow r^2 deteriorates significantly from one step to the next because it mainly samples the lower end of the distribution, where the errors of the HA are most severe. In the Diverse workflow the deterioration of r^2 is typically less significant, as structures are selected across the spectrum. In the Rigorous workflow r^2 tends to increase in the final selection. This may be because the selection is based on self-consistent single point DFT energy evaluations, rather than on the HA. The Random workflow does not show any significant change in r^2, as shown in panels (d), (h), and (l). Exceptions to the r^2 score trends are the Energy workflow for Target XIII, shown in panel (f) and the Diverse workflow for Target XXII, shown in panel (i). In the former case, the deterioration of r^2 is mitigated by sampling a group of structures concentrated towards the higher end of the spectrum. The selection of structures in the higher energy region by the Energy workflow reflects the effect of clustering, which identified this region as containing distinct structural motifs that must be sampled. In the latter case, the step 1 energy-based selection of a large group of structures in the upper-middle range of the spectrum leads to a significant dip of r^2, which is not fully corrected by the final selection step. Additional analyses of the energy and volume distributions of the structures selected by the different workflows are provided in the supplementary material.

In Table II, the outcomes of the Diverse, Energy, Rigorous, and Random workflows of Genarris are compared in terms of the composition of the fully relaxed final pools of Targets II, XIII, and XXII. The Rigorous workflow successfully finds the experimental structure for all three targets, serving its purpose as a global minimum search method. The Energy workflow tends to yield a higher number of duplicates because it systematically samples the low energy regions of the potential energy surface, which increases the likelihood of sampling similar structures that relax to the same local minimum. The Diverse and Rigorous workflows tend to yield a lower number of duplicates because they are designed to sample different regions of the potential energy landscape and similar structures are effectively eliminated by clustering. Target XIII is an exception to these trends, possibly due to its halogen bonds (see also Ref. 48).

TABLE II. Analysis of the final pools for Targets II, XIII and XXII obtained with the Diverse, Energy, Rigorous and Random workflows.

	All	Target II			Target XIII			Target XXII		
Workflow	Found Exp.?	Dup. Pairs	Uniq. Struct.	Avg (STD) RCD Diff.	Dup. Pairs	Uniq. Struct.	Avg (STD) RCD Diff.	Dup. Pairs	Uniq. Struct.	Avg (STD) RCD Diff.
Diverse	No	3	47	13.80 (6.46)	2	48	11.48 (4.91)	1	99	7.22 (2.47)
Energy	No	26	35	13.82 (7.32)	4	43	11.01 (4.56)	28	80	7.30 (2.73)
Rigorous	Yes	1	49	14.84 (7.63)	7	44	11.76 (5.73)	0	100	7.49 (2.53)
Random	No	11	40	14.33 (7.35)	5	45	11.11 (4.78)	9	93	6.89 (2.55)

The differences in the composition of the final pools produced by the Diverse, Energy, Rigorous, and Random workflows are also reflected in the distance matrices, shown in Figure 9. The structures are pre-sorted according to their BE and the distances are calculated based on the RCD, as described in Section 2.3.1. The average distance and standard deviation are given in Table II. Across the three targets, the Rigorous pools consistently have the largest average distance between structures, indicating the most diverse sampling. Graphically, this manifests as overall brighter distance matrices for Target II and XXII in panels (c) and (k). For Target XIII, the larger average may be attributed in part to the two isolated structures, appearing as two bright

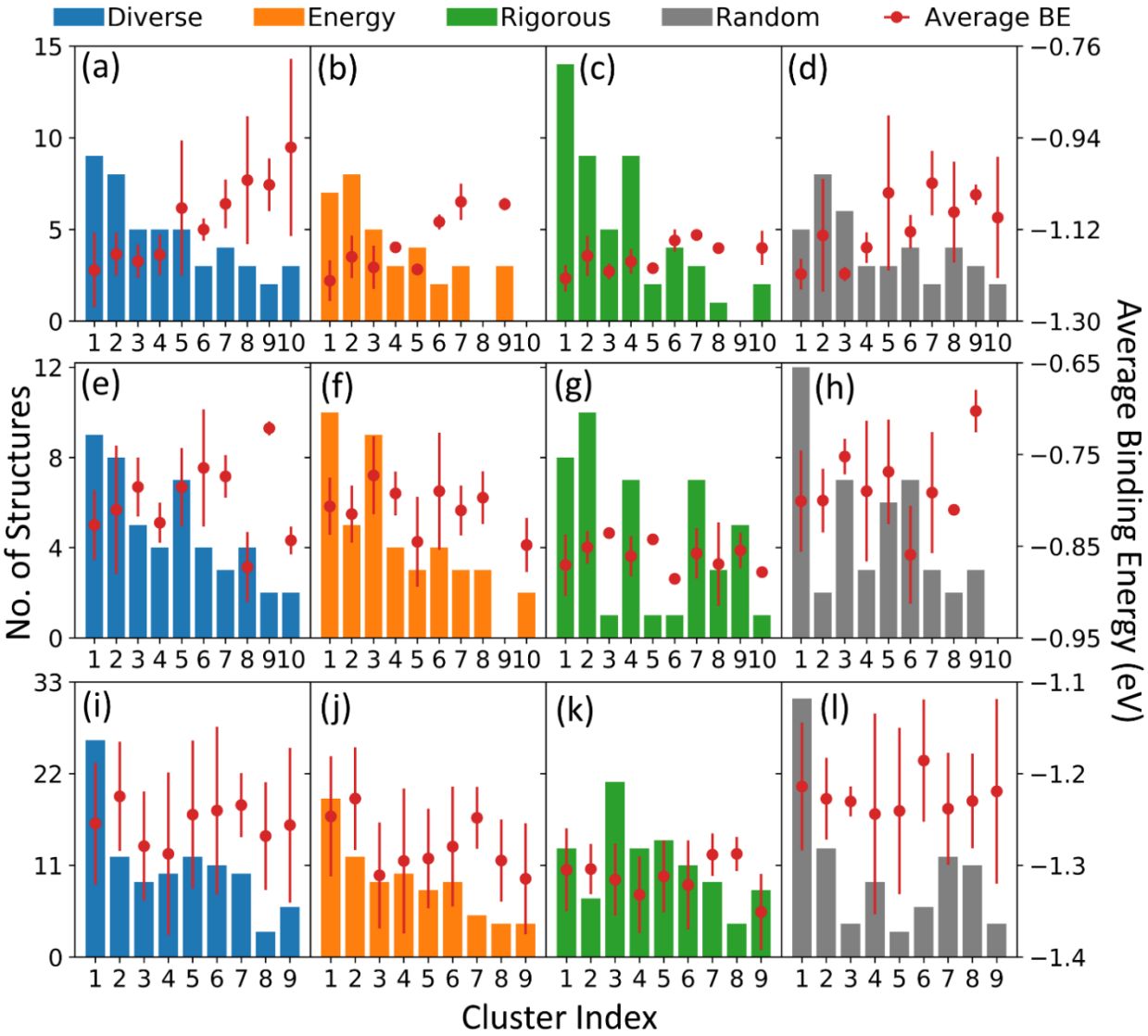


Figure 10: Clustering analysis of the final populations generated by the Diverse, Energy, Rigorous, and Random workflow for Target II (panels a, b, c, d), Target XIII (panels e, f, g, h), and Target XXII (panels i, j, k, l). The histograms show the number of structures that fall into each cluster when the four pools are combined and clustered together. The red markers indicate the average and standard deviation of the BE per molecule for each bin.

lines indicated by the arrows in panel (g). The distance matrices of the Energy pools have a more structured, grid-like appearance. This is particularly obvious for Targets II and XIII, as shown in panels (b) and (f). This indicates groups of structures that are similar within their clusters but different across clusters. This uneven sampling of the configuration space is reflected in the larger standard deviation of distances. For Target XXII, although the grid-like feature is not as prominent (partly due to the larger pool size), clustered sampling is revealed by the darker blocks along the diagonal, framed in red in panel (j), and isolated sampling is revealed by the bright lines, indicated by arrows. The distance matrices of the Diverse pools appear the most even and least structured, as shown in panels (a), (e), and (i). This is corroborated, especially for Targets II and XXII, by a smaller distance standard deviation, which indicates a more uniform sampling.

The Random pools show varied patterns in their distance matrices. For Target II, the Random workflow performed rather poorly, in terms of diverse sampling, except for the two distinct clusters in the lower energy region, framed in red in panel (d). For Targets XIII and XXII, the Random pools, shown in panels (h) and (l), exhibit similar patterns to the Energy pools, shown in panels (f) and (j). This is possibly because some basins of the configuration space are overrepresented in the raw pool and are therefore more likely to be sampled randomly.

The differences in the composition of the final pools produced by the Diverse, Energy, Rigorous, and Random workflows are further elucidated by the clustering analysis, presented in Figure 10. For this analysis, the four final workflow pools of each target were first merged, and RCD-AP clustering was applied to cluster the combined pools into 10 clusters for Target II and XIII, and 9 clusters for Target XXII. Then, histograms were generated by counting the number of structures originating from each workflow in each cluster. The average and standard deviation of the BE per molecule of the structures in each bin are also shown. Overall, the final pools of the Diverse workflow achieve the most uniform sampling across all clusters for all three targets, as shown in panels (a), (e), and (i). For Targets II and XIII, the Energy and Rigorous workflows under-sample or completely miss certain clusters, as shown in panels (b), (c), (f), and (g). The clusters under-sampled by these two energy-selective workflows tend to be higher in energy. The Rigorous workflow consistently provides the lowest energy structures with the smallest standard deviation for all three targets, as shown in panels (c), (g), and (k). In contrast, the Diverse workflow, especially for Target XXII, samples structures across a broader and higher energy range.

Overall, the results presented in this section demonstrate how the different progression of clustering and selection steps in the Diverse, Energy, and Rigorous workflows of Genarris leads to different outcomes in terms of the composition of the final pools. The selection of curated populations of structures based on different criteria may be desirable for different purposes. The user may choose one of the standard workflows suggested here or design their own workflows. In the next section, we demonstrate an application of Genarris for creating an initial population for a genetic algorithm and discuss the effect of the pool composition on the GA search outcomes.

4.4. Genetic Algorithm Performance

The Energy, Diverse, and Random pools generated for each target were used as initial populations for the GAtor genetic algorithm for crystal structure prediction with the settings described in section 3. To illustrate the effect of the initial pool on the behavior of GAtor, a set of low energy structures, representative of the main packing motifs of each target, were selected from Ref. 48. The structures are indexed according to their relative energy, as calculated with the PBE-based hybrid functional, PBE0, and the MBD method therein. In Figure 11, the smallest RCD distance to each of these representatives is plotted as a function of GA iteration to show the convergence towards these structures for Target II (panels a, b, c), Target XIII (panels d, e, f), and Target XXII (panels h, i, j). In each panel the row colors become lighter from left to right as the GA reaches closer towards each representative structure. White indicates that the structure is found. Not all the representative structures were found by the time the GA runs sampled here were stopped (we note that the purpose of this analysis was not to perform an exhaustive GA search, as explained in Section 3). The convergence towards these structures provides useful information on how the composition of the initial pool affects the GA performance.

Overall, starting the GA from the Diverse pool results in the best performance, reaching most of the representative structures and approaching the rest closely within the iteration limit used

here. In particular, these runs consistently find the experimental structures (#7 for Target II and #1 for Targets XIII and XXII). Starting the GA from the Energy pool leads to inconsistent performance. The Energy pool is a good starting point for Target II, finding the experimental structure within a few iterations and also approaching most closely the #1 PBE0+MBD structure. However, for Target XIII and Target XXII, the GA runs started from Energy pool fail to reach most of the representative structures. This inconsistent performance may be a function of whether or not certain packing motifs are adequately represented in the low energy region of the raw pool, as ranked by the HA. The GA runs started from the Random pools consistently exhibit the worse performance, only reaching a few of the representative structures. We therefore conclude that starting a GA from a maximally diverse initial population provides the optimal performance. We recommend using the Diverse workflow of Genarris to produce initial pools for GAtor.

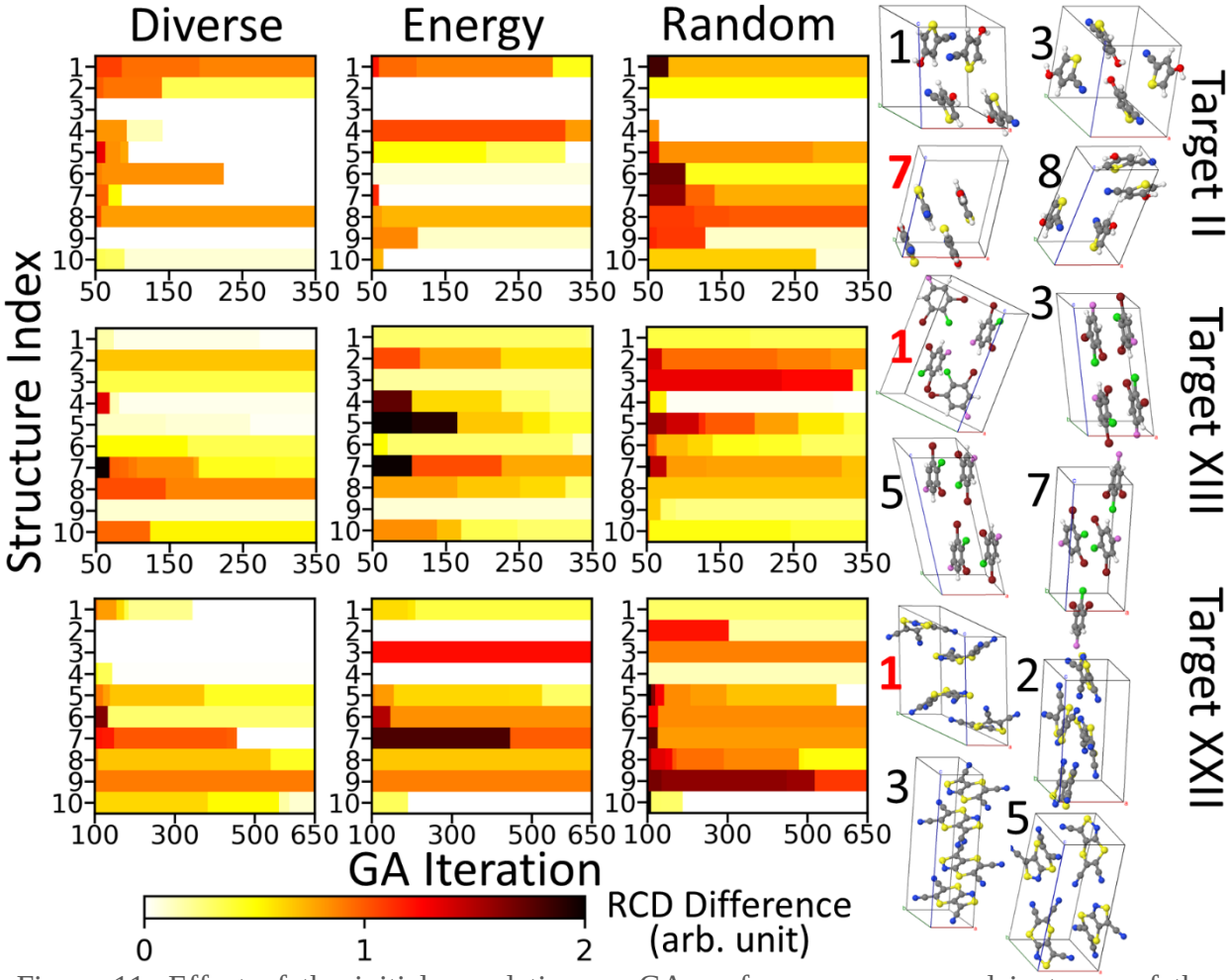


Figure 11: Effect of the initial population on GA performance measured in terms of the minimal RCD distance to a set of representative structures as a function of GA iteration. Some of the representative structures are shown on the right with the experimental structures marked in red.

5. Conclusion

We have introduced Genarris, a Python package for generating random crystal structures of rigid molecules, and demonstrated its application for three past blind test targets. For fast screening of random structures, Genarris relies on the Harris approximation (HA), which has been implemented in the FHI-aims code. The Harris density of a molecular crystal is constructed by superposition of single molecule densities, calculated only once. The DFT total energy is then evaluated for the Harris density without performing a self-consistency cycle. The HA has been validated for binding energy curves of molecular dimers as well as for the ranking of randomly generated molecular crystal structures. The HA is found to be sufficiently reliable in both scenarios, as long as no molecules are unphysically close to each other, in which case the HA fails to capture the strong repulsion between the molecular densities. This situation is avoided in Genarris by imposing a minimum distance between atoms belonging to different molecules during structure generation.

Beyond random structure generation, three standard workflows have been proposed for using Genarris to create curated populations of structures by applying successive steps of clustering and selection to the "raw" pool. The Rigorous workflow is a crystal structure prediction in and of itself, the Energy workflow creates a low-energy pool of structures, and the Diverse workflow balances low energy and maximal diversity. To perform clustering based on structural similarity within the three workflows, we have developed the relative coordinate descriptor (RCD). The RCD is based on the relative positions and orientations of neighboring molecules in the crystal, rather than on interatomic distances. Two machine learning algorithms for clustering, *k*-means and affinity propagation (AP), have been tested here, in conjunction with the RCD and a radial distribution function (RDF) descriptor. RCD-based AP clustering has been found to yield the best performance. AP clustering is better than *k*-means at resolving isolated structurally distinct clusters. RCD-based clustering is better than RDF-based clustering at identifying potential duplicates, resolving packing motif similarity (manifested as space group symmetry) rather than unit cell volume similarity, and achieving a higher silhouette score. Therefore, RCD-AP clustering is the method of choice for all workflows of Genarris.

The outcomes of the Rigorous, Energy, and Diverse workflows have been evaluated with respect to the composition of the final populations of structures and compared to a Random workflow, which selects structures randomly for the final pool. The Rigorous workflow has proven to be an effective structure search method, as it successfully located the experimentally observed structures of all three targets. Based on several indicators, the Diverse workflow provides the most uniform sampling, while the Energy workflow tends to over-sample some regions of the configuration space. The Diverse and Energy workflows have been further evaluated for the purpose of generating an initial pool of structures for a genetic algorithm. For all three targets, launching a genetic algorithm from the Diverse initial pool provides the best performance in terms of convergence towards a representative set of low-energy structures with different packing motifs.

In summary, we have demonstrated versatile applications of Genarris for random structure generation, for crystal structure prediction, and for creating an initial population of structures for a genetic algorithm. Genarris may be applied more broadly for a variety of purposes. For example, Genarris may be used to create curated sets of structures for other optimization algorithms, such as swarm algorithms, Monte Carlo methods, and Bayesian optimization, or to create training sets for machine learning algorithms. To this end, the user may choose one of the workflows proposed here or design their own workflows.

Supplementary Material

See supplementary material for additional details on the calculation of unit cell parameters and molecular rotations, additional details of the Harris approximation implementation in FHI-aims, and additional analyses of the pools of structures generated by the different workflows of Genarris.

Acknowledgements

Work at CMU was funded by the National Science Foundation (NSF) Division of Materials Research through grant DMR-1554428. CS, KS, and HO gratefully acknowledge support from the Solar Technologies GoHybrid initiative of the State of Bavaria. An award of computer time was provided by the Innovative and Novel Computational Impact on Theory and Experiment (INCITE) program. This research used resources of the Argonne Leadership Computing Facility, which is a DOE Office of Science User Facility supported under Contract DE-AC02-06CH11357.

References

[1] H. Hoppe and N.S. Sariciftci, J. Mater. Res. **19**, 1924 (2004).
[2] O. Lavastre, I. Illitchev, G. Jegou, and P.H. Dixneuf, J. Am. Chem. Soc. **124**, 5278 (2002).
[3] T. Tozawa, J.T. a Jones, S.I. Swamy, S. Jiang, D.J. Adams, S. Shakespeare, R. Clowes, D. Bradshaw, T. Hasell, S.Y. Chong, C. Tang, S. Thompson, J. Parker, A. Trewin, J. Bacsa, A.M.Z. Slawin, A. Steiner, and A.I. Cooper, Nat. Mater. **8**, 973 (2009).
[4] J.T.A. Jones, T. Hasell, X. Wu, J. Bacsa, K.E. Jelfs, M. Schmidtmann, S.Y. Chong, D.J. Adams, A. Trewin, F. Schiffman, F. Cora, B. Slater, A. Steiner, G.M. Day, and A.I. Cooper, Nature **474**, 367 (2011).
[5] J. Bernstein, Cryst. Growth Des. **11**, 632 (2011).
[6] A.J. Cruz-Cabeza, S.M. Reutzel-Edens, and J. Bernstein, Chem. Soc. Rev. **44**, 8619 (2015).
[7] J. Bernstein, *Polymorphism in Molecular Crystals* (Oxford University Press, Oxford, UK, 2010).
[8] R.K. Harris, Analyst **131**, 351 (2006).
[9] S.L. Price, D.E. Braun, and S.M. Reutzel-Edens, Chem. Commun. **52**, 7065 (2016).
[10] M. Brinkmann, G. Gadret, M. Muccini, C. Taliani, N. Masciocchi, and A. Sironi, J. Am. Chem. Soc. **122**, 5147 (2000).
[11] R.J. Tseng, R. Chan, V.C. Tung, and Y. Yang, Adv. Mater. **20**, 435 (2008).
[12] R. Pfattner, M. Mas-Torrent, I. Bilotti, A. Brillante, S. Milita, F. Liscio, F. Biscarini, T. Marszalek, J. Ulanski, A. Nosal, M. Gazicki-Lipman, M. Leufgen, G. Schmidt, W.M. Laurens, V. Laukhin, J. Veciana, and C. Rovira, Adv. Mater. **22**, 4198 (2010).
[13] M. Wang, J. Li, G. Zhao, Q. Wu, Y. Huang, W. Hu, X. Gao, H. Li, and D. Zhu, Adv. Mater. **25**, 2229 (2013).
[14] D. Yan and D.G. Evans, Mater. Horizons **1**, 46 (2014).
[15] Y. Li, D. Ji, J. Liu, Y. Yao, X. Fu, W. Zhu, C. Xu, H. Dong, J. Li, and W. Hu, Sci. Rep. **5**, 13195 (2015).
[16] C.H. Pham, E. Kucukbenli, and S. de Gironcoli, arXiv preprint arXiv:1605.00733 (2016).
[17] S.M. Woodley and R. Catlow, Nat. Mater. **7**, 937 (2008).
[18] A.R. Oganov and C.W. Glass, J. Chem. Phys. **124**, 244704 (2006).
[19] Y. Wang, J. Lv, L. Zhu, and Y. Ma, Phys. Rev. B **82**, 94116 (2010).
[20] C.J. Pickard and R.J. Needs, J. Phys. Condens. Matter **23**, 53201 (2011).

[21] C.M. Freeman, J.W. Andzelm, C.S. Ewig, J. Hill, and B. Delley, Chem. Commun. **2**, 2455 (1998).
[22] L. Stievano, F. Tielens, I. Lopes, N. Folliet, C. Gervais, D. Costa, and J.F. Lambert, Cryst. Growth Des. **10**, 3657 (2010).
[23] N. Marom, R.A. Distasio Jr., V. Atalla, S. V. Levchenko, J.R. Chelikowsky, L. Leiserowitz, and A. Tkatchenko, Angew. Chemie Int. Ed. **52**, 6629 (2013).
[24] G.J.O. Beran, Angew. Chemie Int. Ed. **54**, 396 (2015).
[25] G.J.O. Beran, Chem. Rev. **116**, 5567 (2016).
[26] J.P.M. Lommerse, W.D.S. Motherwell, H.L. Ammon, J.D. Dunitz, A. Gavezzotti, D.W.M. Hofmann, F.J.J. Leusen, W.T.M. Mooij, S.L. Price, B. Schweizer, M.U. Schmidt, B.P. van Eijck, P. Verwer, and D.E. Williams, Acta Crystallogr. Sect. B Struct. Sci. Cryst. Eng. Mater. **56**, 697 (2000).
[27] W.D.S. Motherwell, H.L. Ammon, J.D. Dunitz, A. Dzyabchenko, P. Erk, A. Gavezzotti, D.W.M. Hofmann, F.J.J. Leusen, J.P.M. Lommerse, W.T.M. Mooij, S.L. Price, H. Scheraga, B. Schweizer, M.U. Schmidt, B.P. Van Eijck, P. Verwer, and D.E. Williams, Acta Crystallogr. Sect. B **58**, 647 (2002).
[28] G.M. Day, W.D.S. Motherwell, H.L. Ammon, S.X.M. Boerrigter, R.G. Della Valle, E. Venuti, A. Dzyabchenko, J.D. Dunitz, B. Schweizer, B.P. Van Eijck, P. Erk, J.C. Facelli, V.E. Bazterra, M.B. Ferraro, D.W.M. Hofmann, F.J.J. Leusen, C. Liang, C.C. Pantelides, P.G. Karamertzanis, S.L. Price, T.C. Lewis, H. Nowell, A. Torrisi, H.A. Scheraga, Y.A. Arnautova, M.U. Schmidt, and P. Verwer, Acta Crystallogr. Sect. B **61**, 511 (2005).
[29] G.M. Day, T.G. Cooper, A.J. Cruz-Cabeza, K.E. Hejczyk, H.L. Ammon, S.X.M. Boerrigter, J.S. Tan, R.G. Della Valle, E. Venuti, J. Jose, S.R. Gadre, G.R. Desiraju, T.S. Thakur, B.P. Van Eijck, J.C. Facelli, V.E. Bazterra, M.B. Ferraro, D.W.M. Hofmann, M.A. Neumann, F.J.J. Leusen, J. Kendrick, S.L. Price, A.J. Misquitta, P.G. Karamertzanis, G.W.A. Welch, H.A. Scheraga, Y.A. Arnautova, M.U. Schmidt, J. Van De Streek, A.K. Wolf, and B. Schweizer, Acta Crystallogr. Sect. B **65**, 107 (2009).
[30] D.A. Bardwell, C.S. Adjiman, Y. a. Arnautova, E. Bartashevich, S.X.M. Boerrigter, D.E. Braun, A.J. Cruz-Cabeza, G.M. Day, R.G. Della Valle, G.R. Desiraju, B.P. Van Eijck, J.C. Facelli, M.B. Ferraro, D. Grillo, M. Habgood, D.W.M. Hofmann, F. Hofmann, K.V.J. Jose, P.G. Karamertzanis, A. V. Kazantsev, J. Kendrick, L.N. Kuleshova, F.J.J. Leusen, A. V. Maleev, A.J. Misquitta, S. Mohamed, R.J. Needs, M. a. Neumann, D. Nikylov, A.M. Orendt, R. Pal, C.C. Pantelides, C.J. Pickard, L.S. Price, S.L. Price, H. a. Scheraga, J. Van De Streek, T.S. Thakur, S. Tiwari, E. Venuti, and I.K. Zhitkov, Acta Crystallogr. Sect. B **67**, 535 (2011).
[31] A.M. Reilly, R.I. Cooper, C.S. Adjiman, S. Bhattacharya, A.D. Boese, J.G. Brandenburg, P.J. Bygrave, R. Bylsma, J.E. Campbell, R. Car, D.H. Case, R. Chadha, J.C. Cole, K. Cosburn, H.M. Cuppen, F. Curtis, G.M. Day, R.A. DiStasio Jr, A. Dzyabchenko, B.P. van Eijck, D.M. Elking, J.A. can den Ende, J.C. Facelli, M.B. Ferraro, L. Fusti-Molnar, C.-A. Gatsiou, T.S. Gee, R. de Gelder, L.M. Ghiringhelli, H. Goto, S. Grimme, R. Guo, D.W.M. Hofmann, J. Hoja, R.K. Hylton, L. Iuzzolino, W. Jankiewicz, D.T. de Jong, J. Kendrick, N.J.J. de Klerk, H.-Y. Ko, L.N. Kuleshova, X. Li, S. Lohani, F.J.J. Leusen, A.M. Lund, J. Lv, Y. Ma, N. Marom, A.E. Masunov, P. McCabe, D.P. McMahon, H. Meekes, M.P. Metz, A.J. Misquitta, S. Mohamed, B. Monserrat, R.J. Needs, M.A. Neumann, J. Nyman, S. Obata, H. Oberhofer, A.R. Oganov, A.M. Orendt, G.I. Pagola, C.C. Pantelides, C.J. Pickard, R. Podeszwa, L.S. Price, S.L. Price, A. Pulido, M.G. Read, K. Reuter, E. Schneider, C. Schober, G.P. Shields, P. Singh, I.J. Sugden, K. Szaleqicz, C.R. Taylor, A. Tkatchenko, M.E. Tuckerman, F. Vacarro, M. Vasileiadis, A. Vazquez-Mayagoitia, L.

Vogt, Y. Wang, R.E. Watson, G.A. de Wijs, J. Yang, Q. Zhu, and C.R. Groom, Acta Crystallogr. Sect. B **72**, 439 (2016).
[32] A.M. Reilly and A. Tkatchenko, Phys. Rev. Lett. **113**, 55701 (2014).
[33] F. Curtis, X. Wang, and N. Marom, Acta Crystallogr. Sect. B **72**, 562 (2016).
[34] A.M. Reilly and A. Tkatchenko, J. Chem. Phys. **139**, (2013).
[35] A. Tkatchenko, Adv. Funct. Mater. **25**, 2054 (2014).
[36] J. Harris, Phys. Rev. B **31**, 1770 (1985).
[37] G.D. Bellchambers and F.R. Manby, J. Chem. Phys. **135**, (2011).
[38] K. Berland, E. Londero, E. Schröder, and P. Hyldgaard, Phys. Rev. B **88**, 45431 (2013).
[39] D.E. Williams, Acta Crystallogr. Sect. A **52**, 326 (1996).
[40] B.P. Van Eijck and J. Kroon, J. Comput. Chem. **20**, 799 (1999).
[41] D.H. Case, J.E. Campbell, P.J. Bygrave, and G.M. Day, J. Chem. Theory Comput. **12**, 910 (2016).
[42] A. V Dzyabchenko, J. Struct. Chem. **25**, 416 (1984).
[43] I.M. Sobol, Zh. Vychisl. Mat. I Mat. Fiz. 7 **7**, 784 (1967).
[44] P.G. Karamertzanis and C.C. Pantelides, J. Comput. Chem. **26**, 304 (2005).
[45] R.G. Della Valle, E. Venuti, A. Brillante, and A. Girlande, J. Phys. Chem. A **110**, 10858 (2006).
[46] Q. Zhu, A.R. Oganov, C.W. Glass, and H.T. Stokes, Acta Crystallogr. Sect. B **68**, 215 (2012).
[47] A. Supady, V. Blum, and C. Baldauf, J. Chem. Inf. Model. **55**, 2338 (2015).
[48] F. Curtis, X. Li, T. Rose, A. Vazquez-Mayagoitia, S. Bhattacharya, L.M. Ghiringhelli, and N. Marom, to be published.
[49] R.H. Swendsen and J.-S. Wang, Phys. Rev. Lett. **57**, 2607 (1986).
[50] Y.G. Andreev, G.S. MacGlashan, and P.G. Bruce, Phys. Rev. B **55**, 12011 (1997).
[51] D.J. Earl and M.W. Deem, Phys. Chem. Chem. Phys. **7**, 3910 (2005).
[52] D.J. Wales, Science. **285**, 1368 (1999).
[53] S. Goedecker, J. Chem. Phys. **120**, 9911 (2004).
[54] G.G. Rondina and J.L.F. Da Silva, J. Chem. Inf. Model. **53**, 2282 (2013).
[55] J. Pillardy, Y.A. Arnautova, C. Czaplewski, K.D. Gibson, and H.A. Scheraga, Proc. Natl. Acad. Sci. U. S. A. **98**, 12351 (2001).
[56] J.A. Chisholm and S. Motherwell, J. Appl. Crystallogr. **38**, 228 (2005).
[57] T. Mueller, A.G. Kusne, and R. Ramprasad, Rev. Comput. Chem. **29**, 186 (2016).
[58] M. Rupp, Int. J. Quantum Chem. **115**, 1058 (2015).
[59] S. Curtarolo, D. Morgan, K. Persson, J. Rodgers, and G. Ceder, Phys. Rev. Lett. **91**, 135503 (2003).
[60] G. Ceder, D. Morgan, C. Fischer, K. Tibbetts, and S. Curtarolo, MRS Bull. **31**, 981 (2006).
[61] C.C. Fischer, K.J. Tibbetts, D. Morgan, and G. Ceder, Nat. Mater. **5**, 641 (2006).
[62] M. Rupp, A. Tkatchenko, K.-R. Müller, V. Lilienfeld, and O. Anatole, Phys. Rev. Lett. **108**, 58301 (2012).
[63] G. Montavon, M. Rupp, V. Gobre, A. Vazquez-Mayagoitia, K. Hansen, A. Tkatchenko, K.R. Müller, and O. Anatole Von Lilienfeld, New J. Phys. **15**, 95003 (2013).
[64] K. Hansen, G. Montavon, F. Biegler, S. Fazli, M. Rupp, M. Scheffler, O.A. Von Lilienfeld, A. Tkatchenko, and K.R. Müller, J. Chem. Theory Comput. **9**, 3404 (2013).
[65] G. Hautier, C.C. Fischer, A. Jain, T. Mueller, and G. Ceder, Chem. Mater. **22**, 3762 (2010).
[66] Y. Saad, D. Gao, T. Ngo, S. Bobbitt, J.R. Chelikowsky, and W. Andreoni, Phys. Rev. B **85**, 104104 (2012).

[67] K.T. Schütt, H. Glawe, F. Brockherde, A. Sanna, K.R. Müller, and E.K.U. Gross, Phys. Rev. B **89**, 205118 (2014).
[68] O. Isayev, D. Fourches, E.N. Muratov, C. Oses, K. Rasch, A. Tropsha, and S. Curtarolo, Chem. Mater. **27**, 735 (2015).
[69] A. Seko, H. Hayashi, K. Nakayama, A. Takahashi, and I. Tanaka, Phys. Rev. B **95**, 144110 (2017).
[70] M. de Jong, W. Chen, R. Notestine, K. Persson, G. Ceder, A. Jain, M. Asta, and A. Gamst, Sci. Rep. **6**, 34256 (2016).
[71] B.A. Calfa and J.R. Kitchin, AIChE J. **62**, 2605 (2016).
[72] E.O. Pyzer-Knapp, G.N. Simm, and A. Aspuru Guzik, Mater. Horiz. **3**, 226 (2016).
[73] G. Pilania, C. Wang, X. Jiang, S. Rajasekaran, and R. Ramprasad, Sci. Rep. **3**, 2810 (2013).
[74] V. Botu and R. Ramprasad, Int. J. Quantum Chem. **115**, 1074 (2015).
[75] A.P. Bartók, M.C. Payne, R. Kondor, and G. Csányi, Phys. Rev. Lett. **104**, 136403 (2010).
[76] J. Behler, J. Phys. Condens. Matter **26**, 183001 (2014).
[77] C.M. Handley and J. Behler, Eur. Phys. J. B **87**, 152 (2014).
[78] J.R. Boes, M.C. Groenenboom, J.A. Keith, and J.R. Kitchin, Int. J. Quantum Chem. **116**, 979 (2016).
[79] S.A. Ghasemi, A. Hofstetter, S. Saha, and S. Goedecker, Phys. Rev. B **92**, 45131 (2015).
[80] S. Hajinazar, J. Shao, and A.N. Kolmogorov, Phys. Rev. B **95**, 14114 (2017).
[81] A. Seko, A. Takahashi, and I. Tanaka, Phys. Rev. B **92**, 54113 (2015).
[82] J.C. Snyder, M. Rupp, K. Hansen, K.R. Müller, and K. Burke, Phys. Rev. Lett. **108**, 253002 (2012).
[83] Z.D. Pozun, K. Hansen, D. Sheppard, M. Rupp, K.R. Müller, and G. Henkelman, J. Chem. Phys. **136**, 174101 (2012).
[84] T. Stecher, N. Bernstein, and G. Csányi, J. Chem. Theory Comput. **10**, 4079 (2014).
[85] N.J. Browning, R. Ramakrishnan, O.A. von Lilienfeld, and U. Roethlisberger, J. Phys. Chem. Lett. **8**, 1351 (2017).
[86] B.R. Goldsmith, M. Boley, J. Vreeken, M. Scheffler, and L.M. Ghiringhelli, New J. Phys. **19**, 13031 (2017).
[87] S. De, F. Musil, T. Ingram, C. Baldauf, and M. Ceriotti, J. Cheminform. **9**, 1 (2017).
[88] M. Boley, B.R. Goldsmith, L.M. Ghiringhelli, and J. Vreeken, Data Min. Knowl. Discov. **31**, 1391 (2017).
[89] L.J. Nelson, G.L.W. Hart, F. Zhou, and V. Ozoliņš, Phys. Rev. B **87**, 35125 (2013).
[90] L.M. Ghiringhelli, J. Vybiral, S. V. Levchenko, C. Draxl, and M. Scheffler, Phys. Rev. Lett. **114**, 105503 (2015).
[91] L.M. Ghiringhelli, J. Vybiral, E. Ahmetcik, R. Ouyang, S. V. Levchenko, C. Draxl, and M. Scheffler, New J. Phys. **19**, (2017).
[92] P. V. Balachandran, J. Theiler, J.M. Rondinelli, and T. Lookman, Sci. Rep. **5**, 13285 (2015).
[93] E.L. Willighagen, R. Wehrens, P. Verwer, R. De Gelder, and L.M.C. Buydens, Acta Crystallogr. Sect. B **61**, 29 (2005).
[94] V. Blum, R. Gehrke, F. Hanke, P. Havu, V. Havu, X. Ren, K. Reuter, and M. Scheffler, Comput. Phys. Commun. **180**, 2175 (2009).
[95] J.P. Perdew, K. Burke, and M. Ernzerhof, Phys. Rev. Lett. **77**, 3865 (1996).
[96] J.P. Perdew, K. Burke, and M. Ernzerhof, Phys. Rev. Lett. **78**, 1396 (1997).
[97] A. Tkatchenko and M. Scheffler, Phys. Rev. Lett. **102**, 73005 (2009).
[98] M.I. Aroyo, J.M. Perez-Mato, C. Capillas, E. Kroumova, S. Ivantchev, G. Madariaga, A.

Kirov, and H. Wondratschek, Zeitschrift Für Krist. Mater. **221**, 15 (2006).
[99] S. Batsanov, Inorg. Mater. **37**, 871 (2001).
[100] A. Sharapov and G. Hutchison, in *Abstr. Pap. Am. Chem. Soc.* (2012).
[101] G.R. Hutchison, in *Abstr. Pap. Am. Chem. Soc.* (2013).
[102] C. Schober, K. Reuter, and H. Oberhofer, J. Chem. Phys. **144**, 54103 (2016).
[103] M.A. Blanco, M. Flórez, and M. Bermejo, J. Mol. Struct. THEOCHEM **419**, 19 (1997).
[104] B. Huang and O.A. Von Lilienfeld, J. Chem. Phys. **145**, (2016).
[105] K. Hansen, F. Biegler, R. Ramakrishnan, W. Pronobis, O.A. Von Lilienfeld, K.R. Müller, and A. Tkatchenko, J. Phys. Chem. Lett. **6**, 2326 (2015).
[106] P. Verwer and F.J.J. Leusen, in *Rev. Comput. Chem.* (John Wiley & Sons, Inc., 2007), pp. 327–365.
[107] J.B. MacQueen, in *Proc. 5-Th Berkeley Symp. Math. Stat. Probab.* (1967), pp. 281–297.
[108] D. Dueck and B.J. Frey, Science. **315**, 972 (2007).
[109] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, G. Louppe, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and É. Duchesnay, J. Mach. Learn. Res. **12**, 2825 (2012).
[110] P.J. Rousseeuw, J. Comput. Appl. Math. **20**, 53 (1987).
[111] Ref. 31; supplemental information of submission 12.

3.4 Submitted Manuscript: GAtor: A First Principles Genetic Algorithm for Molecular Crystal Structure Prediction

This manuscript details the methodology and applications of the GAtor genetic algorithm package, written in Python. GAtor is the main subject of this thesis and I contributed to the majority of the work presented. I developed the code from scratch, and implemented and tested all the different crossover, mutation, selection, fitness evaluation, data analysis, and duplicate check modules. I implemented the module that interfaces with FHI-aims. I also developed the various parallel workflows of GAtor so that GAtor can run on a variety of HPC supercomputing environments. I ran the GA for the four molecules studied and performed the majority of the energetic post processing for re-ranking of the final structures. I wrote the entire manuscript and I made all the figures included.

GAtor: A First Principles Genetic Algorithm for Molecular Crystal Structure Prediction

Farren Curtis,[†] Xiayue Li,[‡] Timothy Rose,[¶] Álvaro Vázquez-Mayagoitia,[§] Saswata Bhattacharya,[‖] Luca M. Ghiringhelli,[⊥] and Noa Marom[*,¶,†,#]

†*Department of Physics, Carnegie Mellon University, Pittsburgh, PA 15213, USA*
‡*Google, Mountain View, CA 94030, USA*
¶*Department of Materials Science and Engineering, Carnegie Mellon University, Pittsburgh, PA 15213, USA*
§*Argonne Leadership Computing Facility, Argonne National Laboratory, Lemont, Illinois, 60439, USA.*
‖*Department of Physics, Indian Institute of Technology Delhi, Hauz Khas, New Delhi 110016, India*
⊥*Fritz-Haber-Institut der Max-Planck-Gesellschaft, Faradayweg 4-6, 14195, Berlin, Germany*
#*Department of Chemistry, Carnegie Mellon University, Pittsburgh, PA 15213, USA*

E-mail: nmarom@andrew.cmu.edu

Abstract

We present the implementation of GAtor, a massively parallel, first principles genetic algorithm (GA) for molecular crystal structure prediction. GAtor is written in Python and currently interfaces with the FHI-aims code to perform local optimizations and energy evaluations using dispersion-inclusive density functional theory (DFT). GAtor offers a variety of fitness evaluation, selection, crossover, and mutation schemes. Breeding operators designed specifically for molecular crystals provide a balance between exploration and exploitation. Evolutionary niching is implemented in GAtor by using machine learning to cluster the dynamically updated population by structural similarity then employing a cluster-based fitness function. Evolutionary niching promotes uniform sampling of the potential energy surface by evolving small sub-populations, which helps overcome initial pool biases and selection biases (genetic drift). The various settings offered by GAtor increase the likelihood of locating all low-energy minima, including those located in disconnected, hard to reach regions of the potential energy landscape. The best structures generated are re-relaxed and re-ranked using a hierarchy of increasingly accurate DFT functionals and dispersion methods. GAtor is applied to a chemically diverse set of four past blind test targets, characterized by different types of intermolecular interactions. The experimentally observed structures and other low-energy structures are found for all four targets. In particular, for Target II, 5-cyano-3-hydroxythiophene, the top ranked putative crystal structure is a Z'=2 structure with $P\bar{1}$ symmetry and a scaffold packing motif, which has not been reported previously.

1 Introduction

Molecular crystals are a unique class of materials with diverse applications in pharmaceuticals, organic electronics, pigments, and explosives.[1–11] The molecules comprising these crys-

tals are bound by weak dispersion (van der Waals) interactions. As a result, the same molecule may crystallize in several different solid forms, known as polymorphs. Because the structure of a molecular crystal governs its physical properties, polymorphism may drastically impact the desired functionality for a given application. For pharmaceuticals, different polymorphs may display varying stability, solubility, and compressibility, affecting the drug's manufacturability, bioavailability, and efficacy.[1,12,13] For applications in organic electronics and organic photovoltaics (OPV), different polymorphs possess different optoelectronic properties,[14,15] directly impacting device performance.[16–18]

Because molecular crystals have a wide range of applications, there has been increasing interest in the fundamental challenge of crystal structure prediction (CSP), or the computation of a molecule's putative crystal structure(s) solely from its two-dimensional chemical diagram, examples of which are shown in Fig. 1. This challenge is embodied by CSP blind tests, organized periodically by the Cambridge Crystallographic Data Centre.[19–24] CSP can reveal the general behavior of a target molecule, predict the existence of new polymorphs, and serve as a complementary tool for experimental investigations.[13,25,26] Once considered unachievable,[27] CSP is still an extremely challenging task because it requires combining highly accurate electronic structure methods with efficient algorithms for configuration space exploration.

The energy differences between molecular crystal polymorphs are typically within a few kJ/mol,[28–31] which calls for the accuracy of a quantum mechanical approach. Reaching the required accuracy has become more practical thanks to a decade of development in dispersion-inclusive density functional theory (DFT), including exchange-correlation functionals[32–43] and pairwise methods that add the leading order C^6/R^6 dispersion term to the inter-nuclear energy.[44–55] Notably, the recently developed many-body dispersion (MBD) method[56–58] accurately describes the structure, energetics, dielectric properties, and mechanical properties of molecular crystals[3,14,28,59–64] by accounting for long range electrostatic screening and non-pairwise-additive contributions of many-body dispersion interactions. Using dispersion-inclusive DFT for the final ranking of relative stabilities has become a CSP best practice.[24] Vibrational contributions to the zero-point energy and free energy of the system at finite temperature have also been shown to be important, and may be further included.[3,61,65–67]

Approaches to configuration space exploration in CSP include molecular dynamics,[68,69] Monte Carlo methods,[25,70] particle swarm optimization,[71] and (quasi)-random searches.[72,73] Genetic algorithms (GAs) are a versatile class of optimization algorithms inspired by the evolutionary principle of survival of the fittest.[74–76] A GA starts from an initial pool of locally optimized trial structures. The scalar descriptor (or combination of descriptors) being optimized is mapped onto a fitness function and structures with higher fitness values are assigned higher probabilities for mating. Breeding operators create offspring structures by combining the structural genes[i] of one or more parent structure(s). The child structure is locally optimized and added to the population. The cycle of local optimization, fitness evaluation, and offspring generation propagates structural features associated with the property being optimized and repeats to "convergence" (A GA is not guaranteed to find the global minimum. For practical purposes, convergence may be defined as when the GA can no longer find any new low-energy structures in a large number of iterations).

GAs can be applied robustly to complex multidimensional search spaces, including those with many extrema or discontinuous derivatives. They provide a good balance between exploration and exploitation by introducing randomness in the mating step followed by local optimization. Furthermore, they are conceptually simple algorithms, ideal for parallelization,

[i]The term "genetic algorithm" is sometimes reserved for an evolutionary algorithm that purely encodes an individual's genes with bit-string representations. For our purposes we make no such distinction between genetic and evolutionary algorithms.

and can lead to unbiased and unintuitive solutions. In the context of structure prediction, the target function being optimized is typically the total or free energy. GAs have been used extensively to find the global minimum structures of crystalline solids[77–89] and clusters.[74–76,90–98] Advantageously, the GA fitness function may be based on any property of interest, not necessarily the energy.[74,99–104] For organic molecular crystals the goal is not just to locate the most stable structure but also any potential polymorphs. In the most recent CSP blind test,[24] GAs were used by us[ii] and others (see submissions #8, #12, #21).

Here, we present GAtor, a new, massively parallel, first principles genetic algorithm (GA) specifically designed for structure prediction of crystal structures of (semi-)rigid molecules. GAtor is written in Python with a modular structure that allows the user to switch between and/or modify core GA routines for specialized purposes. For initial pool generation, GAtor relies on a separate package, Genarris, reported elsewhere[105] and briefly described in Section 3.1. GAtor offers a variety of features that enable the user to customize the search settings as needed for chemically diverse systems, including different fitness, selection, crossover, and mutation schemes. GAtor is designed to fully utilize high performance computing (HPC) architectures by spawning several parallel GA replicas that read from and write to a common population of structures. This approach does not require a full "generation" of candidates to complete before performing a new selection.[96,97,104] For energy evaluations and local optimization of trial structures, GAtor employs dispersion-inclusive DFT by interfacing with the ab initio, all-electron electronic structure code FHI-aims.[106,107]

The paper is organized as follows: Section 2 describes the DFT methods and numerical settings of FHI-aims used in conjunction with GAtor; Section 3 details GAtor's parallelization scheme and the features currently available in the code; Section 4 showcases applications of GAtor for a chemically diverse set of four past blind test targets, 3,4-cyclobutylfuran (Target I[20]), 5-cyano-3-hydroxythiophene (Target II[20]), 1,3-dibromo-2-chloro-5-fluorobenzene (Target XIII[22]), and tricyano-1,4-dithiino[c]-isothiazole (Target XXII[24]) shown in Fig. 1. Finally, Section 5 provides concluding remarks and best practices.

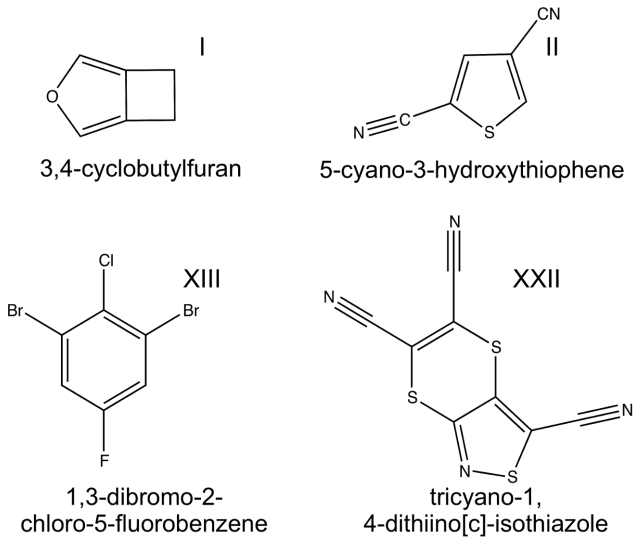


Figure 1: Two-dimensional molecular diagrams of four past blind test targets, Target I,[20] Target II,[20] Target XIII,[22] and Target XXII.[24]

2 DFT Settings

Because first principles calculations are computationally expensive, lighter DFT settings are employed within the GA search, with the intention of locating the experimental structure and any potential polymorphs among the lowest energy structures. To obtain more precise rankings, the best structures produced from the GA are postprocessed with higher-level functionals and dispersion corrections. Hierarchal screening approaches have become a common practice in CSP.[24] For GAtor, the user has the option to input FHI-aims control files for any desired level(s) of theory. The DFT settings used in the present study are detailed below.

For local structural optimizations within the GA, the generalized gradient approximation of Perdew-Burke-Ernzerhof (PBE)[108,109] is used with the pairwise Tkatchenko-Scheffler (TS)

[ii] In the sixth blind test we used a preliminary version of GAtor.

dispersion-correction[54] with *lower-level* numerical settings, which correspond to the light numerical settings and tier 1 basis sets of FHI-aims.[106] Additionally, a $2 \times 2 \times 2$ k-point grid and reduced angular grids are used. A convergence value of 10^{-5} electrons is set for the change in charge density in the self-consistent field (SCF) cycle and SCF forces and stress evaluations are not computed. These settings are implemented in order to accelerate geometry relaxations within the GA. For Target XIII, atomic ZORA scalar relativity[106] settings are used for the heavier halogen elements.

For postprocessing, the best 5-10% of the final structures produced by the GA are re-relaxed and re-ranked using a $3 \times 3 \times 3$ k-point grid, PBE+TS, and *higher-level* numerical settings, which correspond to the tight/tier2 default settings of FHI-aims.[106] Next, single point energy (SPE) evaluations are performed using PBE with the MBD method[56–58] for the best structures as ranked by PBE+TS. The final re-ranking is performed using the hybrid functional PBE0[110,111] with the MBD correction. The inclusion of 25% exact exchange in PBE0 mitigates the self-interaction error, leading to a more accurate description of electron densities and multipoles.[60,61] For some molecular crystals the correct polymorph ranking is reproduced only when using PBE0+MBD.[14,28] The PBE0+MBD ranking is considered to be the most reliable of the methods used here. Thermal contributions to the total energy, shown to change the energy ranking in approximately 9% of organic compounds,[67] are not further included in the present study.

3 Code Description

GAtor is written in Python and uses the spglib[112] crystal symmetries library, scikit learn[113] machine learning package, and pymatgen[114] library for materials analysis. GAtor is available for download from http://software.noamarom.com under a BSD-3 license. The code is modular by design, such that core GA tasks, such as selection, similarity checks, crossover, and mutation can be interchanged in the user input file and/or modified. For energy evaluations and local optimization GAtor currently interfaces with the all-electron DFT code FHI-aims,[106,107] and may be modified to interface with other electronic structure and molecular dynamics packages.

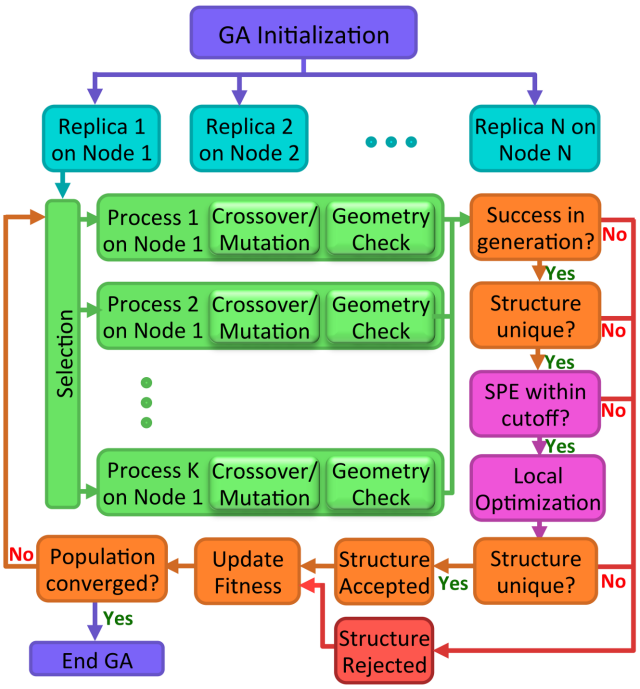


Figure 2: An example workflow of GAtor on a high performance computing cluster. In the diagram, N independent GA replicas run on N computing nodes, with K core processing units per node. Single point energy (SPE) evaluations and local optimizations are performed using FHI-aims.

GAtor takes advantage of high performance computing (HPC) architectures by avoiding processor idle time and effectively utilizing all available resources. An example workflow is shown in Fig. 2. After initialization, the master process spawns a user-defined number of GA replicas across N nodes. Each independent replica performs the core genetic algorithm tasks independently while reading from and writing to a dynamically updated pool of structures.[96,97,104]

Additional multiprocessing may be utilized within each replica for child generation. Two classes of breeding operators are implemented in GAtor, crossover and mutation, described in detail in sections 3.4 and 3.5. Crossover op-

erators create a child by combining the structural genes of two parents, whereas mutation operators create a child by altering the structural genes of one parent. After selection, either crossover or mutation is performed with a user-defined probability. When multiprocessing is used, the same set of parents (crossover) or single parent (mutation) undergo the same breeding operation, but with different random parameters. If a child cannot pass the geometry checks after a user-defined number of attempts, a new selection is performed. Otherwise, the first child that passes the geometry checks proceeds to the first uniqueness check. If a candidate structure successfully passes all geometry checks, uniqueness checks, and energy cutoffs, it is added to the common population. The fitness of each structure in the population is updated, and a new selection can be performed immediately. A detailed account of the core tasks and features of the GA is provided below.

3.1 GA Initialization

During GA initialization GAtor reads in an initial pool of structures generated by the Genarris random structure generation package.[105] Genarris generates random symmetric crystal structures in the 230 crystallographic space groups, then combines fragment-based DFT with clustering techniques from machine learning to produce a high-quality, diverse starting population at a relatively low computational cost, as described in detail in Ref. 105. The initial pool structures are pre-relaxed with PBE+TS and *lower-level* numerical settings as described in Section 2 and their total energies are stored beforehand. GAtor updates their starting fitness values, as described below, before performing selection.

3.2 Fitness Evaluation

In GAtor, the fitness of an individual determines its likelihood of being chosen for crossover or mutation. GAtor provides a traditional energy-based fitness function, in which structures with lower relative stabilities are assigned higher fitness values. Additionally, GAtor provides the option of a cluster-based fitness function, which can use various clustering techniques to perform evolutionary niching. Using cluster-based fitness can reduce genetic drift, as explained below, by suppressing the over-sampling of certain regions of the potential energy surface and promoting the evolution of several subpopulations simultaneously.

3.2.1 Energy-based Fitness

Within energy-based fitness, the total energy E_i of the ith structure in the population is evaluated using dispersion-inclusive DFT as detailed in Section 2. The fitness f_i of each structure is defined as,

$$f_i = \frac{\epsilon_i}{\sum_i \epsilon_i} \qquad 0 \leq f \leq 1 \tag{1}$$

$$\epsilon_i = \frac{E_{\max} - E_i}{E_{\max} - E_{\min}} \tag{2}$$

where ϵ_i is the ith structure's relative energy, and $E_{\max}$ and $E_{\min}$ correspond to the structures with the dynamically updated highest and lowest total energies in the population, respectively.[96,97,104] Hence, structures with lower relative energies have higher fitness values.

3.2.2 Cluster-Based Fitness

When using a traditional energy-based fitness function, a GA may be prone to exploring the same region(s) of the potential energy surface, which may or may not include the experimental structure(s) or the global minimum structure. This may be due to a number of factors, including lack of diversity in the common population and selection biases towards or against certain packing motifs over time, a phenomenon known as genetic drift. Genetic drift can result from biases in the initial pool[105] and from the topology of the potential energy landscape (e.g. a desirable packing motif for a given molecule could be located in narrow well that is rarely visited). The search may also be influenced by systematic biases of the energy method used (e.g., the exchange-correlation functional and dispersion method) towards or against certain packing motifs.[14]

GAs may be adapted to be more suitable for multi-modal optimization using evolutionary niching methods.[115–119] Niching methods support the formation of stable subpopulations in the neighborhood of several optimal solutions. For molecular crystal structure prediction, incorporating niching techniques may increase diversity and diminish the effect of inherent or initial pool biases. The goal is for the GA to locate all low-energy polymorphs which may or may not have similar structural motifs to the experimentally observed crystal structure(s) or the most stable crystal structure present in the population.

GAtor provides the option to dynamically cluster the common population of molecular crystals into groups (niches) of structural similarity, using pre-defined feature vectors for each target molecule and clustering algorithms implemented in the sci-kit learn machine learning Python package.[113] Currently, GAtor offers the use of radial distribution function (RDF) vectors of interatomic distances for user-defined species, relative coordinate descriptor (RCD) vectors,[105] or a simple lattice parameter based descriptor, L, given by:

$$L = \frac{1}{\sqrt[3]{V}}(a, b, c) \tag{3}$$

where V is the unit cell volume and a, b, and c are the structure's lattice parameters after employing Niggli reduction[120–123] and unit cell standardization. Niggli reduction produces a unique representation of the translation vectors of the unit cell but does not define a standard orientation. Therefore, all unit cell lattice vectors are standardized such that $\vec{a}$ points along the $\hat{x}$ direction, $\vec{b}$ lies in the xy plane, and the convention $a \leq b \leq c$ is used. The lattice parameter based descriptor encourages the sampling of under-represented lattices in the population (e.g. structures which are almost 2D which may have one lattice parameter significantly shorter than the others). GAtor offers k-means clustering[124] and affinity propagation (AP),[125] and may be adapted to use other clustering algorithms implemented in sci-kit learn. AP is a clustering method that determines the number of clusters in a data set, based on a structure similarity metric, rather than defining the number of clusters *a priori*. This has the advantage of resolving small, structurally distinct clusters.[105] Once the common population has been clustered into niches, a fitness sharing scheme[117] is applied such that a structure's scaled fitness, f_i', is given by

$$f_i' = \frac{f_i}{m_i} \tag{4}$$

where m_i is a cluster-based scaling parameter, currently determined by the number of structures in each individual's shared cluster. This clustering scheme increases the fitness of under-sampled low-energy motifs within the population, and suppresses the over-sampling of densely populated regions. One example of evolutionary niching is discussed in Section 4.1 for Target XXII. Further investigations of the effect of the descriptor and the fitness function will be the subject of future work.

There are a variety of other strategies for incorporating niching or clustering into an evolutionary algorithm. Refs. 126,127 use fingerprint functions based on inter-atomic distances to prevent too dissimilar structures from mating. Recently, Ref. 98 explored incorporating agglomerative hierarchical clustering (AHC) into an evolutionary algorithm applied to organic molecules and surfaces. AHC detects the number of clusters in the given data set, similar to AP. One of their methods promoted selection of cluster outliers, while another utilized a fitness function that combined the structure's cluster size with its energy, similar to the technique employed in GAtor.

3.3 Selection

Selection is inspired by the evolutionary principle of survival of the fittest. In GAtor, individuals with structural motifs associated with higher fitness values have a higher probability of being selected for mating. GAtor currently offers a choice of two genetic algorithm selection strategies: roulette wheel selection and tournament selection.

3.3.1 Roulette wheel selection

This selection technique[128] simulates a roulette wheel, where fitter individuals in the population conceptually take up larger slots on the wheel, and therefore have a higher probability of being selected when the wheel is spun. In GAtor, the procedure is as follows: First, a random number r is chosen, uniform in the interval [0, 1]. Then, a parent structure is selected for mating if it has the first sorted, normalized fitness value with $f_i > r$.[96,97,104]

3.3.2 Tournament Selection

In tournament selection,[128] a user-defined number of individuals are randomly selected from the common population to form a tournament. In GAtor, the two structures with the highest fitness values in the tournament (i.e. the winner and the runner-up) are selected for mating. Tournament selection is efficient (requiring no sorting of the population) and gives the user control over the selection pressure via control of the tournament size.[129] Using a larger tournament size gives preference to the best structures in the population, while a smaller tournament has a higher probability of selecting less fit individuals for the purposes of maintaining diversity.

3.4 Crossover

Crossover is an operator that combines the structural genes of two parent structures selected for mating to form a single offspring. The crossover operators implemented in GAtor were developed specifically for organic molecular crystals. The popular 'cut-and-splice'[130] crossover operator used in other genetic algorithms, takes a random fraction of the each parent's unit cell (and the motifs within) and pastes them together. While this approach is successful for structure prediction of clusters and inorganic crystals,[77,78,83,84,93,95–98,104,131,132] it may not be the most natural choice for molecular crystals because it can break important space group symmetries that may be associated with, e.g., efficient packing and lower total energies. Initialization of the starting population within random symmetric space groups has been shown to increase the efficiency of evolutionary searches.[86,127,133] In the same vein, further steps can be taken to design the breeding operators themselves to exploit and explore the symmetry of the starting population and to reduce the number of expensive first principles calculations on structures far from equilibrium. Therefore, several mutation and crossover operators implemented in GAtor can preserve or break certain space group symmetries of the parent structure(s), as detailed below.

3.4.1 Standard Crossover

In this crossover scheme each parent's genes are represented by the Niggli-reduced, standardized unit cell lattice parameters and angles $(a, b, c, \alpha, \beta, \gamma)$ as well as the molecular geometry[iii], orientation $\Phi = (\theta_z, \theta_y, \theta_x)$, and center of mass (COM) position in fractional coordinates, R_{COM}, of each molecule within the unit cell. The orientation of each molecule within the unit cell is defined by computing the θ_z, θ_y, and θ_x Euler angles, respectively, which rotate a Cartesian reference frame to a reference frame aligned with each molecule's principal axes of rotation. When generating a child structure, the molecules in the unit cell of each parent structure are randomly paired together. The fractional COM positions for each molecule in the child structure are directly inherited from one randomly selected parent. The lattice parameters from each parent are combined with random fractions to form the lattice parameters of the child structure. The child's molecular geometries are inherited from one randomly selected parent and initially centered at the origin with their principal axes of rotation aligned with the Cartesian axes. The final orientations of the molecules in the child structure are constructed by combining the orientation angles of

[iii]The geometry of the molecules are allowed to relax during local optimization. This is important for semi-rigid molecules, such as Target XXII. This extra degree of freedom is accounted for in the crossover process by randomly selecting the relaxed molecular geometry from one parent.

the paired molecules from the parent structures with random fractions.

Fig. 3, panel (a) shows an example of standard crossover for two selected parent structures of Target XXII with space groups $P2_1/c$ and $Pca2_1$, respectively. Four molecules from each parent are randomly selected (circled in blue) and paired together. The molecular geometries and COM positions of the child structure are both inherited from the $P2_1/c$ parent structure. The orientation angles of the molecules paired from each parent structure are combined with random fractions. The lattice parameters are also combined with random fractions. In this specific example, a child structure is created with a $Z' = 2$ motif that has lower symmetry than either of its parents, $P\bar{1}$, but still contains inversion symmetry before local optimization.

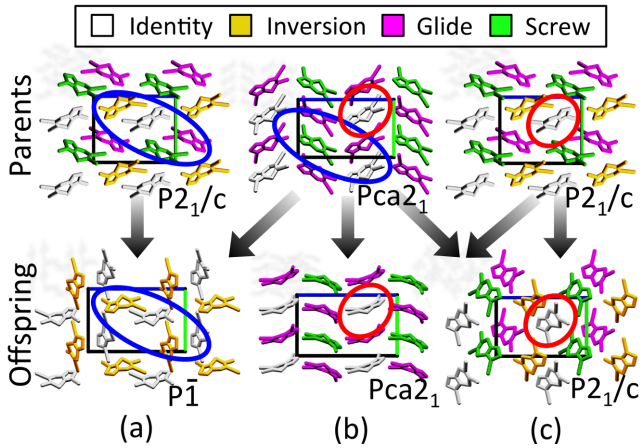


Figure 3: Examples of (a) standard crossover and (b-c) symmetric crossover applied to selected parent structures of Target XXII. The colors of the molecules correspond to the symmetry operations applied to the asymmetric unit of each structure, shown in white. The a, b, and c crystallographic lattice vectors are displayed in red, green, and blue, respectively.

3.4.2 Symmetric Crossover

In this crossover scheme each parent's genes are represented by the orientation and COM position of their respective crystallographic asymmetric units as well as their respective space group operations and unit cell lattice parameters. For the explicit computation of each parent's asymmetric unit and space group operations, GAtor relies on the pymatgen[114] package, which utilizes the spglib crystal symmetries library.[112] When generating a child structure, the genes of the parents are combined strategically to preserve one parent's space group symmetries as follows. First, the asymmetric unit and corresponding space group operations are deduced for both parents. If the two asymmetric units contain the same number of molecules, then the respective molecules in each unit are paired together. If the asymmetric units contain a different number of molecules, then one parent's asymmetric unit is used as a reference and paired with an equivalent number of molecules in the second parent's unit cell. If the asymmetric units contain different relaxed molecular geometries, then the molecular conformations in the child's asymmetric unit may be randomly inherited from one parent. The orientation and COM position of the molecule(s) within the child's asymmetric unit are constructed by combining the orientation and COM position of the paired molecule(s) from each parent with random fractions. If both parents possess the same Bravais lattice type then their lattice parameters may be combined with random fractions. Otherwise, the child's lattice is randomly inherited from one parent. Finally, the symmetry operations (containing specific translations, reflections, and rotations of the asymmetric unit in fractional coordinates) are selected from one parent and applied to the child's generated asymmetric unit and lattice. Either parent's space group operations may be randomly selected and applied to the child's asymmetric unit when both parents possess the same number of molecule's in the asymmetric unit and the same Bravais lattice type. Otherwise, one parent's space group operations will be compatible with the symmetry of the generated lattice and asymmetric unit by construction and are thus applied. This crossover procedure ensures the space group of the child is directly inherited from one of its parents, at least before local optimization.

Examples of symmetric crossover are shown in Fig. 3, panels (b) and (c). The participat-

ing asymmetric units of the parent and child structures are circled in red. In panel (b), the child structure inherits the molecular geometry from the $Pca2_1$ parent structure, which is more planar than the molecular geometry of the $P2_1/c$ structure. The orientations of the asymmetric units (both Z'=1) and the parent lattice vectors are combined with random weights. The space group symmetry operations from the $Pca2_1$ parent are applied to the child's asymmetric unit on the generated lattice. In panel (c), the child structure inherits the molecular geometry and symmetry operations from the $P2_1/c$ parent structure. The orientations of the parents' asymmetric units and their lattice vectors are combined with random weights. The randomness used when creating the orientation of the motif in the asymmetric unit explains why the child shown in panel (b) has a different orientation of the asymmetric unit as the one shown in panel (c), and allows for more diversity in the generated offspring. In these specific examples, both child structures produced using symmetric crossover have higher symmetry than the child produced with standard crossover, before local optimization.

3.5 Mutation

Mutation operators are applied to the genes of single parent structures to form new offspring. In GAtor, certain mutations may promote exploration of the potential energy surface via dramatic structural changes, while others may exploit promising regions via subtle changes. The user chooses the percentage of selected structures that undergo mutation, and may select specific or random mutations to be applied. GAtor also provides an option that allows a percentage of structures to undergo a combination of any two mutation operations before local optimization. This approach encourages exploration and may reduce the number of duplicate structures generated in the search.[84]

3.5.1 Strains

GAtor offers a variety of strain operators that produce child structures by acting upon the lattice vectors of the selected parent structure. Similar to Refs.,[77,84,86] the strain tensor is represented using the symmetric Voigt strain matrix $\boldsymbol{\epsilon}$,

$$\boldsymbol{\epsilon} = \begin{bmatrix} \epsilon_{11} & \frac{\epsilon_{12}}{2} & \frac{\epsilon_{13}}{2} \\ \frac{\epsilon_{12}}{2} & \epsilon_{22} & \frac{\epsilon_{23}}{2} \\ \frac{\epsilon_{13}}{2} & \frac{\epsilon_{23}}{2} & \epsilon_{33} \end{bmatrix}. \quad (5)$$

The strain matrix is applied to each lattice vector $\vec{a}_{\text{parent}}$ of the chosen parent structure to produce the lattice vector of the child $\vec{a}_{\text{child}}$ via

$$\vec{a}_{\text{child}} = \vec{a}_{\text{parent}} + \boldsymbol{\epsilon}\vec{a}_{\text{parent}} \quad (6)$$

The components of ϵ_{ij} are chosen to produce different modes of strain. To apply completely random strains, all six unique ϵ_{ij} components are randomly selected from a normal distribution with a user-defined standard deviation that determines the strength of the applied mutation. To apply random deformations in certain crystallographic directions, one or more random ϵ_{ij} may be chosen while the others are set to 0. Strains that preserve the overall unit cell volume of the parent structure, or change a single unit cell angle, may also be applied. When applying a strain, the COM of each molecule is moved according to its fractional coordinates. An example of strain mutation is shown in Fig. 4 panel (a). Here, a random strain is applied that transforms the lattice of the parent structure from monoclinic ($\alpha = \gamma = 90; \beta \neq 90$) to triclinic ($\alpha \neq \beta \neq \gamma \neq 90$). The COM of each molecule is moved accordingly, breaking the glide and screw symmetry of the parent structure and creating a Z'=2 child structure.

3.5.2 Molecular Rotations

Rotation mutations change the orientations of the molecules in the selected parent structure. Different random rotations can be applied to the Cartesian coordinates of the atoms in selected molecules centered at the origin, or the same random rotation can be applied about each molecule's principal axes of rotation. For Z'=1 structures, the latter type of rotation is equivalent to randomly changing the orientation of molecule in the asymmetric unit, as shown in Fig. 4, panel (b). Here, each molecule

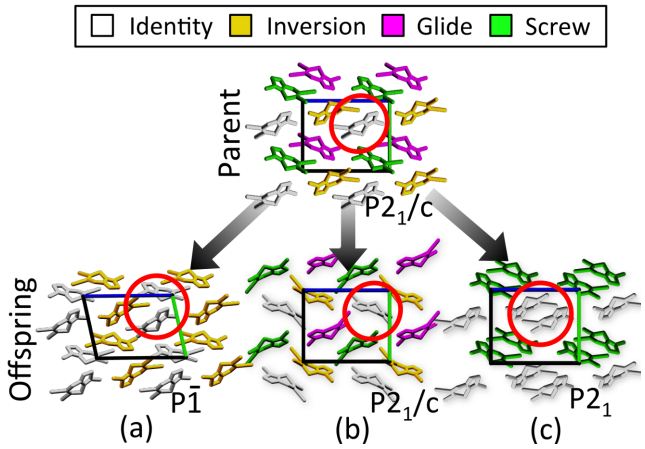


Figure 4: Examples of (a) random strain, (b) rotation, and (c) translation mutations applied to a $P2_1/c$ structure of Target XXII. The colors of the molecules correspond to the symmetry operations applied to the asymmetric unit of each structure, shown in white and circled in red. The a, b, and c crystallographic lattice vectors are displayed in red, green, and blue, respectively.

from the parent structure receives the same random rotation about its principal axes of rotation, rotating the asymmetric unit and preserving the parent's $P2_1/c$ symmetry in the resulting offspring.

3.5.3 Translations

Translational mutations change the position of R_{COM} for each molecule within the unit cell. They are either applied randomly to the COM (in Cartesian coordinates) of selected molecules, or in a random direction according to each molecule's inertial reference frame. An example of the latter type of mutation is depicted in Fig. 4, panel (c). Here, each molecule from the parent structure receives the same random translation according to the orientation of its inertial reference frame. In this example, paired enantiomers are translated in equal and opposite directions, which breaks the glide symmetry of the parent structure, and forms an asymmetric unit containing two molecules in a tightly packed dimer.

3.5.4 Permutations

Permutation mutations swap R_{COM} for randomly selected molecules within the unit cell. Depending on the point group symmetry of the molecule, the lattice, and the permutation, this operator can preserve, add, or break certain space group symmetries of the parent structure. An example permutation mutation that preserves the parent's space group symmetry is shown in Fig. 5, panel (a). Here, a permutation is applied which effectively swaps R_{COM} of the highlighted asymmetric unit (shown in white) and its nearest neighbor (shown in yellow), as well as swapping R_{COM} of the two other molecules in the unit cell related by screw and glide symmetry (shown in green and fuchsia, respectively). As a result, the child structure inherits the $P2_1/c$ symmetry of the parent structure.

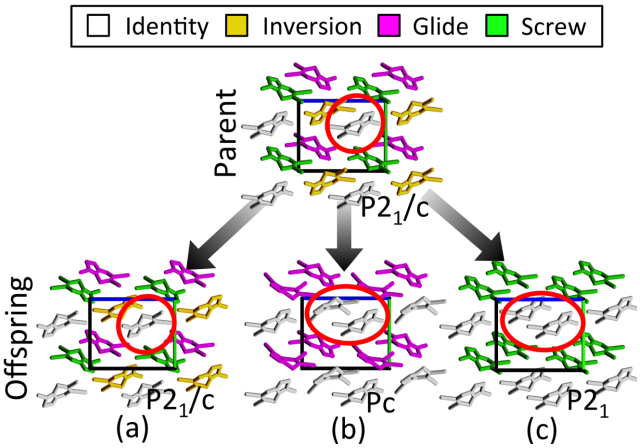


Figure 5: Examples of (a) permutation, (b) permutation-rotation, and (c) permutation-reflection mutations applied to a $P2_1/c$ structure of Target XXII. The colors of the molecules correspond to the symmetry operations applied to the asymmetric unit of each structure, shown in white and circled in red. The a, b, and c crystallographic lattice vectors are displayed in red, green, and blue, respectively.

3.5.5 Permutation-Rotations and Permutation-Reflections

Permutation-rotation mutations swap randomly selected molecules within the unit cell and then a random rotation is applied about

their principal axes of rotation. Fig. 5, panel (b) shows an example of permutation-rotation. Here, the two molecules in the parent unit cell colored in yellow and green swap position and undergo a random rotation, while the others remain fixed. As a result, the structure produced (space group Pc) no longer contains the exact two fold screw symmetry of the parent structure (space group $P2_1/c$), and effectively contains an asymmetric unit consisting of two molecules with the same chirality. In the permutation-reflection mutation, half of the molecules in the unit cell swap positions and then undergo a reflection in the xy, yz, or zx Cartesian planes centered at their COM. Fig. 5, panel (c) shows an example of permutation-reflection. Here, the two molecules in the parent unit cell colored in yellow and fuchsia swap positions and undergo a reflection about the zx plane pointing out of the page, while the others remain fixed. As a result, the structure produced (space group $P2_1$) no longer contains the glide symmetry of the parent structure (space group $P2_1/c$), and effectively contains an asymmetric unit consisting of two molecules of the same chirality. For crystals containing chiral molecules, such as Target XXII, this mutation can be especially effective because it can swap the relative positioning of enantiomers within the unit cell.

3.6 Rejection Criteria

Because in the mating step crossover or mutation operations are performed randomly on a diverse set of structures, the offspring generated may be unphysical or duplicates of existing structures. GAtor applies various criteria for rejecting a child structure before performing local optimization. This preserves the diversity of the population by preventing uncontrolled multiplication of similar structures and avoids computationally expensive local optimization of unreasonable or redundant structures.

3.6.1 Geometry Checks

Structures may be rejected if any two intermolecular contacts are too close. The minimum distance d_{min} between any two atoms A and B belonging to different molecules is given by:

$$d_{\text{min}} = s_{\text{r}}(r_{\text{A}} + r_{\text{B}}) \tag{7}$$

where r_A and r_B are the vdW radii of the atoms A and B, respectively, and s_{r} is user-defined and typically set between 0.6-0.9. Additionally, the user may constrain how close the COMs of any two molecules are allowed to be, or specify the allowed unit cell volume range for the generated structures. If the children produced by a parent or set of parents do not pass the geometry checks after a user-defined number of attempts, a new selection is performed.

3.6.2 Similarity Checks

Identifying duplicate crystal structures is critical for maintaining diversity and preventing a GA from getting stuck in a specific region of the potential energy surface. Furthermore, it is imperative to identify structures that are too similar to others in the existing population before local optimization to avoid expensive and redundant DFT calculations. Checking for duplicates is complicated by the fact that multiple representations exist for the same crystal structure. To address this issue, Niggli reduction[120–123] and cell standardization are used for all structures within GAtor, as previously described in Section 3.2.2.

GAtor performs a similarity check on all generated offspring before and after local optimization. The pre-relaxation similarity check prevents the local optimization of any structures too similar to others in the population, using loose site and lattice tolerances in pymatgen's StructureMatcher class.[114] The post-relaxation similarity check identifies whether any optimized structures relaxed into *bona fide* duplicates of existing structures in the population, using stricter site and lattice tolerance settings. If the candidate structure is found to have a similar lattice to another in the common pool (within the user-defined tolerances for the lattice parameter lengths and angles), then the root mean square (RMS) distances are computed between equivalent atomic sites. If the maximum, normalized RMS distance is within

the user-defined tolerance, then the two structures are determined to be duplicates.

3.6.3 Single Point Energy (SPE) Cutoff

Single point DFT calculations, using PBE+TS and *lower-level* numerical settings, are performed on unrelaxed offspring to decide whether they should undergo local optimization, as shown in Fig. 2. If the energy of the unrelaxed structure is higher than the user-defined cutoff, it is immediately rejected. This reserves computational resources for the local optimization of structures with energies that are more likely to have desirable genetic features. The energy cutoff can be fixed or set relative to the current global minimum. Typically, the relative energy cutoff is set to 70-120 kJ/mol per molecule, however it may be system dependent. A recommended best practice is to set the cutoff to prevent the addition of structures worse in energy than those in the diverse initial pool.

3.7 Termination

Because there is no unique way of converging a genetic algorithm, the user specifies simple conditions for when the code should terminate. One option is choosing to terminate the algorithm if a certain number of the best structures in the common population have not changed in a user-defined amount of iterations (e.g. if the top 20 structures have not changed in 50 iterations of the GA). This tracks whether all low-energy structures have been located in a reasonable number of iterations. Alternatively, the user may choose to terminate after the total population has reached a certain size. Additionally, the user may terminate the code manually at any time. If GAtor stops due to, e.g. wall time limits or hardware failures, there is an option to restart the code and finish all calculations leftover from the previous run before performing new selection. Code restarts can also be used strategically to modify the GA settings (e.g. to tighten the energy cutoffs or change mutation schemes) without affecting the common population of structures.

4 Applications

GAtor was used to perform crystal structure prediction for the four chemically diverse blind test targets shown in Fig. 1. The initial pool for each target was generated with Genarris[105] to create a starting population of diverse, high-quality structures.[105] The generated initial pool structures were locally optimized with the same DFT settings used in the GA and checked for duplicates. For each molecule, a variety of crossover, mutation, and selection parameters were tested using the same initial population. For testing purposes, GA searches were performed only with the same number of molecules per unit cell as the experimental structure(s). The number of molecules in the asymmetric unit was not constrained. In all cases, the experimental structures were generated as well as several other low-energy structures that may be viable polymorphs.

4.1 Target XXII

Target XXII ($C_8S_3N_4$) was selected from the sixth blind test.[24] It belongs to a unique class of compounds, called thiacyanocarbons, which only contain carbon, nitrogen, sulfur and a plurality of cyano groups.[134] The molecule contains no rotatable bonds, however it can bend about the S-S axis of the six-membered ring. The energy barrier between its chiral forms is small, leading to the appearance of many structures with planar or near-planar conformations in the computed crystalline energy landscape.[14,24] The correct crystal structure of Target XXII was generated by 12 out of 21 groups that participated in category 1 of the most recent blind test,[24] and ranked as the most stable structure by 4 groups.

Fig. 6 shows an analysis of GA runs of Target XXII with different settings. The initial pool contained 100 structures, and all runs were stopped when the number of structures added to the common population from the GA reached 550 structures. The shorthand notation is as follows: standard crossover (SC), symmetric crossover (SymC), tournament selection (T), and roulette wheel selection (R). The percent-

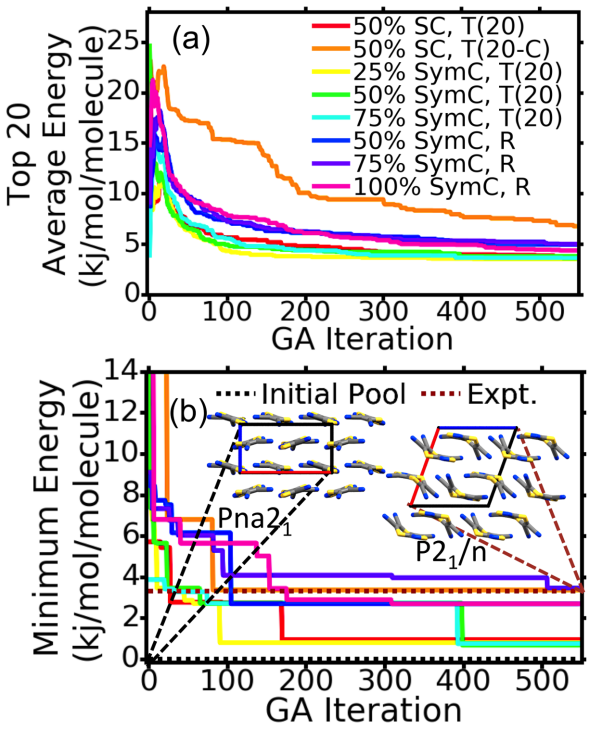


Figure 6: (a) The average energy of the top 20 Target XXII structures as a function of GA iteration and (b) the global minimum structure generated as a function of GA iteration, shown for different GA runs. S, N, and C atoms are colored in yellow, blue, and grey, respectively. The a, b, and c crystallographic lattice vectors are displayed in red, green, and blue, respectively.

age (e.g. 75%) indicates the crossover probability, with the remaining percentage (e.g. 25%) indicating mutation probability. For runs using tournament selection, the tournament size is shown in parentheses. Cluster-based fitness is denoted by a C after the selection type. Here, affinity propagation clustering was used with the descriptor given by Eq. 3, which promotes the selection of structures with under-sampled lattice parameters. Although this descriptor is simple, it provides insight into the behavior of cluster-based fitness in the GA and was successful in generating the experimental structure of Target XXII.

The average energy of the top 20 structures per GA iteration for the different runs is shown in Fig. 6, panel (a). The energies shown are relative to the global minimum structure evaluated with PBE+TS and *lower-level* numerical settings. For the 7 runs that used energy-based fitness, the average energy of the top 20 structures smoothly converges to within approximately 5 kJ/mol per molecule of the global minimum structure upon GA termination. The runs that used tournament selection had a slightly lower average energy of the top 20 structures over time compared to the runs using roulette wheel selection. The run that used clustering, depicted in orange, shows a larger average energy than the other runs and a slower, more erratic convergence of the top 20 structures to within 7 kJ/mol per molecule of the global minimum structure upon GA termination. This behavior is not unusual because the cluster-based fitness explicitly promotes under-represented structures in the population, which may have higher energies.

For all runs, the minimum energy structure as a function of GA iteration is shown in Fig. 6, panel (b). The energies of the experimental structure and the lowest energy structure in the initial pool are also indicated. The latter happened to correspond to the PBE+TS global minimum structure using *lower-level* numerical settings. We note that the initial pool produced by Genarris is not random, but rather consists of a diverse set of high-quality structures, as detailed in Ref. 105. All runs generated the experimental structure (located approximately 3.3 kJ/mol per molecule above the global minimum) but at different GA iterations. Most runs located structures lower in energy than the experimental, but only those that used tournament selection and energy-based fitness (shown in red, yellow, green, and cyan) generated the second to the global minimum structure. GA runs that used symmetric crossover, tournament selection, and energy-based fitness (shown in yellow, green, and cyan) found the experimental structure in fewer GA iterations on average than the runs that used energy-based fitness and roulette wheel selection (shown in blue, purple, and pink).

Fig. 7 depicts different evolutionary routes that generated the experimental structure in different GA runs. Each route starts from an initial pool structure and details the various

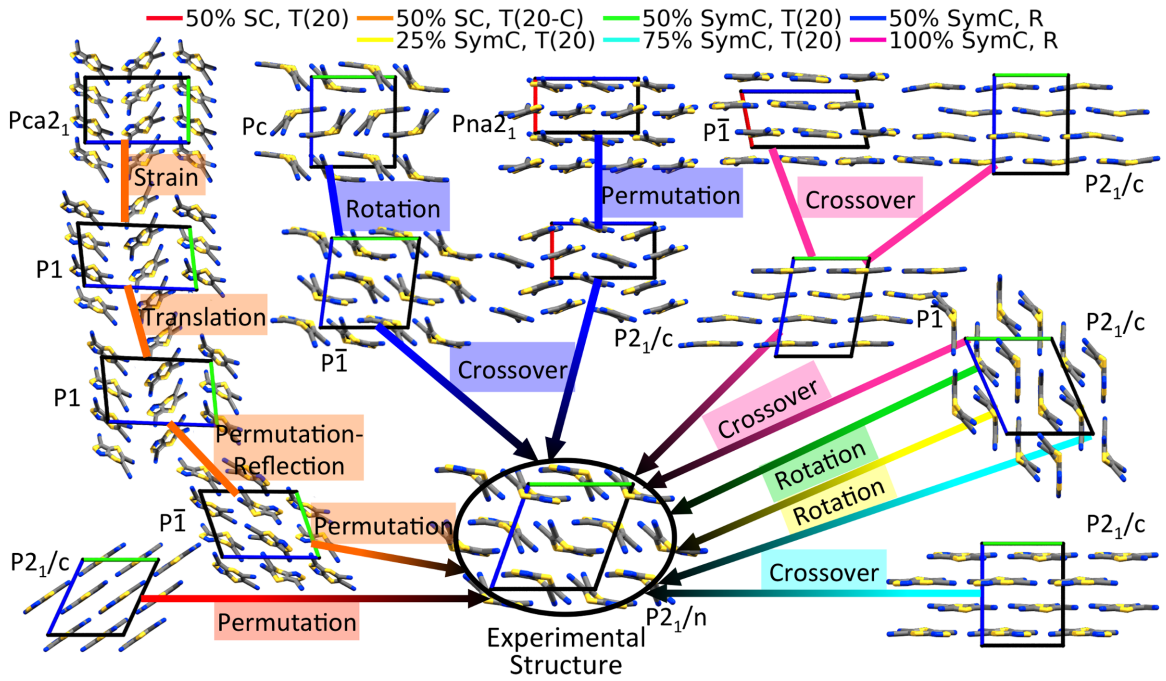


Figure 7: The different evolutionary routes which generated the experimental structure of Target XXII for different runs of the GA. The a, b, and c crystallographic lattice vectors are displayed in red, green, and blue, respectively.

breeding operations (followed by local optimizations), which ultimately generate the experimental structure. The variety of evolutionary routes and paths highlights the flexibility and randomness of the GA. The runs that used tournament selection, an energy-based fitness function, and 25% and 50% symmetric crossover (shown in yellow and green, respectively) generated the experimental structure by performing a rotation mutation to the same structure from the initial pool, followed by local optimization. The run that used 50% symmetric crossover, tournament selection, and an energy-based fitness function (shown in red) also generated the experimental structure with a single mutation by performing a permutation mutation to a structure with a planar molecular conformation, followed by local optimization, which produced the bent molecular conformation of the experimental structure. In fact, several structures with planar or near-planar conformations are found in the evolutionary routes. The run that used 50% symmetric crossover and roulette wheel selection (shown in blue) performed a final crossover between a parent structure with $P\bar{1}$ symmetry and a bent molecular conformation, with another structure with $P2_1/c$ symmetry and a nearly-planar conformation. In this case, the lattice parameters of the two parent structures were combined, the $P2_1/c$ symmetry of the parent with the planar conformer was selected, while the bent conformation of the $P\bar{1}$ parent was chosen for the asymmetric unit of the child structure. Symmetric crossover operations combining structures with planar and bent conformations were also performed in the final mating step of the runs with 100% symmetric crossover and 75% symmetric crossover, shown in pink and cyan, respectively. The run that that utilized the cluster-based fitness function, shown in orange, took a unique path to the experimental structure. A crucial mutation along this route was permutation-reflection, which introduced an inversion center and created a $P\bar{1}$, Z'=2 structure. This $P\bar{1}$ structure subsequently underwent permutation followed by local optimization to reach the experimental structure. Overall, the combination of symmetric crossover and mutation was highly effective for Target XXII.

A detailed comparison between the runs that used tournament selection and 50% percent

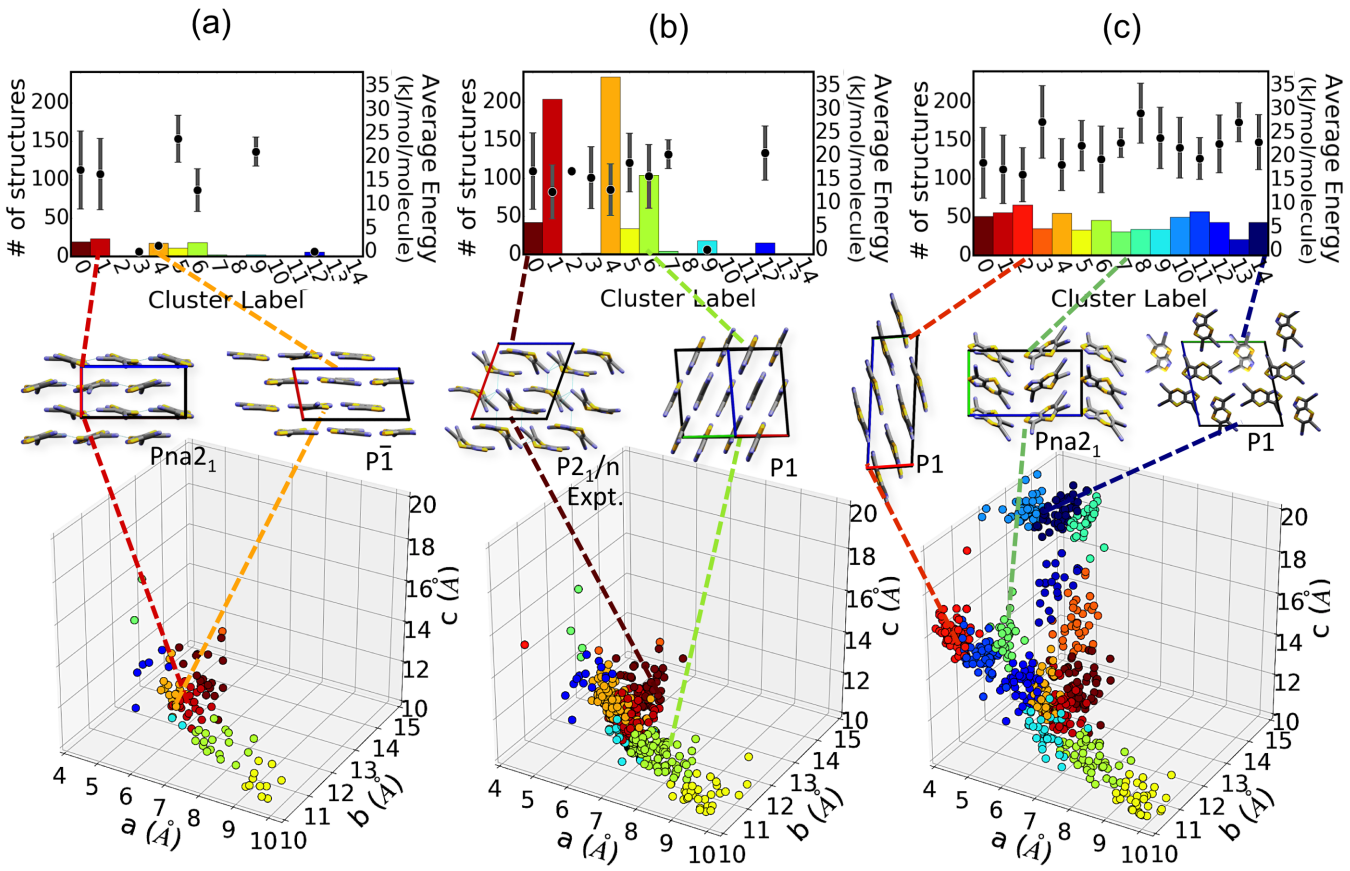


Figure 8: A comparison of the different clusters and structural motifs found in the initial pool (a), the common population evolved using energy-based fitness (b), and the common population evolved with cluster-based fitness (c). The a, b, and c crystallographic lattice vectors are displayed in red, green, and blue, respectively.

standard crossover, with and without cluster-based fitness, is shown in Fig. 8. The final structures produced from cluster-based fitness run, including the initial pool, formed 15 clusters, as computed using affinity propagation[125] with the lattice parameter based descriptor and a Euclidean metric. The structures from the run which used energy-based fitness were assigned to one of the 15 clusters from the cluster-based run. Panel (a) depicts the population of the initial pool, while panels (b) and (c) depict the independent evolution of the initial population for the energy and cluster-based fitness runs, respectively. The initial pool contained several low-energy structures with planar or near-planar conformations, which tend to have shorter a parameters than structures with bent conformations, such as the experimental structure. Panel (b) reveals initial pool bias and genetic drift in the run that used energy-based fitness. Initial pool bias is evident from the fact that the GA hardly explores regions not represented in the initial pool. Genetic drift is apparent from the preferential exploration of the clusters labeled 1, 4, and 6, compared to other clusters represented in the initial pool. These clusters contain layered structures with planar or near-planar conformations, examples of which are shown in panels (a) and (b). Such structures likely correspond to large, shallow basins of the energy landscape that are frequently visited. In addition, these structural motifs are systematically favored by PBE+TS, as discussed in detail in Ref.[14] and below. Cluster 0, which contains structures with a bent conformation, including the experimental structure, is sampled less frequently, possibly because such structures correspond to narrow wells in the potential energy surface that are more difficult to locate. Panel (c) demonstrates that evolution-

ary niching helps overcome initial pool biases and genetic drift. In this case, a more uniform sampling of the potential energy landscape is achieved. Clusters 1, 4, and 6 have fewer members than in the energy-based run, while cluster 0 has more members. Evidently, for Target XXII, utilizing cluster-based fitness with the lattice parameter descriptor suppressed the over-selection of crystal structures with planar or near-planar conformations. This descriptor was effective for Target XXII because in this case the unit cell shape is correlated with the molecular conformation. Furthermore, several clusters outside the boundaries of the initial pool were only explored with the cluster-based fitness function. These clusters include, for example, structures with more elongated unit cell shapes (a representative structure is shown for cluster 2). This demonstrates that evolutionary niching can correct initial pool biases and explore novel regions of the potential energy surface (this may be particularly useful if the initial pool is not as optimal as the pools produced by Genarris). However, it does so at the price of an increased computational cost, and in this case generates more high-energy structures that may or may not be useful for the purpose of maintaining diversity.

All structures generated were combined into a final set of 200 unique structures evaluated with PBE+TS and *lower-level* numerical settings. The structures were re-relaxed using PBE+TS with *higher-level* numerical settings and subsequently re-checked for duplicates. The final 100 PBE+TS structures were then re-ranked with PBE+MBD and PBE0+MBD, as shown in panel (a) of Fig. 9. The re-ranking of Target XXII structures generated within the sixth CSP blind test has been discussed extensively in Ref.[14] It has been demonstrated therein that different exchange-correlation functionals and dispersion methods systematically favor specific packing motifs. The experimental structure (which was not generated during the blind test due to the constraints imposed on the unit cell angles in the preliminary version of GAtor used therein) was ranked as the top structure only by PBE0+MBD. The same trends are observed here. Within the present study, the top 100 Z=4 structures are located within relative energy windows of 6.7, 7.5, and 9.6 kJ/mol per molecule using PBE+TS, PBE+MBD, and PBE0+MBD, respectively. The number of structures generated in these intervals shows significant improvement compared to our submission to the sixth blind test. In particular, an important low-energy structure (ranked as #3 by PBE0+MBD) was located in the present study. These improvements may be attributed to a number of factors including updated crossover, mutation, and similarity checks, as well as the use of a more diverse and comprehensive initial pool as generated by Genarris.[105] Panel (b) of Fig. 9 shows the PBE0+MBD energy versus density for the structures. Structures with bent molecular conformations, including the experimental structure, have lower densities than structures with planar or near-planar conformations.

4.2 Target II

Target II (C_5H_3NOS) was selected from the second blind test.[20,135] At the time, no participating groups used *ab initio* methods for the structure prediction of this molecule, and only one group submitted the correct experimental structure, ranking it as their second most thermodynamically stable structure. Fig. 10 shows an analysis of the different GA runs that successfully generated the experimental crystal structure of Target II. The initial pool contained 45 structures. Each run was stopped when the number of additions to the common pool reached 350. The average energy of the top 20 structures as a function of GA iteration is shown in panel (a) of Fig. 10. All energies shown are relative to the energy of the global minimum structure as ranked by PBE+TS with the *lower-level* numerical settings used within the GA. All runs converged the top 20 structures to within 4 kJ/mol per molecule when the GA was terminated. The run that used 100% pure mutation (denoted by 100% M) with tournament selection, shown in purple, consistently exhibited the lowest average energy of the top 20 structures. In panel (b), the minimum energy structure added by the GA as a function

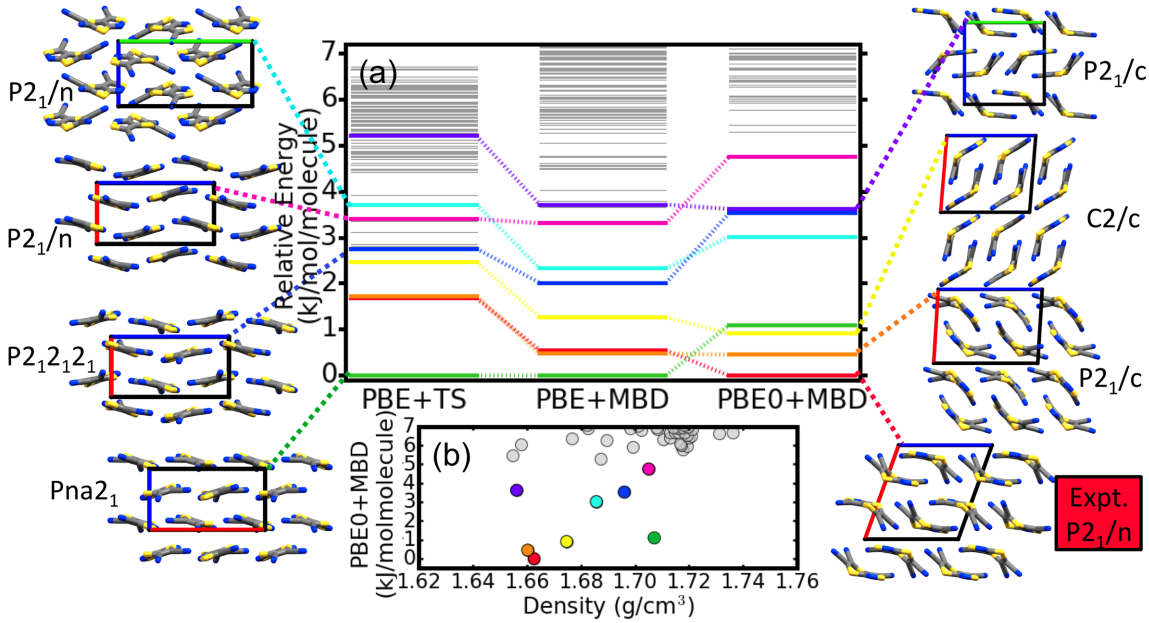


Figure 9: (a) The relative total energies as obtained by different dispersion-inclusive DFT methods and (b) the PBE0+MBD energy versus density of putative crystal structures of Target XXII. The top 8 predicted structures, as ranked by PBE0+MBD, are shown in color. The a, b, and c crystallographic lattice vectors are displayed in red, green, and blue, respectively.

of GA iteration is shown along with the lowest-energy structure from the initial population and the experimental structure. All runs generated the experimental structure, as well as at least one other structure lower in energy. Two runs, shown in orange and yellow, generated the most structures with lower energies than the experimental. These runs used 50% standard crossover and tournament selection with tournament sizes of 10 and 20, respectively. Strain mutations were particularly effectively at generating new low-energy structures for this target.

All structures produced by the GA runs were combined into a final set of 200 non-duplicate structures as evaluated with PBE+TS and *lower-level* numerical settings. The structures were re-relaxed with PBE+TS and *higher-level* numerical settings and subsequently re-checked for duplicates. The final 100 PBE+TS structures were then re-ranked with PBE+MBD and PBE0+MBD, as shown in panel (a) of Fig. 11. The top 10 structures as ranked by PBE0+MBD are highlighted in color. The top 100 structures are found in relative energy windows of 5.3, 5.5, and 6.1 kJ/mol per molecule using PBE+TS, PBE+MBD, and PBE0+MBD, respectively. Interestingly, the experimental structure becomes less stable with the increasingly accurate DFT methods and is ranked as #10 with PBE0+MBD. Structures ranked as #4-#10 with PBE0+MBD display layered packing motifs in several different space groups, within an energy window of approximately 0.6 kJ/mol per molecule. The layered motif of Target II is characterized by hydrogen-bonds that form 1D chains between the hydroxyl group of one molecule and the nitrile group of another (O−H···N) which are stacked on top of one another as shown in Fig. 12. The prediction of nearly energetically degenerate crystal structures consisting of the same sheet stacked in different ways is a common phenomena.[12,136,137] While the structures ranked #4-#10 are determined as distinct lattice energy minima, they likely converge to a lower number of minima on the free energy surface.[12,138]

The structure ranked as #3 by PBE0+MBD (shown in yellow) was not reported by any participating group during the second blind test and has the highest computed density of the low-energy structures, as shown in panel (b) of Fig. 11. This structure contains the same 1D hydrogen-bonded patterns as the experimental structure, but with zig-zag stacking. Ref.

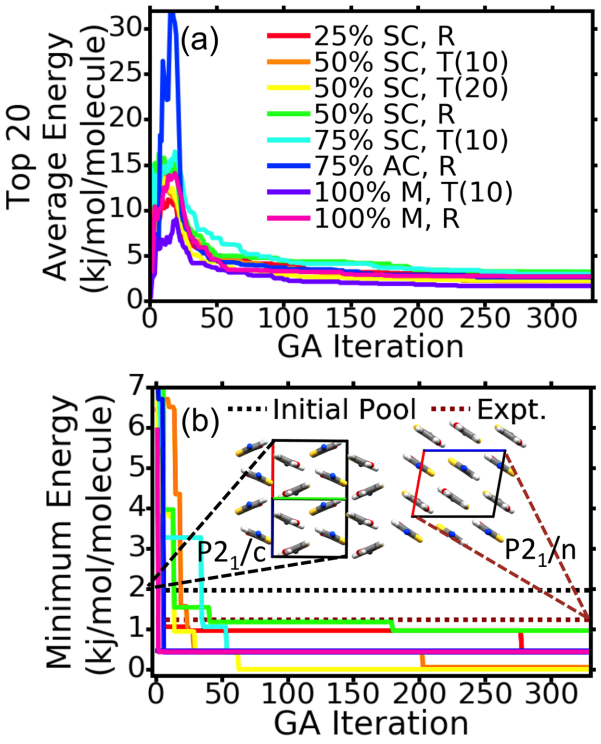


Figure 10: (a) The average energy of the top 20 Target II structures as a function of GA iteration and (b) the global minimum structure generated as a function of GA iteration, shown for different GA runs. S, N, O, C, and H atoms are colored in yellow, blue, red, grey, and white, respectively. The a, b, and c crystallographic lattice vectors are displayed in red, green, and blue, respectively.

139 later performed an additional CSP study on Target II using a tailor-made force field[140] within the GRACE software. This methodology has been highly successful at CSP and predicted all five targets in the most recent blind test.[24] Searching structures with Z'=1, this study predicted the #3 PBE0+MBD zigzag structure for the first time, ranking it as the global minimum structure when re-ranked using DFT with a pairwise dispersion correction.[141] Furthermore, it was shown that this form became more stable with increasing pressure, suggesting it could be an unobserved high-pressure polymorph of Target II. Our #2 $P2_1$ PBE0+MBD structure with a scaffold packing motif was also discussed in Ref. 139 and ranked as #3. Ref. 138 computed the relative stability of the $P2_1$ scaffold structure and the experimental structure using the B86bPBE density functional[108,142] combined with the exchange-hole dipole moment (XDM)[143,144] dispersion model and found the $P2_1$ scaffold structure to be more stable. When a quasi-harmonic thermal correction was further included, the experimental structure was ranked as the more stable structure.

The $P\bar{1}$, Z'=2 structure with a scaffold packing motif ranked as the global minimum by all three DFT methods has not been previously reported in any CSP studies of Target II. It has a higher computed density than the experimental form, as shown in panel (b) of Fig.11. A detailed comparison between the packing motif of the #1 PBE0+MBD scaffold structure and the experimental structure is shown in Fig. 12. Panels (a) and (b) compare the the experimental structure and $P\bar{1}$, Z'=2 structure, respectively, projected along the a crystallographic direction. From this vantage point, a similar motif of stacked pairs of molecules related by inversion symmetry is observed. The experimental structure exhibits O−H···N hydrogen-bonds, colored in fuchsia, which connect adjacent molecules along the b direction to form 1D chains. The molecules comprising the chains are cross-linked by close S···O contacts, colored in orange. The $P\bar{1}$, Z'=2 structure shows the O−H···N hydrogen bonded chains cross-linked by C−H···O and C−H···S close contacts shown in cyan and green, respectively. Panels (c) and (d) compare the two structures projected along the b direction. The experimental structure exhibits stacked layers of the 1D hydrogen-bonded chains. The $P\bar{1}$, Z'=2 structure is characterized by alternating stacked layers related by a tilt of approximately 45° about the c direction, resulting in C−H···O and C−H···S close contacts between multiple neighboring molecules shown in cyan and green, respectively. Panels (e) and (f) compare the two structures projected along the c direction. This provides another viewpoint of the experimental structure's stacked layers, while the $P\bar{1}$, Z'=2 panel (f) showcases a scaffold packing motif.

The #1 PBE0+MBD scaffold structure would not have been found without GAtor's ability to generate crystal structures with Z′ >1

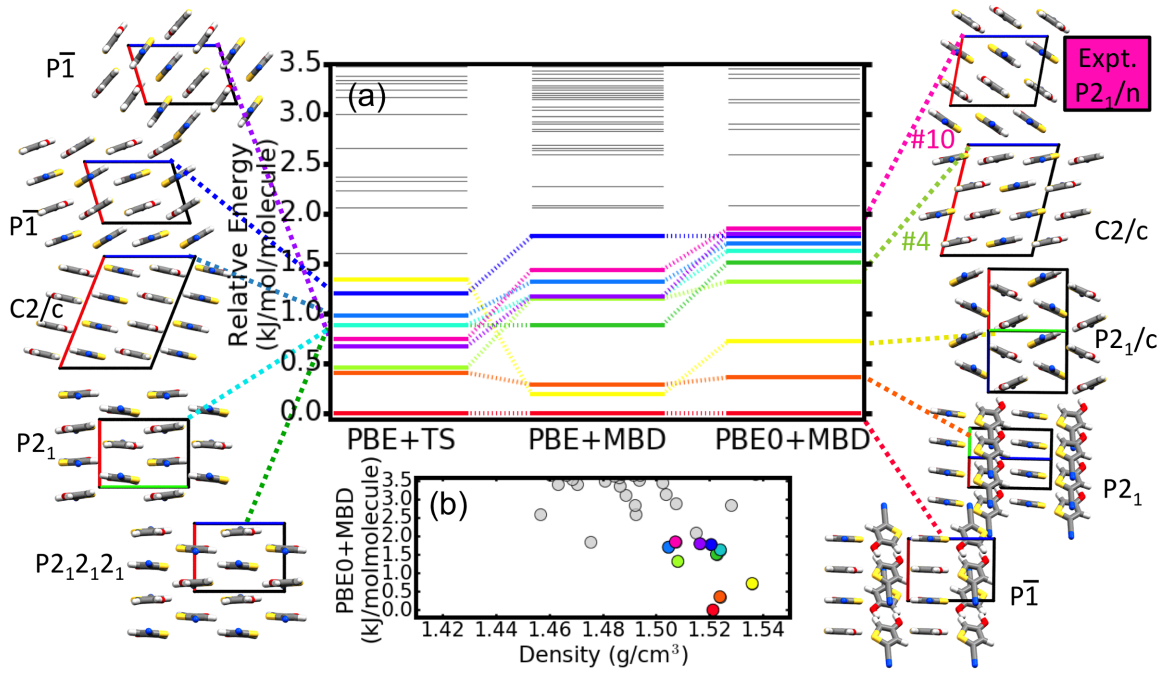


Figure 11: (a) The relative total energies as obtained by different dispersion-inclusive DFT methods and (b) the PBE0+MBD energy versus density of putative crystal structures of Target II. The top 10 predicted structures, as ranked by PBE0+MBD, are shown in color. The a, b, and c crystallographic lattice vectors are displayed in red, green, and blue, respectively.

through the various crossover and mutation operators. As emphasized in Ref. 145, stable crystal structures are formed when intermolecular interactions are optimized through close packing. While these requirements favor highly symmetric structures, symmetry can be sacrificed in favor of forming particularly stabilizing intermolecular interactions.[145–147] Future investigations incorporating finite temperature and pressure effects will add further insight into the relative stability of the #1 PBE0+MBD scaffold structure and the other predicted low-energy structures, including the experimental structure.

4.3 Target XIII

Target XIII ($C_6H_2Br_2ClF$) was selected from the fourth blind test,[22] in which it was categorized as a rigid molecule containing challenging elements for modeling methods. Target XIII contains three different halogens, allowing for a variety of halogen bonds. Many common electronic structure theory methods do not accurately capture halogen bonds because they require a precise treatment of both electrostatic and dispersion interactions.[148–152] During the fourth blind test, the correct experimental structure was successfully predicted and ranked as #1 by 4/14 groups. The methodology used in one of the successful submissions is further detailed in Ref. 153.

Indeed, predicting the correct crystal structure of Target XIII proved challenging. The various crossover, mutation, and selection settings used in different GA runs of Target XIII are shown Fig. 13. The initial pool for all runs contained 48 structures, and the runs were stopped after 1400 iterations, the first 900 of which are shown. Of the six runs attempted, only one run (50% SymC, R), colored in green was able to locate the experimental structure, although all runs found crystal structures lower in energy than the experimental using PBE+TS and *lower-level* numerical settings. Panel (a) shows the average energy of the top 40 structures as a function of GA iteration, relative to the global minimum energy with PBE+TS and *lower-level* settings. The run that used standard crossover (50% SC, R), colored in red con-

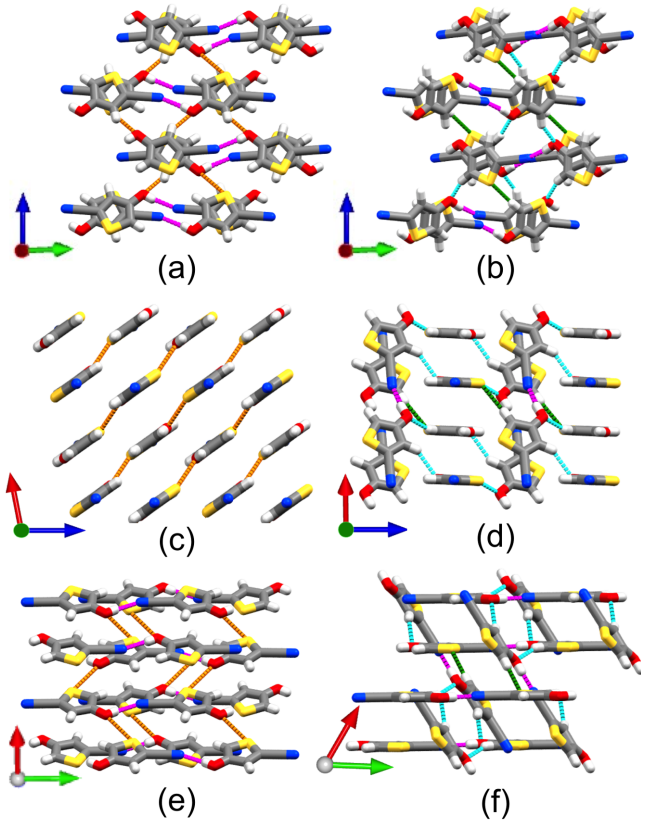


Figure 12: A detailed comparison between the experimental crystal structure (left) and the predicted #1 PBE0+MBD crystal structure (right) for Target II depicted along the three crystallographic directions. The a, b, and c crystallographic basis vectors are shown in red, green, and blue, respectively. Close intermolecular contacts less than the sum of their respective vdW radii minus 0.1 Å are colored as follows: O−H· · ·N (fuchsia), S· · ·O (orange), C−H· · ·O (cyan), C−H· · ·S (green).

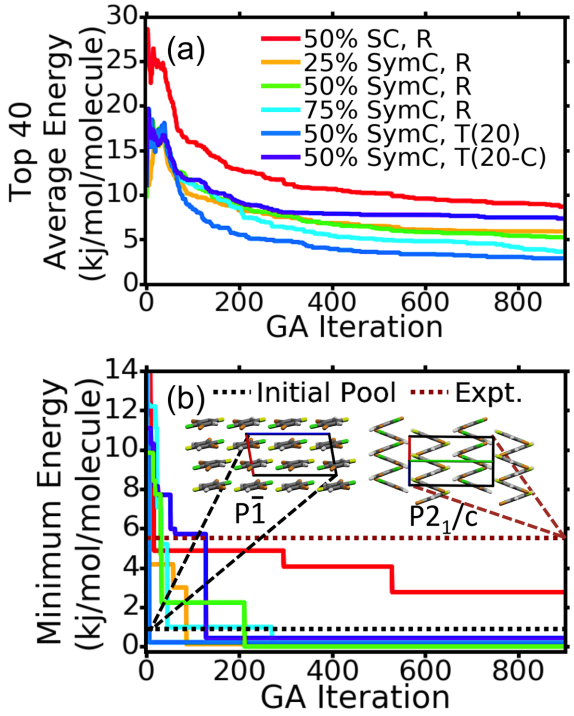


Figure 13: (a) The average energy of the top 40 Target XIII structures and (b) the global minimum structure produced as a function of GA iteration for different GA runs. C, H, Br, Cl, and F atoms are colored in grey, white, brown, green, and yellow, respectively. The a, b, and c crystallographic lattice vectors are displayed in red, green, and blue, respectively.

sistently had the highest average energy, even higher than the run that used cluster-based fitness with affinity propagation and the lattice parameter based descriptor, shown in indigo. Panel (b) shows the minimum energy structure as a function of GA iteration. All runs converged the top structure to within 1 kJ/mol per molecule within 300 iterations, except for the run that used standard crossover. For this target symmetric crossover was essential in producing low-energy structures.

All structures generated were combined into a final set of 200 unique structures evaluated with PBE+TS and *lower-level* numerical settings. The top 150 structures were re-relaxed with PBE+TS with *higher-level* numerical settings and subsequently re-checked for duplicates. The final top 90 structures as ranked by PBE+TS and *higher-level* settings, were then re-ranked with PBE+MBD and PBE0+MBD. The top 90 structures are located within relative energy windows of 6.8, 7.8, and 6.5 kJ/molecule per molecule when ranked by PBE+TS, PBE+MBD, and PBE0+MBD, respectively. Panel (a) of Fig. 14 shows the ranking of the structures found within a window of 4.2 kJ/mol per molecule of the global minimum. The top 8 crystal structures as ranked by PBE0+MBD are highlighted in color. After optimization with PBE+TS and *higher-level* numerical settings, the experimental structure is ranked as #1. It is consistently predicted as the

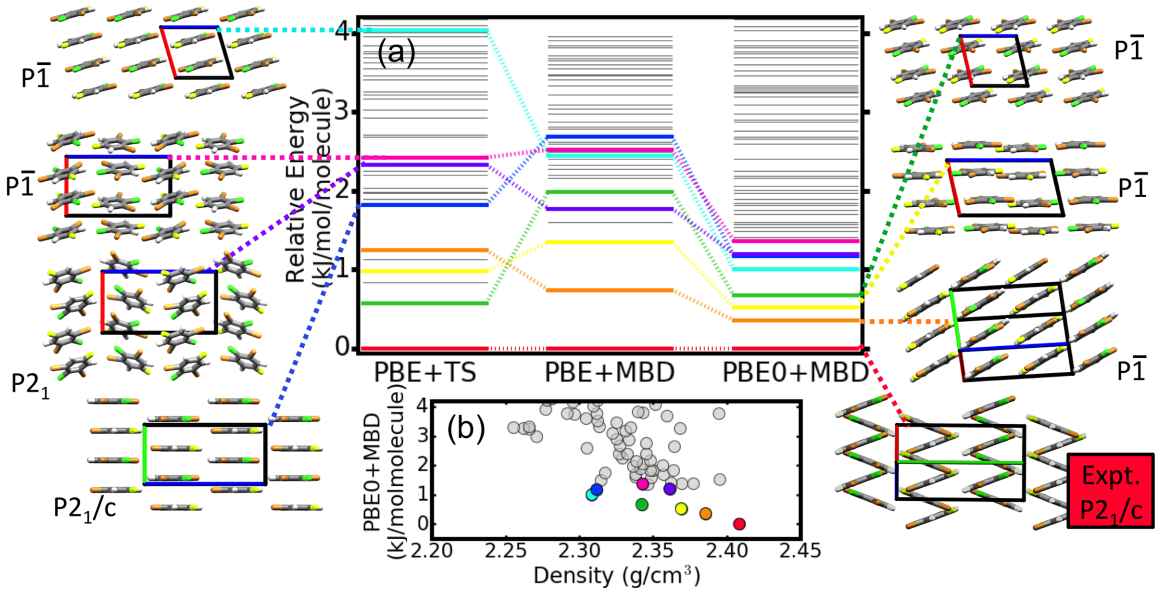


Figure 14: (a) The relative total energies as obtained by different dispersion-inclusive DFT methods and (b) the PBE0+MBD energy versus density of putative crystal structures of Target XIII. The top 8 predicted structures, as ranked by PBE0+MBD, are shown in color. The a, b, and c crystallographic lattice vectors are displayed in red, green, and blue, respectively.

most stable crystal structure by PBE+MBD and PBE0+MBD. Focusing on the top 8 crystal structures as ranked by PBE0+MBD, only the experimental structure contains a zig-zag packing motif. Additionally, 4/8 of the top structures have Z'=2. Panel (b) of Fig. 14 shows the PBE0+MBD energy versus density of the top structures. This reveals the experimental structure with the zig-zag motif has the highest density. For the experimental structure, close bromine-bromine contacts are found perpendicular to the zig-zag stacking direction, while 7/8 of the other top structures generated show π-stacking and/or close halogen bonds that stabilize the stacking of the layers.

Although many low-energy structures were generated, 5/6 of the GA runs did not successfully locate the experimental structure. This may be attributed to two primary factors. First, it is possible that the *lower-level* numerical settings used to save computational time in the GA search, were not sufficiently accurate for this halogenated molecule. When using PBE+TS and *lower-level* numerical settings, the experimental structure was nearly 6 kJ/mol per molecule higher than the global minimum, and ranked as #39 when all structures generated from the different GA runs were combined. When these structures were postprocessed with PBE+TS and *higher-level* numerical settings, the experimental structure was ranked as #1. As lower energy structures have a higher probability of being selected, this could have systematically biased the searches. This highlights the complications that may arise when using a hierarchical approach. Second, while most low-energy structures of Target XIII have a layered packing motif, the experimental structure has a unique zig-zag packing motif and an oblong unit cell. Such oblong unit cells were rarely generated in the search. In fact, even the run that used cluster-based fitness with the lattice parameter descriptor failed to locate the experimental structure. Although candidate child structures with similar lattices to the experimental were frequently generated in this run, they were subsequently rejected by the geometric and energetic constraints before local optimization. This suggests that the experimental structure is located in a narrow well in the potential energy surface, while layered structures exist in wider, more-accessible basins. When studying halogen-bonded systems in the future, it may be beneficial to use cluster-based fitness

with a descriptor based on halogen-halogen or hydrogen-halogen intermolecular contacts.

4.4 Target I

Target I (C_6H_6O) was selected from the second blind test.[20,135] It has two reported polymorphs, a stable form, which crystallizes in $P2_1/c$ with Z=4, and a metastable form which crystallizes in *Pbca* with Z=8. At the time of the second blind test, no participating groups submitted the more stable Z=4 form. 4/11 groups submitted the metastable Z=8 form, with 3/4 groups ranking it as the most stable structure.

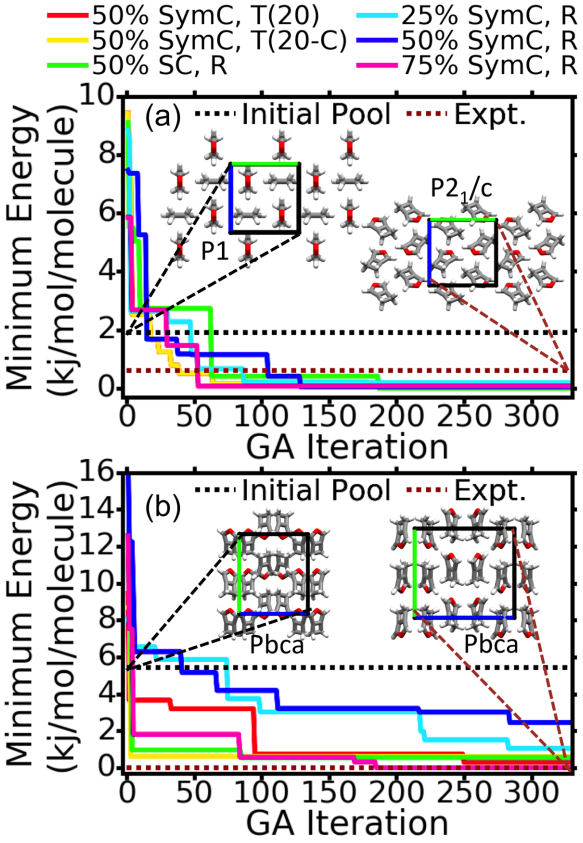


Figure 15: The global minimum structure produced by the GA runs as a function of GA iteration for runs that used (a) Z=4 and (b) Z=8. C, H, and O atoms are colored in grey, white, and red, respectively. The a, b, and c crystallographic lattice vectors are displayed in red, green, and blue, respectively.

For Target I, independent GA searches were conducted starting from initial pools with Z=4 and Z=8. These contained 45 and 96 structures, respectively. The GA runs were stopped when the number of additions to the common pool reached 650 and 350, respectively. During evolution, the Z=4 runs also generated structures with Z=2, and the Z=8 runs generated structures with Z=4 and Z=2. The minimum energy as a function of GA iteration, relative to the global minimum using PBE+TS with *lower-level* numerical settings, is shown in Fig. 15, panels (a) and (b), for the Z=4 and Z=8 runs, respectively. For the Z=4 runs, the convergence behavior of the minimum energy structure was similar for all settings tested, including the run that used lattice parameter based clustering, shown in orange. All runs located structures lower in energy than the Z=4 experimental polymorph at this level of theory. For the Z=8 runs, the runs that used 25% and 50% symmetric crossover with roulette wheel selection were slower to converge, and did not locate the Z=8 polymorph when the GA was stopped.

All structures produced by the Z=4 and Z=8 GA runs were combined into a final set of 200 unique structures, as evaluated with PBE+TS and *lower-level* numerical settings. Supercells were allowed in the pymatgen duplicate check. The final top 100 structures were re-relaxed using PBE+TS with higher-level settings and subsequently re-ranked using PBE+MBD. The structures located within 2 kJ/mol per molecule of the global minimum are shown in panel (a) of Fig. 16. The top 6 structures as ranked by PBE+MBD were also re-ranked using PBE0+MBD and are highlighted in color. Of these top 6 structures, 4/6 display similar packing motifs to the metastable *Pbca* polymorph, shown in green with co-facial dimers oriented in opposite directions, stacked in slightly different ways. To highlight structural differences, intermolecular close-contacts are displayed in cyan.

The metastable Z=8 *Pbca* polymorph is ranked as #1 with PBE+TS, #4 with PBE+MBD, and #3 when re-ranked with PBE0+MBD. Using PBE+TS and PBE0+MBD, this polymorph is determined to be practically energetically degenerate with the putative Z=8 $P2_1/c$ structure, shown in yellow. However, the Z=8 $P2_1/c$ structure

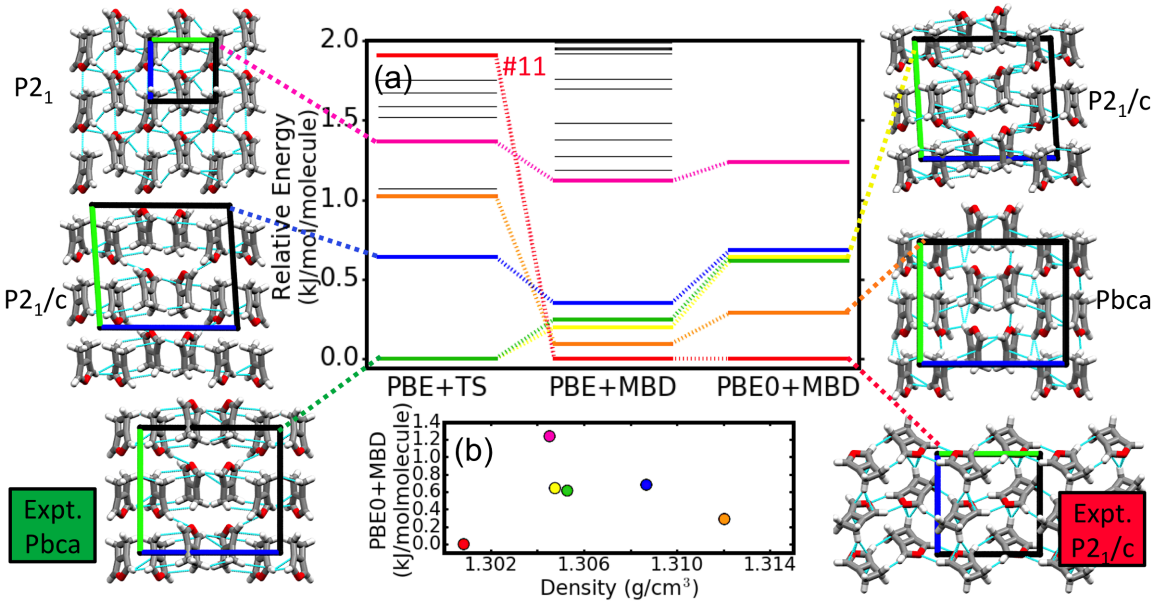


Figure 16: (a) The relative total energies as obtained by different dispersion-inclusive DFT methods and (b) the PBE0+MBD energy versus density of selected crystal structures of Target I. The top 6 predicted structures, as ranked by PBE+MBD, are highlighted in color. Intermolecular contacts less than the sum of vdW radii are shown in cyan. The a, b, and c crystallographic lattice vectors are displayed in red, green, and blue, respectively.

has Z'=2 and a slightly different lattice from the metastable polymorph, and hence was determined to be a unique lattice energy minima. The experimental $P2_1/c$ polymorph with Z=4, highlighted in red, is ranked as #11 with PBE+TS, but #1 with PBE+MBD and PBE0+MBD. There is no significant re-ranking between PBE+MBD and PBE0+MBD for the structures considered. The relative energy differences between these structures increased when re-ranked by PBE0+MBD, as compared to PBE+MBD. Panel (b) of Fig. 16 shows the PBE0+MBD energy versus density of the highlighted structures. The six structures have very similar densities, but the most stable $P2_1/c$ experimental structure has the lowest density.

Several computational studies conducted after the second blind test[154–156] consistently ranked the Z=8 *Pbca* polymorph as the most stable form. However, attempts at its recrystallization only lead to the stable Z=4 $P2_1/c$ form. Ref. 67 suggests that the Z=8 *Pbca* structure is located on a saddle point of the potential energy surface, and that symmetry breaking produces a stable Z'=2 structure. This could be the Z'=2 $P2_1/c$ structure, colored in yellow and ranked as #4 with PBE0+MBD, as discussed above. It should be noted, however, that the nature of the potential energy landscape, including whether certain structures are determined as minima or saddle points, may depend strongly on the energy method used.[157,158] In the present study, PBE+MBD and PBE0+MBD rank the experimental Z=4 $P2_1/c$ structure as the most stable polymorph. This highlights the importance of accounting for many-body dispersion interactions and long-range screening effects in the MBD method. Ref. 138 also computed the $P2_1/c$ experimental structure as more stable than the *Pbca* form using B86bPBE-XDM.

5 Conclusion and Best Practices

We have introduced GAtor, a first principles genetic algorithm for molecular crystal structure prediction. GAtor currently interfaces with FHI-aims and is optimized for HPC environments. The code offers a variety of features that enable the user to customize the GA search set-

tings, including energy-based and cluster-based fitness (evolutionary niching), roulette wheel and tournament selection, symmetric and standard crossover, different mutation schemes, and various tunable parameters related to energy cutoffs, similarity checks, and geometric constraints. GAtor's crossover and mutation operators, specifically tailored for molecular crystals, provide a balance between exploration and exploitation. These operators enable the generation and exploration of high Z' structures.

GAtor was applied to predict the structures of a chemically diverse set of four past blind test targets. The known structures of all four targets were successfully predicted, as well as several additional low-energy structures. Different GA settings were found to be more effective for different targets. Target XXII contains only C, N, and S atoms and has a small energy barrier between its two enantiomers, related by a bending degree of freedom. For this target, symmetric crossover and tournament selection were particularly effective. Evolutionary niching with respect to a descriptor based on lattice parameters uniformly explored the potential energy surface, including regions outside the initial pool, and suppressed the oversampling of structures with a planar molecular conformation (genetic drift). Target II forms various hydrogen-bonds. Its known experimental structure was located with a variety of GA settings, including runs that purely used mutations. For this molecule, standard crossover was more effective than symmetric crossover. Target XIII contains several halogens (Br, Cl, F), which make it challenging due to the presence of halogen bonds. In addition, the experimental structure comprises a ziz-zag packing motif unlike the layered packing motifs found in most of the low-energy structures in the population. This may explain why the experimental structure was generated only once. For Target XIII, symmetric crossover was critical for the production of low-energy structures. Target I forms mainly weak C$\cdots$H and C$-$H$\cdots$O interactions. It has two known polymorphs with Z=4 and Z=8, the latter of which is a less stable "disappearing polymorph". All GA settings tested were found to be equally effective in locating the Z=4 structure. For the Z=8 structure, the combination of 25% or 50% symmetric crossover with roulette wheel selection was less effective.

Low-energy structures found in different GA runs were grouped together, re-relaxed, and re-ranked with increasingly accurate dispersion-inclusive DFT methods: PBE+TS, PBE+MBD, and PBE0+MBD. For Target XIII, all three methods ranked the experimental structure as #1. For Target I, PBE+MBD and PBE0+MBD correctly ranked the Z=4 polymorph as #1 and the Z=8 polymorph as less stable, at #4 and #3, respectively, and very close in energy to a structure with Z'=2 and a similar packing motif. The MBD method was instrumental in obtaining the correct ordering of the two known polymorphs of Target I based solely on lattice energy without considering vibrational and thermal contributions. For Target XXII, only PBE0+MBD ranked the experimental structure as #1. Target II is an exception because the relative energy of its experimental structure increases, rather than decreases, with increasing accuracy. It is ranked as #10 with PBE0+MBD. The structures ranked #4-#9 exhibit a variety of layered packing motifs, similar to the experimental structure. The structure consistently ranked as #1 with all three methods was predicted for the first time using GAtor. It is a Z'=2 structure with $P\bar{1}$ symmetry and a scaffold packing motif, whose lattice energy is 1.8 kJ/mol per molecule lower than the known experimental form. The #2 structure, which also has a scaffold packing motif, and the #3 structure with a zig-zag packing motif have been previously reported by others. Several of the low-lying putative structures of Target II have higher densities than the observed structure, therefore it may be possible to crystallize them under high pressure conditions. This may motivate further experimental investigations of Target II. Further computational studies considering finite temperature and pressure effects may provide additional insight into the relative stability of the putative low-energy structures identified here and the possibility of growing them experimentally.

Several best practices for the usage of GAtor

have emerged from the results reported here. First, because the GA exhaustively explores regions of the configuration space represented in the initial pool (unless evolutionary niching is used), it is recommended to start GAtor from a carefully crafted initial pool, containing a diverse set of high-quality structures. Such an initial pool may be generated by Genarris[105] or by other means. Second, rather than running GAtor with predetermined settings for a large number of iterations, we recommend running GAtor with several different settings for a smaller number of iterations, and then combining the structures found in all searches for post-processing. As each system is unique and it is difficult to know *a priori* which settings will be the most effective, running the GA with different settings increases the likelihood of success. Third, it is recommended to use evolutionary niching in at least one of the runs. Overall, the goal is to locate all the low-lying minima including those found in disconnected, hard to reach regions of the potential energy surface. For this reason, cluster-based fitness can be a useful tool for uniformly sampling the potential energy landscape and for overcoming initial pool biases and selection biases (genetic drift). In the future, we plan to implement increasingly sophisticated capabilities in GAtor to treat increasingly complex systems. We expect GAtor to be a useful tool for the computational chemistry, materials science, and condensed matter communities.

Acknowledgement Work at CMU was funded by the National Science Foundation (NSF) Division of Materials Research through grant DMR-1554428. An award of computer time was provided by the Innovative and Novel Computational Impact on Theory and Experiment (INCITE) program. This research used resources of the Argonne Leadership Computing Facility, which is a DOE Office of Science User Facility supported under Contract DE-AC02-06CH11357.

Supporting Information Available: The supporting information provides a comparison between the experimental structures predicted by GAtor and the published experimental forms, including RMS differences computed with the Mercury[159] software. For each target, CIF files and total energies of the structures used for re-ranking are also included. This material is available free of charge via the Internet at `http://pubs.acs.org/`.

References

(1) Bernstein, J. *Polymorphism in molecular crystals*; Oxford University Press, 2002; Vol. 14.

(2) Day, G. M.; S Motherwell, W. D.; Jones, W. A strategy for predicting the crystal structures of flexible molecules: the polymorphism of phenobarbital. *Phys Chem Chem Phys* **2007**, *9*, 1693–704.

(3) Reilly, A. M.; Tkatchenko, A. Role of dispersion interactions in the polymorphism and entropic stabilization of the aspirin crystal. *Phys Rev Lett* **2014**, *113*, 055701.

(4) Elder, D. P.; Patterson, J. E.; Holm, R. The solid-state continuum: A perspective on the interrelationships between different solid-state forms in drug substance and drug product. *Journal of Pharmacy and Pharmacology* **2015**, *67*, 757–772.

(5) Reese, C.; Bao, Z. Organic single-crystal field-effect transistors. *Materials Today* **2007**, *10*, 20–27.

(6) Hasegawa, T.; Takeya, J. Organic field-effect transistors using single crystals. *Science and Technology of Advanced Materials* **2009**, *10*, 024314.

(7) Bergantin, S.; Moret, M. Rubrene polymorphs and derivatives: The effect of chemical modification on the crystal structure. *Crystal Growth and Design* **2012**, *12*, 6035–6041.

(8) Cudazzo, P.; Gatti, M.; Rubio, A. Excitons in molecular crystals from first-principles many-body perturbation the-

ory: Picene versus pentacene. *Phys. Rev. B* **2012**, *86*, 195307.

(9) Cudazzo, P.; Sottile, F.; Rubio, A.; Gatti, M. Exciton dispersion in molecular solids. *J Phys Condens Matter* **2015**, *27*, 113204.

(10) Panina, N.; Leusen, F. J. J.; Janssen, F. F. B. J.; Verwer, P.; Meekes, H.; Vlieg, E.; Deroover, G. Crystal structure prediction of organic pigments: quinacridone as an example. *J Appl Crystallogr* **2007**, *40*, 105–114.

(11) Fitzgerald, M.; Gardiner, M. G.; Armitt, D.; Dicinoski, G. W.; Wall, C. Confirmation of the molecular structure of tetramethylene diperoxide dicarbamide (TMDD) and its sensitiveness properties. *J Phys Chem A* **2015**, *119*, 905–10.

(12) Price, S. L. Why don't we find more polymorphs? *Acta Crystallographica Section B: Structural Science, Crystal Engineering and Materials* **2013**, *69*, 313–328.

(13) Price, S. L.; Braun, D. E.; Reutzel-Edens, S. M. Can computed crystal energy landscapes help understand pharmaceutical solids? *Chem. Commun.* **2016**, *52*, 7065–7077.

(14) Curtis, F.; Wang, X.; Marom, N. Effect of packing motifs on the energy ranking and electronic properties of putative crystal structures of tricyano-1,4-dithiino[c]-isothiazole. *Acta Crystallogr B Struct Sci Cryst Eng Mater* **2016**, *72*, 562–70.

(15) Wang, X.; Garcia, T.; Monaco, S.; Schatschneider, B.; Marom, N. Effect of crystal packing on the excitonic properties of rubrene polymorphs. *CrystEngComm* **2016**, *18*, 7353–7362.

(16) Giri, G.; Verploegen, E.; Mannsfeld, S. C. B.; Atahan-Evrenk, S.; Kim, D. H.; Lee, S. Y.; Becerril, H. A.; Aspuru-Guzik, A.; Toney, M. F.; Bao, Z. Tuning charge transport in solution-sheared organic semiconductors using lattice strain. *Nature* **2011**, *480*, 504–508.

(17) Mei, J.; Diao, Y.; Appleton, A. L.; Fang, L.; Bao, Z. Integrated materials design of organic semiconductors for field-effect transistors. *Journal of the American Chemical Society* **2013**, *135*, 6724–6746.

(18) Diao, Y.; Tee, B. C.-K.; Giri, G.; Xu, J.; Kim, D. H.; Becerril, H. A.; Stoltenberg, R. M.; Lee, T. H.; Xue, G.; Mannsfeld, S. C. B.; Bao, Z. Solution coating of large-area organic semiconductor thin films with aligned single-crystalline domains. *Nat. Mater.* **2013**, *12*, 665–671.

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

(20) 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.; van Eijck, B. P.; Verwer, P.; Williams, D. E. Crystal structure prediction of small organic molecules: a second blind test. *Acta Cryst. B* **2002**, *58*, 647–661.

(21) 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.

(22) 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.

(23) 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.

(24) 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.; Hoja, J.; Hylton, R. K.; Iuzzolino, L.; Jankiewicz, W.; de Jong, D. T.; Kendrick, J.; de Klerk, N. J.; Ko, H.-Y.; Kuleshova, L. N.; Li, X.; Lohani, S.; Leusen, F. 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.; Vàzquez-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**,

(25) Neumann, M. a.; van de Streek, J.; Fabbiani, F. P. a.; Hidber, P.; Grassmann, O. Combined crystal structure prediction and high-pressure crystallization in rational pharmaceutical polymorph screening. *Nature Communications* **2015**, *6*, 7793.

(26) Shtukenberg, A. G.; Zhu, Q.; Carter, D. J.; Vogt, L.; Hoja, J.; Schneider, E.; Song, H.; Pokroy, B.; Polishchuk, I.; Tkatchenko, A.;

Oganov, A. R.; Rohl, A. L.; Tuckermanejk, M. E.; Kahral, B. Powder diffraction and crystal structure prediction identify four new coumarin polymorphs. *Chemical Science* **2017**,

(27) Gavezzotti, A. Are crystal structures predictable? *Accounts of chemical research* **1994**, *27*, 309–314.

(28) Marom, N.; DiStasio, R. A.; Atalla, V.; Levchenko, S.; Reilly, A. M.; Chelikowsky, J. R.; Leiserowitz, L.; Tkatchenko, A. Many-Body Dispersion Interactions in Molecular Crystal Polymorphism. *Angew. Chem. Int. Ed.* **2013**, *52*, 6629–6632.

(29) Cruz-Cabeza, A. J.; Reutzel-Edens, S. M.; Bernstein, J. Facts and fictions about polymorphism. *Chemical Society Reviews* **2015**, *44*, 8619–8635.

(30) Beran, G. J. A new era for ab initio molecular crystal lattice energy prediction. *Angewandte Chemie International Edition* **2015**, *54*, 396–398.

(31) Beran, G. J. Modeling Polymorphic Molecular Crystals with Electronic Structure Theory. *Chem. Rev.* **2016**, *116*, 5567–5613.

(32) Dion, M.; Rydberg, H.; Schröder, E.; Langreth, D. C.; Lundqvist, B. I. Van der Waals density functional for general geometries. *Physical review letters* **2004**, *92*, 246401.

(33) Lee, K.; Murray, É. D.; Kong, L.; Lundqvist, B. I.; Langreth, D. C. Higher-accuracy van der Waals density functional. *Physical Review B* **2010**, *82*, 081101.

(34) Vydrov, O. A.; Van Voorhis, T. Nonlocal van der Waals density functional made simple. *Physical review letters* **2009**, *103*, 063004.

(35) Peverati, R.; Truhlar, D. G. M11-L: A local density functional that provides improved accuracy for electronic structure calculations in chemistry and physics. *The Journal of Physical Chemistry Letters* **2011**, *3*, 117–124.

(36) Peverati, R.; Truhlar, D. G. An improved and broadly accurate local approximation to the exchange–correlation density functional: The MN12-L functional for electronic structure calculations in chemistry and physics. *Physical Chemistry Chemical Physics* **2012**, *14*, 13171–13174.

(37) Zhao, Y.; Truhlar, D. G. The M06 suite of density functionals for main group thermochemistry, thermochemical kinetics, noncovalent interactions, excited states, and transition elements: two new functionals and systematic testing of four M06-class functionals and 12 other functionals. *Theoretical Chemistry Accounts: Theory, Computation, and Modeling (Theoretica Chimica Acta)* **2008**, *120*, 215–241.

(38) Vydrov, O. A.; Van Voorhis, T. Nonlocal van der Waals density functional: The simpler the better. *The Journal of chemical physics* **2010**, *133*, 244103.

(39) Vydrov, O. A.; Van Voorhis, T. Dispersion interactions from a local polarizability model. *Physical Review A* **2010**, *81*, 062708.

(40) Berland, K.; Cooper, V. R.; Lee, K.; Schröder, E.; Thonhauser, T.; Hyldgaard, P.; Lundqvist, B. I. van der Waals forces in density functional theory: a review of the vdW-DF method. *Reports on Progress in Physics* **2015**, *78*, 066501.

(41) Thonhauser, T.; Zuluaga, S.; Arter, C.; Berland, K.; Schröder, E.; Hyldgaard, P. Spin signature of nonlocal correlation binding in metal-organic frame-

works. *Physical review letters* **2015**, *115*, 136402.

(42) Peng, H.; Yang, Z.-H.; Perdew, J. P.; Sun, J. Versatile van der Waals Density Functional Based on a Meta-Generalized Gradient Approximation. *Physical Review X* **2016**, *6*, 041005.

(43) Sun, J.; Ruzsinszky, A.; Perdew, J. P. Strongly constrained and appropriately normed semilocal density functional. *Physical review letters* **2015**, *115*, 036402.

(44) Riley, K. E.; Pitonák, M.; Jurecka, P.; Hobza, P. Stabilization and structure calculations for noncovalent interactions in extended molecular systems based on wave function and density functional theories. *Chemical Reviews* **2010**, *110*, 5023–5063.

(45) Grimme, S. Semiempirical GGA-type density functional constructed with a long-range dispersion correction. *Journal of computational chemistry* **2006**, *27*, 1787–1799.

(46) Grimme, S.; Antony, J.; Ehrlich, S.; Krieg, H. A consistent and accurate ab initio parametrization of density functional dispersion correction (DFT-D) for the 94 elements H-Pu. *The Journal of chemical physics* **2010**, *132*, 154104.

(47) Johnson, E. R.; Becke, A. D. A post-Hartree–Fock model of intermolecular interactions. *The Journal of chemical physics* **2005**, *123*, 024101.

(48) Otero-De-La-Roza, A.; Johnson, E. R. A benchmark for non-covalent interactions in solids. *The Journal of chemical physics* **2012**, *137*, 054103.

(49) Jurečka, P.; Černỳ, J.; Hobza, P.; Salahub, D. R. Density functional theory augmented with an empirical dispersion term. Interaction energies and geometries of 80 noncovalent complexes compared with ab initio quantum mechanics calculations. *Journal of computational chemistry* **2007**, *28*, 555–569.

(50) Wu, Q.; Yang, W. Empirical correction to density functional theory for van der Waals interactions. *The Journal of chemical physics* **2002**, *116*, 515–524.

(51) Wu, X.; Vargas, M.; Nayak, S.; Lotrich, V.; Scoles, G. Towards extending the applicability of density functional theory to weakly bound systems. *The Journal of Chemical Physics* **2001**, *115*, 8748–8757.

(52) Steinmann, S. N.; Corminboeuf, C. A generalized-gradient approximation exchange hole model for dispersion coefficients. *The Journal of chemical physics* **2011**, *134*, 044117.

(53) Steinmann, S. N.; Corminboeuf, C. Comprehensive benchmarking of a density-dependent dispersion correction. *Journal of chemical theory and computation* **2011**, *7*, 3567–3577.

(54) Tkatchenko, A.; Scheffler, M. Accurate Molecular Van Der Waals Interactions from Ground-State Electron Density and Free-Atom Reference Data. *Phys. Rev. Lett.* **2009**, *102*, 073005.

(55) Brandenburg, J.; Bates, J.; Sun, J.; Perdew, J. Benchmark tests of a strongly constrained semilocal functional with a long-range dispersion correction. *Physical Review B* **2016**, *94*, 115144.

(56) DiStasio, R. A.; von Lilienfeld, O. A.; Tkatchenko, A. Collective many-body van der Waals interactions in molecular systems. *Proceedings of the National Academy of Sciences* **2012**, *109*, 14791–14795.

(57) Tkatchenko, A.; DiStasio, R. A.; Car, R.; Scheffler, M. Accurate and Efficient Method for Many-Body van der Waals Interactions. *Phys. Rev. Lett.* **2012**, *108*, 236402.

(58) Ambrosetti, A.; Reilly, A. M.; DiStasio, R. A.; Tkatchenko, A. Long-range correlation energy calculated from coupled atomic response functions. *J. Chem. Phys.* **2014**, *140*, 18A508.

(59) Schatschneider, B.; Liang, J.-J.; Reilly, A. M.; Marom, N.; Zhang, G.-X.; Tkatchenko, A. Electrodynamic response and stability of molecular crystals. *Physical Review B* **2013**, *87*, 060104.

(60) Reilly, A. M.; Tkatchenko, A. Understanding the role of vibrations, exact exchange, and many-body van der Waals interactions in the cohesive properties of molecular crystals. *J. Chem. Phys.* **2013**, *139*, 024705.

(61) Reilly, A. M.; Tkatchenko, A. Seamless and Accurate Modeling of Organic Molecular Materials. *J. Phys. Chem. Lett.* **2013**, *4*, 1028–1033.

(62) Tkatchenko, A. Current understanding of van der Waals effects in realistic materials. *Advanced Functional Materials* **2015**, *25*, 2054–2061.

(63) Hermann, J.; DiStasio Jr, R. A.; Tkatchenko, A. First-Principles Models for van der Waals Interactions in Molecules and Materials: Concepts, Theory, and Applications. **2017**,

(64) Flores-Huerta, A. G.; Tkatchenko, A.; Galván, M. Nature of Hydrogen Bonds and S··· S Interactions in the l-Cystine Crystal. *The Journal of Physical Chemistry A* **2016**, *120*, 4223–4230.

(65) Hoja, J.; Reilly, A. M.; Tkatchenko, A. First-principles modeling of molecular crystals: structures and stabilities, temperature and pressure. *Wiley Interdisciplinary Reviews: Computational Molecular Science* **2017**, *7*.

(66) Rossi, M.; Gasparotto, P.; Ceriotti, M. Anharmonic and quantum fluctuations in molecular crystals: a first-principles study of the stability of paracetamol. *Physical Review Letters* **2016**, *117*, 115702.

(67) Nyman, J.; Day, G. M. Static and lattice vibrational energy differences between polymorphs. *CrystEngComm* **2015**, *17*, 5154–5165.

(68) Yu, T.-Q.; Tuckerman, M. E. Temperature-accelerated method for exploring polymorphism in molecular crystals based on free energy. *Physical review letters* **2011**, *107*, 015701.

(69) Schneider, E.; Vogt, L.; Tuckerman, M. E. Exploring polymorphism of benzene and naphthalene with free energy based enhanced molecular dynamics. *Acta Crystallographica Section B: Structural Science, Crystal Engineering and Materials* **2016**, *72*, 542–550.

(70) Akkermans, R. L.; Spenley, N. A.; Robertson, S. H. Monte Carlo methods in materials studio. *Molecular Simulation* **2013**, *39*, 1153–1164.

(71) Wang, Y.; Lv, J.; Zhu, L.; Ma, Y. CALYPSO: A method for crystal structure prediction. *Computer Physics Communications* **2012**, *183*, 2063–2070.

(72) Pickard, C. J.; Needs, R. Ab initio random structure searching. *Journal of Physics: Condensed Matter* **2011**, *23*, 053201.

(73) Case, D. H.; Campbell, J. E.; Bygrave, P. J.; Day, G. M. Convergence properties of crystal structure prediction by quasi-random sampling. *Journal of chemical theory and computation* **2016**, *12*, 910–924.

(74) Johnston, R. L. Evolving better nanoparticles: Genetic algorithms for optimising cluster geometries. *Dalton Transactions* **2003**, 4193–4207.

(75) Sierka, M. Synergy between theory and experiment in structure resolution of low-dimensional oxides. *Progress in Surface Science* **2010**, *85*, 398–434.

(76) Heiles, S.; Johnston, R. L. Global optimization of clusters using electronic structure methods. *International Journal of Quantum Chemistry* **2013**, *113*, 2091–2109.

(77) Oganov, A. R.; Glass, C. W. Crystal structure prediction using ab initio evolutionary techniques: Principles and applications. *The Journal of chemical physics* **2006**, *124*, 244704.

(78) Glass, C. W.; Oganov, A. R.; Hansen, N. USPEX—evolutionary crystal structure prediction. *Computer Physics Communications* **2006**, *175*, 713–720.

(79) Abraham, N. L.; Probert, M. I. J. A periodic genetic algorithm with real-space representation for crystal structure and polymorph prediction. *Phys. Rev. B* **2006**, *73*, 224104.

(80) Trimarchi, G.; Zunger, A. Global space-group optimization problem: Finding the stablest crystal structure without constraints. *Physical Review B* **2007**, *75*, 104113.

(81) Wu, S.; Umemoto, K.; Ji, M.; Wang, C.-Z.; Ho, K.-M.; Wentzcovitch, R. M. Identification of post-pyrite phase transitions in SiO 2 by a genetic algorithm. *Physical Review B* **2011**, *83*, 184102.

(82) Woodley, S.; Battle, P.; Gale, J.; Catlow, C. A. The prediction of inorganic crystal structures using a genetic algorithm and energy minimisation. *Physical Chemistry Chemical Physics* **1999**, *1*, 2535–2542.

(83) Trimarchi, G.; Zunger, A. Finding the lowest-energy crystal structure starting from randomly selected lattice vectors and atomic positions: first-principles evolutionary study of the Au–Pd, Cd–Pt, Al–Sc, Cu–Pd, Pd–Ti, and Ir–N binary systems. *Journal of Physics: Condensed Matter* **2008**, *20*, 295212.

(84) Lonie, D. C.; Zurek, E. XtalOpt: An open-source evolutionary algorithm for crystal structure prediction. *Computer Physics Communications* **2011**, *182*, 372–387.

(85) Jóhannesson, G. H.; Bligaard, T.; Ruban, A. V.; Skriver, H. L.; Jacobsen, K. W.; Nørskov, J. K. Combined electronic structure and evolutionary search approach to materials design. *Physical Review Letters* **2002**, *88*, 255506.

(86) Zhu, Q.; Oganov, A. R.; Glass, C. W.; Stokes, H. T. Constrained evolutionary algorithm for structure prediction of molecular crystals: methodology and applications. *Acta Crystallogr B* **2012**, *68*, 215–26.

(87) Lund, A. M.; Pagola, G. I.; Orendt, A. M.; Ferraro, M. B.; Facelli, J. C. Crystal structure prediction from first principles: The crystal structures of glycine. *Chemical physics letters* **2015**, *626*, 20–24.

(88) Avery, P.; Falls, Z.; Zurek, E. XtalOpt Version r10: An open–source evolutionary algorithm for crystal structure prediction. *Computer Physics Communications* **2017**, *217*, 210 – 211.

(89) Falls, Z.; Lonie, D. C.; Avery, P.; Shamp, A.; Zurek, E. XtalOpt version r9: An open-source evolutionary algorithm for crystal structure prediction. *Computer Physics Communications* **2016**, *199*, 178 – 179.

(90) Morris, J.; Deaven, D.; Ho, K.; Wang, C.; Pan, B.; Wacker, J.; Turner, D. Genetic algorithm optimization of atomic clusters. Evolutionary Algorithms. 1999; pp 167–175.

(91) Alexandrova, A. N.; Boldyrev, A. I. Search for the Li n 0/+ 1/-1 (n= 5-7) Lowest-Energy Structures Using the ab Initio Gradient Embedded Genetic Algorithm (GEGA). Elucidation of the Chemical Bonding in the Lithium Clusters. *Journal of chemical theory and computation* **2005**, *1*, 566–580.

(92) Marques, J. M. C.; Pereira, F. B. An evolutionary algorithm for global minimum search of binary atomic clusters. *Chemical Physics Letters* **2010**, *485*, 211–216.

(93) Hartke, B. Global cluster geometry optimization by a phenotype algorithm with Niches: Location of elusive minima, and low-order scaling with cluster size. *Journal of computational chemistry* **1999**, *20*, 1752–1759.

(94) Catlow, C. R. A.; Bromley, S. T.; Hamad, S.; Mora-Fonz, M.; Sokol, A. A.; Woodley, S. M. Modelling nano-clusters and nucleation. *Physical Chemistry Chemical Physics* **2010**, *12*, 786–811.

(95) Bazterra, V. E.; Oña, O.; Caputo, M. C.; Ferraro, M. B.; Fuentealba, P.; Facelli, J. C. Modified genetic algorithms to model cluster structures in medium-size silicon clusters. *Physical Review A* **2004**, *69*, 053202.

(96) Bhattacharya, S.; Levchenko, S. V.; Ghiringhelli, L. M.; Scheffler, M. Stability and Metastability of Clusters in a Reactive Atmosphere: Theoretical Evidence for Unexpected Stoichiometries of Mg M O x. *Physical review letters* **2013**, *111*, 135501.

(97) Bhattacharya, S.; Levchenko, S. V.; Ghiringhelli, L. M.; Scheffler, M. Efficient ab initio schemes for finding thermodynamically stable and metastable atomic structures: Benchmark of cascade genetic algorithms. *New Journal of Physics* **2014**, *16*, 123016.

(98) Jørgensen, M. S.; Groves, M. N.; Hammer, B. Combining evolutionary algorithms with clustering toward rational global structure optimization at the atomic scale. *Journal of Chemical Theory and Computation* **2017**, *13*, 1486–1493.

(99) O'Boyle, N. M.; Campbell, C. M.; Hutchison, G. R. Computational design and selection of optimal organic photovoltaic materials. *The Journal of Physical Chemistry C* **2011**, *115*, 16200–16210.

(100) Jain, A.; Castelli, I. E.; Hautier, G.; Bailey, D. H.; Jacobsen, K. W. Performance of genetic algorithms in search for water splitting perovskites. *Journal of Materials Science* **2013**, *48*, 6519–6534.

(101) d'Avezac, M.; Luo, J.-W.; Chanier, T.; Zunger, A. Genetic-algorithm discovery of a direct-gap and optically allowed superstructure from indirect-gap Si and Ge semiconductors. *Physical review letters* **2012**, *108*, 027401.

(102) Zhang, L.; Luo, J.-W.; Saraiva, A.; Koiller, B.; Zunger, A. Genetic design of enhanced valley splitting towards a spin qubit in silicon. *Nature communications* **2013**, *4*.

(103) Chua, A. L.-S.; Benedek, N. A.; Chen, L.; Finnis, M. W.; Sutton, A. P. A genetic algorithm for predicting the structures of interfaces in multicomponent systems. *Nature materials* **2010**, *9*, 418–422.

(104) Bhattacharya, S.; Sonin, B. H.; Jumonville, C. J.; Ghiringhelli, L. M.; Marom, N. Computational design of nanoclusters by property-based genetic algorithms: Tuning the electronic properties of (TiO2) n clusters. *Physical Review B* **2015**, *91*, 241115.

(105) Li, X.; Curtis, F.; Rose, T.; Schober, C.; Vázquez-Mayagoitia, A.; Marom, N.

Genarris: Random Generation of Molecular Crystal Structures and Fast Screening with a Harris Approximation. *To be published*

(106) 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. *Computer Physics Communications* **2009**, *180*, 2175–2196.

(107) Havu, V.; Blum, V.; Havu, P.; Scheffler, M. Efficient O (N) integration for all-electron electronic structure calculation using numeric basis functions. *Journal of Computational Physics* **2009**, *228*, 8367–8379.

(108) Perdew, J. P.; Burke, K.; Ernzerhof, M. Generalized Gradient Approximation Made Simple. *Phys. Rev. Lett.* **1996**, *77*, 3865–3868.

(109) Perdew, J. P.; Burke, K.; Ernzerhof, M. Generalized Gradient Approximation Made Simple [Phys. Rev. Lett. 77, 3865 (1996)]. *Phys. Rev. Lett.* **1997**, *78*, 1396–1396.

(110) Perdew, J. P.; Ernzerhof, M.; Burke, K. Rationale for mixing exact exchange with density functional approximations. *J. Chem. Phys.* **1996**, *105*, 9982–9985.

(111) Adamo, C.; Barone, V. Toward reliable density functional methods without adjustable parameters: The PBE0 model. *J. Chem. Phys.* **1999**, *110*, 6158–6170.

(112) Togo, A. `https://atztogo.github.io/spglib/`.

(113) Pedregosa, F.; Varoquaux, G.; Gramfort, A.; Michel, V.; Thirion, B.; Grisel, O.; Blondel, M.; Prettenhofer, P.; Weiss, R.; Dubourg, V.; Vanderplas, J.; Passos, A.; Cournapeau, D.; Brucher, M.; Perrot, M.; Duchesnay, E. Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research* **2011**, *12*, 2825–2830.

(114) Ong, S. P.; Richards, W. D.; Jain, A.; Hautier, G.; Kocher, M.; Cholia, S.; Gunter, D.; Chevrier, V. L.; Persson, K. A.; Ceder, G. Python Materials Genomics (pymatgen): A robust, open-source python library for materials analysis. *Computational Materials Science* **2013**, *68*, 314–319.

(115) Holland, J. H. Adaptation in natural and artificial systems. An introductory analysis with application to biology, control, and artificial intelligence. *Ann Arbor, MI: University of Michigan Press* **1975**,

(116) others,, et al. Genetic algorithms with sharing for multimodal function optimization. Genetic algorithms and their applications: Proceedings of the Second International Conference on Genetic Algorithms. 1987; pp 41–49.

(117) Sareni, B.; Krahenbuhl, L. Fitness sharing and niching methods revisited. *IEEE Transactions on Evolutionary computation* **1998**, *2*, 97–106.

(118) Shir, O. M. *Handbook of Natural Computing*; Springer, 2012; pp 1035–1069.

(119) Preuss, M. *Multimodal optimization by means of evolutionary algorithms*; Springer, 2015.

(120) Niggli, P. *Krystallographische und strukturtheoretische Grundbegriffe*; Akademische verlagsgesellschaft mbh, 1928; Vol. 1.

(121) Gruber, B. The relationship between reduced cells in a general Bravais lattice. *Acta Crystallographica Section A: Crystal Physics, Diffraction, Theoretical and General Crystallography* **1973**, *29*, 433–440.

(122) Křivỳ, I.; Gruber, B. A unified algorithm for determining the reduced (Niggli) cell. *Acta Crystallographica Section A: Crystal Physics, Diffraction, Theoretical and General Crystallography* **1976**, *32*, 297–298.

(123) Grosse-Kunstleve, R. W.; Sauter, N. K.; Adams, P. D. Numerically stable algorithms for the computation of reduced unit cells. *Acta Crystallographica Section A: Foundations of Crystallography* **2004**, *60*, 1–6.

(124) Kanungo, T.; Mount, D. M.; Netanyahu, N. S.; Piatko, C. D.; Silverman, R.; Wu, A. Y. An efficient k-means clustering algorithm: Analysis and implementation. *IEEE transactions on pattern analysis and machine intelligence* **2002**, *24*, 881–892.

(125) Frey, B. J.; Dueck, D. Clustering by passing messages between data points. *science* **2007**, *315*, 972–976.

(126) Lyakhov, A. O.; Oganov, A. R.; Valle, M. How to predict very large and complex crystal structures. *Computer Physics Communications* **2010**, *181*, 1623–1632.

(127) Lyakhov, A. O.; Oganov, A. R.; Stokes, H. T.; Zhu, Q. New developments in evolutionary structure prediction algorithm USPEX. *Computer Physics Communications* **2013**, *184*, 1172–1182.

(128) Goldberg, D. E. *Genetic Algorithms in Search, Optimization and Machine Learning*, 1st ed.; Addison-Wesley Longman Publishing Co., Inc.: Boston, MA, USA, 1989.

(129) Blickle, T.; Thiele, L. A comparison of selection schemes used in evolutionary algorithms. *Evolutionary Computation* **1996**, *4*, 361–394.

(130) Deaven, D. M.; Ho, K.-M. Molecular geometry optimization with a genetic algorithm. *Physical review letters* **1995**, *75*, 288.

(131) Froltsov, V. A.; Reuter, K. Robustness of cut and splice genetic algorithms in the structural optimization of atomic clusters. *Chemical Physics Letters* **2009**, *473*, 363 – 366.

(132) Ji, M.; Wang, C.-Z.; Ho, K.-M. Comparing efficiencies of genetic and minima hopping algorithms for crystal structure prediction. *Physical Chemistry Chemical Physics* **2010**, *12*, 11617–11623.

(133) Wang, Y.; Lv, J.; Zhu, L.; Ma, Y. Crystal structure prediction via particle-swarm optimization. *Phys. Rev. B* **2010**, *82*, 094116.

(134) Simmons, H. E.; Vest, R. D.; Blomstrom, D. C.; Roland, J. R.; Cairns, T. L. Thiacyanocarbons. I. Tetracyano-1,4-dithiin, Tetracyanothiophene and Tricyano-1,4-dithiino [c]isothiazole. *J. Am. Chem. Soc.* **1962**, *84*, 4746–4756.

(135) Blake, A. J.; Clark, B. A.; Gierens, H.; Gould, R. O.; Hunter, G. A.; McNab, H.; Morrow, M.; Sommerville, C. C. Intramolecular and intermolecular geometry of thiophenes with oxygen-containing substituents. *Acta Crystallographica Section B: Structural Science* **1999**, *55*, 963–974.

(136) Braun, D. E.; Tocher, D. A.; Price, S. L.; Griesser, U. J. The complexity of hydration of phloroglucinol: a comprehensive structural and thermodynamic characterization. *The Journal of Physical Chemistry B* **2012**, *116*, 3961–3972.

(137) Braun, D. E.; Nartowski, K. P.; Khimyak, Y. Z.; Morris, K. R.; Byrn, S. R.; Griesser, U. J. Structural Properties, Order–Disorder Phenomena, and Phase Stability of Orotic Acid Crystal Forms. *Molecular pharmaceutics* **2016**, *13*, 1012–1029.

(138) Whittleton, S. R.; Otero-de-la Roza, A.; Johnson, E. R. Exchange-Hole Dipole Dispersion Model for Accurate Energy Ranking in Molecular Crystal Structure Prediction. *Journal of chemical theory and computation* **2017**, *13*, 441–450.

(139) Chan, H. S.; Kendrick, J.; Leusen, F. J. Predictability of the polymorphs of small

organic compounds: Crystal structure predictions of four benchmark blind test molecules. *Physical Chemistry Chemical Physics* **2011**, *13*, 20361–20370.

(140) Neumann, M. A. Tailor-made force fields for crystal-structure prediction. *The Journal of Physical Chemistry B* **2008**, *112*, 9810–9829.

(141) Grimme, S. Density functional theory with London dispersion corrections. *Wiley Interdisciplinary Reviews: Computational Molecular Science* **2011**, *1*, 211–228.

(142) Becke, A. On the large-gradient behavior of the density functional exchange energy. *The Journal of chemical physics* **1986**, *85*, 7184–7187.

(143) Becke, A. D.; Johnson, E. R. Exchange-hole dipole moment and the dispersion interaction revisited. *The Journal of chemical physics* **2007**, *127*, 154108.

(144) Otero-de-la Roza, A.; Johnson, E. R. Van der Waals interactions in solids using the exchange-hole dipole moment model. *The Journal of chemical physics* **2012**, *136*, 174109.

(145) Steed, K. M.; Steed, J. W. Packing problems: high Z crystal structures and their relationship to cocrystals, inclusion compounds, and polymorphism. *Chemical reviews* **2015**, *115*, 2895–2933.

(146) Vande Velde, C. M.; Tylleman, B.; Zeller, M.; Sergeyev, S. Structures of alkyl-substituted Tröger's base derivatives illustrate the importance of Z′ for packing in the absence of strong crystal synthons. *Acta Crystallographica Section B: Structural Science* **2010**, *66*, 472–481.

(147) Taylor, R.; Cole, J. C.; Groom, C. R. Molecular Interactions in Crystal Structures with $Z' > 1$. *Crystal Growth & Design* **2016**, *16*, 2988–3001.

(148) Riley, K. E.; Hobza, P. Investigations into the Nature of Halogen Bonding Including Symmetry Adapted Perturbation Theory Analyses. *J Chem Theory Comput* **2008**, *4*, 232–42.

(149) Cavallo, G.; Metrangolo, P.; Milani, R.; Pilati, T.; Priimagi, A.; Resnati, G.; Terraneo, G. The halogen bond. *Chem. Rev* **2016**, *116*, 2478–2601.

(150) Kozuch, S.; Martin, J. M. L. Halogen Bonds: Benchmarks and Theoretical Analysis. *J Chem Theory Comput* **2013**, *9*, 1918–31.

(151) Řezáč, J.; Riley, K. E.; Hobza, P. Benchmark Calculations of Noncovalent Interactions of Halogenated Molecules. *J Chem Theory Comput* **2012**, *8*, 4285–92.

(152) Otero-de-la Roza, A.; Johnson, E. R.; DiLabio, G. A. Halogen Bonding from Dispersion-Corrected Density-Functional Theory: The Role of Delocalization Error. *J Chem Theory Comput* **2014**, *10*, 5436–47.

(153) Misquitta, A. J.; Welch, G. W.; Stone, A. J.; Price, S. L. A first principles prediction of the crystal structure of C6Br2ClFH2. *Chemical Physics Letters* **2008**, *456*, 105–109.

(154) Mooij, W. T.; Leusen, F. J. Multipoles versus charges in the 1999 crystal structure prediction test. *Physical Chemistry Chemical Physics* **2001**, *3*, 5063–5066.

(155) Day, G. M.; Chisholm, J.; Shan, N.; Motherwell, W. S.; Jones, W. An assessment of lattice energy minimization for the prediction of molecular organic crystal structures. *Crystal growth & design* **2004**, *4*, 1327–1340.

(156) Asmadi, A.; Neumann, M. A.; Kendrick, J.; Girard, P.; Perrin, M.-A.; Leusen, F. J. Revisiting the blind tests in crystal structure prediction: accurate energy ranking of molecular crystals.

The Journal of Physical Chemistry B **2009**, *113*, 16303–16313.

(157) Wales, D. J.; Salamon, P. Observation time scale, free-energy landscapes, and molecular symmetry. *Proceedings of the National Academy of Sciences* **2014**, *111*, 617–622.

(158) Carr, J. M.; Mazauric, D.; Cazals, F.; Wales, D. J. Energy landscapes and persistent minima. *The Journal of chemical physics* **2016**, *144*, 054109.

(159) Macrae, C. F.; Bruno, I. J.; Chisholm, J. A.; Edgington, P. R.; McCabe, P.; Pidcock, E.; Rodriguez-Monge, L.; Taylor, R.; van de Streek, J.; Wood, P. A. *Mercury CSD 2.0* – new features for the visualization and investigation of crystal structures. *Journal of Applied Crystallography* **2008**, *41*, 466–470.

Chapter 4

Summary and Outlook

Molecular crystal structure prediction (CSP) is a coveted computational tool because it can predict the existence of previously unobserved polymorphs and serve as an important resource for experimental studies of organic solid forms. In this thesis, I have developed and applied a first principles genetic algorithm called GAtor, which performs molecular crystal structure prediction for small (semi-)rigid molecules. Accurately computing the relative stabilities of potential polymorphs is particularly challenging. Therefore, the best structures produced from the GA are ranked in energy using state-of-the-art dispersion inclusive DFT methods, including PBE+TS, PBE+MBD, and PBE0+MBD.

Section 3.1 presents the main publication from the sixth CCDC blind test. Blind tests are held periodically to showcase the advances and remaining challenges of cutting edge CSP methods. In these tests participating researchers have a year to submit putative crystal structures of molecules with unpublished crystal structures solely from the molecule's 2D chemical diagram. Our group participated in the sixth blind test, attempting structure prediction of a unique molecule, tricyano-1,4-dithiino[c]-isothiazole (Target XXII), which contains carbon, nitrogen, sulfur and a plurality of cyano groups and has partial hinge flexibility in its six-membered ring. The experimental structure was not generated at the time of submission mainly due to the constraints imposed on the unit cell angles within the preliminary version of the genetic algorithm. However, several other important low-energy structures were generated, including a structure that was very similar to the experimental in terms of its crystal packing motif and practically degenerate in energy when computed using PBE+TS and PBE+MBD.

Section 3.2 presents an additional analysis of the structures generated within the sixth blind test along with the experimental form. The top 100 low-energy structures generated were categorized into four main packing motifs, including cyclic dimer, catemer, bilayer, and planar. The PBE+TS, PBE+MBD, and PBE0+MBD energies were computed for these structures, and it was found that certain dispersion-inclusive DFT methods systematically favored particular packing motifs. Structures with catemer-like chain motifs were destabilized with respect to cyclic dimer structures by the TS pairwise dispersion method and stabilized by the inclusion of many-body dispersion interactions in the MBD method. Structures with layered motifs were overstabilized by the semi-local PBE functional, compared with the hybrid PBE0 functional, possibly due to the self- interaction error. Only with PBE0+MBD is the experimentally observed $P2_1/n$ cyclic dimer structure found to be the most stable structure. Several electronic and optical properties of a computed low energy structure with a bilayer packing motif, within 1.9 kJ/mol of the experimental structure, were compared to those of the experimental structure with a cyclic dimer motif. Namely, the bilayer structure showed a narrower band gap, enhanced band dispersion, and a broader optical absorption. This demonstrates how the the crystal packing of an organic semiconductor can be significantly modified by only changing the crystal packing, important for applications in organic electronics.

Section 3.3 presents the methodology and application of a general purpose molecular crystal generation package called Genarris. In this thesis, Genarris is used to prepare the initial pool for the GAtor genetic algorithm. However, it may also be used for other purposes such as creating molecular crystal training sets for machine learning applications with a modest computational cost. In Genarris, several thousand structures (e.g. 5,000) are generated randomly for a given molecule within the 230 crystallographic space groups. The energies of the raw pool are evaluated using a Harris approximation, in which the Harris density of a molecular crystal is constructed by a superposition of single molecule densities. The single molecule density is converged self-consistently and only needs to be computed once. The dispersion-inclusive DFT total energy is evaluated for the Harris density without performing a self-consistent cycle, allowing for fast energy evaluations of the large pool of structures. Different workflows are created, e.g. energy, diverse, and rigorous, which determine the final pool of structures. The diverse workflow is used for creating the initial pools for GAtor, in which affinity propagation (AP) clustering is used wth a relative coordi-

nate descriptor that captures the packing motif for a given molecular crystal. The lowest-energy structures are selected from each cluster of structures in order to generate a smaller subset of the raw pool (e.g. 500 structures). Then, the smaller set of structures is clustered again and only the exemplars (the centers of each cluster) are chosen for the diverse initial pool (e.g. 50 structures). The energy workflow employs fully-self consistent DFT calculations after the initial clustering stage, and then selects the top 50 structures of lowest energy for targeted sampling of low-energy structures. The rigorous approach additionally incorporates local relaxation into its workflow, and is a CSP algorithm in and of itself, generating the experimental structures of 5-cyano-3-hydroxythiophene (Target II), 1,3-dibromo-2-chloro-5-fluorobenzene (Target XIII), and tricyano-1,4-dithiino[c]-isothiazole (Target XXII). The diverse workflow is shown to perform best in the GAtor genetic algorithm, as compared to the energy workflow and a randomly selected pool of structures.

Section 3.4 presents the up-to-date methodology and applications of the GAtor genetic algorithm, the main subject of this thesis. GAtor is optimized for high performing computing (HPC) environments by having a workflow that runs several GA replicas in parallel which read and write to a common pool of structures. GAtor has been successfully test on up to 262,144 cores at the Argonne Leadership Facility (ALCF). The code offers a variety of features that enable the user to customize the GA search settings, including energy-based and cluster-based fitness (evolutionary niching), roulette wheel and tournament selection, symmetric and standard crossover, different mutation schemes, and various tunable parameters related to energy cutoffs, similarity checks, and geometric constraints. The crossover and mutation operators are specifically tailored for molecular crystals and provide a balance between exploration and exploitation of the potential energy surface. They also have the ability to generate high Z' structures. Specifically, symmetric crossover is a novel operator that ensures the space group symmetries of one parent structure are inherited in the produced child structure. Evolutionary niching via cluster-based fitness aids in evenly sampling the potential energy surface by learning from the accumulated data and suppressing the over-sampling of densely populated regions. To validate the algorithm, GAtor was used to perform structure prediction for a chemically diverse set of four past blind test targets, 4-cyclobutylfuran (Target I), 5-cyano-3-hydroxythiophene (Target II), 1,3-dibromo-2-chloro-5-fluorobenzene (Target XIII), and tricyano-1,4-dithiino[c]-isothiazole (Target XXII). The experimental structure(s)

as well as several other important low-energy structures were generated for all four targets. For Target XXII, the cluster-based fitness function was employed with a simple lattice-parameter descriptor that uniformly explored the potential energy surface, including regions outside the initial pool, and suppressed the oversampling of structures with planar molecular conformations. For Target II, the structure consistently ranked as #1 with PBE+TS, PBE+MBD, and PBE0+MBD was predicted for the first time using GAtor. It is a Z'=2 structure with $P\bar{1}$ symmetry and a scaffold packing motif, whose lattice energy is 1.8 kJ/mol per molecule lower than the known layered experimental form. This structure, as well as several other low-lying putative structures of Target II have higher densities than the observed experimental structure, therefore it may be possible to crystallize them under high pressure conditions. This may motivate further experimental investigations of Target II. Further computational studies considering finite temperature and pressure effects may provide additional insight into the relative stability of the putative low-energy structures identified and the possibility of growing them experimentally.

GAtor is still a relatively new code, with many extensions and applications planned for the future that will be briefly discussed. The cluster-based fitness scheme shows much promise for overcoming classic hurdles of genetic algorithms (e.g. getting stuck local minima) and is particularly relevant for polymorph prediction where one is not interested in only generating the experimental structure or the lowest energy structure. Rather one wants to locate all low-energy structures, including those in narrow or hard-to-reach wells in the potential energy surface. To this end, tests are currently being carried out that further analyze how the chosen descriptor and cluster-based fitness function affect the GA search. Target XIII is being used as a test case because its experimental structure exhibits a packing motif quite unlike many of those found in the energy-based GA searches, as described in Section 3.4. Additionally, GAtor will soon be extended for the structure prediction of co-crystals, which contain more than one type of molecule in the asymmetric unit. Much of the machinery of Genarris and GAtor will remain the same, but additional measures need to be taken to vary the relative orientation of the different molecules within the asymmetric unit, especially in the initial pool generation. In the future, GAtor will also be extended to have functionality for structure prediction of flexible-molecules. As this significantly increases the complexity of CSP, approaches for generating tailor-made force fields may need to be developed for configuration space screening purposes.

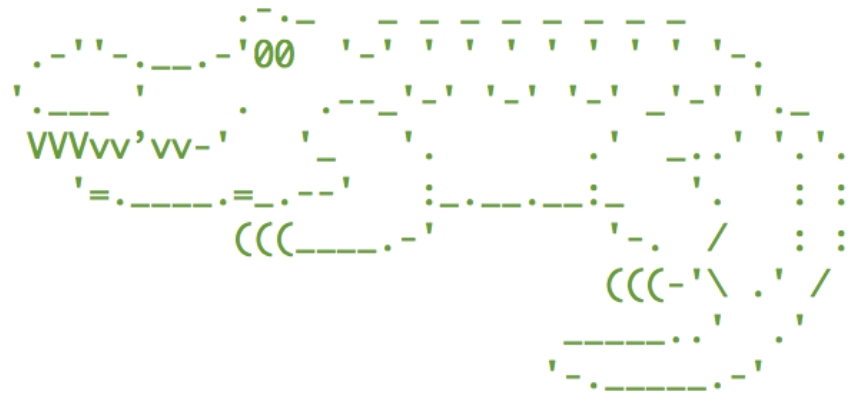
GAtor will also be extended to perform optimization of molecular interfaces, for applications in organic electronics. This will take significant effort as new schemes need to be developed for generating the initial population and new breeding operators need to be designed for these systems. Finally, GAtor will soon be extended to perform property-based optimization, as opposed to energy-based optimization, for the purposes of molecular crystal engineering for organic electronics applications. For example, descriptors will be determined which correlate a property of interest (e.g. high charge carrier mobility) with the electronic structure of a given molecular crystal. These descriptors will be incorporated into a property-based fitness function in order to guide the GA to generate structures that optimize the property of interest. Overall, GAtor is expected to be a useful and flexible tool for the condensed matter and material's science communities.

Appendix A

Appendix

A.1 GAtor Genetic Algorithm User Manual

GAtor Genetic Algorithm for Molecular Crystal Structure Prediction



A User's Manual

Farren Curtis

Department of Physics and
Department of Materials Science and Engineering
Carnegie Mellon University, Pittsburgh, PA 15213, USA

December 2017

Contents

1 Basic Installation and Tutorial **1**

1.1 Introduction 1

1.2 Installation Requirements for GAtor 1

1.3 Structure of the Code 2

1.4 Basic Tutorial 2

1.4.1 The ui.conf file 2

1.4.2 Basic ui.conf file settings 2

1.4.3 Filling the initial pool 6

1.4.4 Running the GA 7

1.4.5 Individual and Combined Replica Outputs 7

1.4.6 GAtor time log 7

1.4.7 Temp Directories for FHI-aims evaluations 8

1.4.8 Structures directory 8

1.4.9 Energy Hierarchy 8

1.4.10 Duplicates 9

2 Full Configuration File Parameters **10**

Chapter 1

Basic Installation and Tutorial

1.1 Introduction

Welcome to the GAtor genetic algorithm for molecular crystal structure prediction, which finds the most energetically stable crystal structures for (semi-)rigid molecules. GAtor uses principles from evolutionary theory such as survival of the fittest, crossover, and mutation that are implemented as operators acting on individual molecules and/or lattice vectors of the fittest crystal structures selected for mating. Energy evaluations and structural relaxations are performed using dispersion-inclusive density functional theory (DFT). For this purpose, GAtor currently interfaces with the all-electron numerical atom-centered orbital DFT code FHI-aims.

1.2 Installation Requirements for GAtor

GAtor is written is python and interfaces with the all-electron electronic structure theory code FHI-aims. GAtor can be downloaded from http://software.noamarom.com. To run GAtor the user will need to install:

- Python 2.7 (http://www.python.org).
- NumPy version $\geq$ 1.9 (http://www.numpy.org).
- Pymatgen version $\geq$ 4.4.0 (http://www.pymatgen.org).
- Sci-kit learn version $\geq$ 0.17.1 (http://scikit-learn.org).
- A build of FHI-aims (MPI and scalapack-supported, if possible).

Check that python and the dependent packages can be successfully imported:

```
$ python
» import numpy
» import sklearn
» import pymatgen
```

1.3 Structure of the Code

GAtor contains the main directories `/src`, `/tests` and `/tutorial`. The `/src` folder contains the master script `GAtor_master.py` along with core modules of the GA such as `/selection` and `/crossover`. The `/tutorial` folder provides an example GA run (see Section 1.4). The `/tests` folder contains different tests one should run when running a new molecule for the first time.

To run GAtor, one invokes the main script and inputs a user-defined configuration file `ui.conf` via `$ python gator/GAtor_master.py ui.conf`. For an explanation of the `ui.conf` file see Section 1.4.1.

1.4 Basic Tutorial

This section takes the user through the example calculation provided in `/tutorial`. This folder contains a `ui.conf` file, a sample initial pool, and sample FHI-aims control files `control.in.SPE.tier_1` and `control.in.FULL.tier_1`. The first control file is used just for single point energy evaluations, while the second is used for local optimization[1]. The energy cutoffs corresponding to each control file are detailed in Section 1.4.2.

1.4.1 The ui.conf file

The `ui.conf` is the only file the user needs to modify in order to use GAtor. It can be named anything as long as it ends with ".conf". A simple example for the molecule 3-4-cyclobutylfuran can be found in `/gator/tutorial/ui.conf`. The conf file contains the parameters that control all aspects of GAtor including parallelization options, paths to the user-input initial pool, options for interfacing with FHI-aims, and tuning parameters for GA tasks such as mutation probability and duplicate-check tolerances. For an explanation of the simple keywords shown in `/gator/example_calc/ui.conf` see Section 1.4.5. For a full catalog of all the possible keywords that can go into this file, see Section 2.

1.4.2 Basic ui.conf file settings

This section details the parameters you will see in the basic configuration file `ui.conf`. All parameters are grouped into main sections. This conf file runs GAtor for the molecule 3-4-cyclobutylfuran.

[1]To accelerate the DFT calculations for tutorial purposes, limitations are placed on the number self-consistent iterations and max relaxation steps in the control files. These control files should not be used for production purposes, as the calculations will not be fully converged.

`[GAtor_master]`

The GAtor_master section controls the main procedures of initial pool filling and running the GA.

- `fill_initial_pool = TRUE`
  - Fills the user-defined initial pool into the common pool of structures before running the rest of the GA tasks.
- `run_ga = TRUE`
  - Executes of the main GA procedure

`[modules]`

The modules section details the names of the individual GA modules used in the subfolders of /src/. Some of these modules (e.g. selection_module=tournament_selection) may be set to alternative options (e.g. selection_module=roulette_selection). For more information on alternative modules see Section 2.

`[initial_pool]`

- `user_structures_dir = initial_pool`
  - Path to the pre-prepared initial pool, as generated by *Genarris*. Structures are in a JSON format.
- `stored_energy_name = energy_tier1`
  - Stored energy name in initial pool json files (if other than "energy"). This should be at the same level of theory as the last control file listed in `control_in_filelist`.

`[run_settings]`

This section controls general GA run settings.

- `num_molecules = 4`
  - This is the number of molecules in the unit cell to be run (must match number of molecules per unit cell in the initial pool). This must be specified by the user.
- `end_GA_structures_added = 5`
  - Setting which stops GAtor after a certain amount of children (in this case 5) have been added to the common pool. For more options for stopping/converging the GA refer to Section 2.
- `output_all_geometries = TRUE`

  - Prints the FHI-aims-style geometries of parents, children, and mutations to the main output `GAtor.out`. This setting is set to TRUE for easy visualization of the structures using Jmol (copy/paste) but may be uncommented for a less verbose output.

- `#skip_energy_evaluations = TRUE`
  - This parameter is uncommented by default, but may be used to skip FHI-aims energy evaluations of the generated structures (giving them an energy of 0 eV). This can be used to verify the structure selection and generation works (e.g. on a laptop) without having to run FHI-aims.

`[parallel_settings]`

- `parallelization_method = subprocess`
  - The parallelization setting of the GAtor replica(s) being run (not for FHI-aims). Subprocessing uses Python's Subprocess module (to run FHI-aims) and can be used on a laptop, or on a cluster (where the Python subprocess will run on the job scheduler nodes). See Chapter 2 for alternative options.

- `number_of_replicas = 1`
  - Number of GAtor replicas being run by user.

- `processes_per_replica = 1`
  - Number of parallel Python processes used per replica. This sets the number of python processes used for parallel GA tasks such as child generation. This make the child generation process faster but should be set with an awareness of the number of processes available on the given machine being used.

- `aims_processes_per_replica = 64`
  - Number of parallel processes used per replica to run FHI-aims for a given replica. For example, this is set to 64 if one is running 1 replica on 1 node of a cluster with 64 processes.

`[FHI-aims]`

- `execute_command = mpirun`
  - Command to run the FHI-aims binary. Since we will be using the scalable version of aims the command is `mpirun`.

- `path_to_aims_executable = aims.mpi.x`
  - Path to FHI-aims executable being used for energy evaluations and/or structural relaxations.

- `control_in_directory = control`
  - Name of the directory in the main calculations folder which can contain multiple FHI-aims control.in files (which can be named arbitrarily).
- `control_in_filelist = control.in.SPE.tier_1 control.in.FULL.tier_1`
  - Name of control file(s) in control directory being used within GAtor. The GA evaluates the control files in order, and the user can set energy cutoffs for each control file (see below). In the tutorial a dummy single-point energy and full relaxation control file are included for demonstrative purposes.
- `relative_energy_thresholds = 3 3`
  - Relative energy cutoffs (in eV) from the current global minimum structure for each control file specified in `control_in_filelist`. If a structure has a relative energy less than the minimum energy structure minus this cutoff, it is immediately rejected. For more energy cutoff options see Section 2.
- `save_failed_calc = TRUE`
  - If uncommented, saves aims calculations if they fail for some reason in `/tmp`
- `save_successful_calc = TRUE`
  - If uncommented, saves full aims calculations data for successful GA structures. Should be commented out if space is an issue.

`[selection]`

This section controls parameters related to the specific `selection_module` chosen.

- `tournament_size = 3`
  - For tournament selection, this controls the tournament size.

`[crossover]`

This section controls parameters related to the specific `crossover_module` chosen.

- `crossover_probability = 0.5`
  - This parameter controls the probability of crossover for a child structure. If set to 0.5 the child has a 50% chance of undergoing crossover, and a 50% chance of undergoing mutation. A separate parameter is not needed for mutation.

`[mutation]`

This section controls parameters related to the specific `mutation_module` chosen.

- `stand_dev_trans = 0.5`
  - Standard deviation (in Angstrom) of the random translation mutations applied.
- `stand_dev_rot = 30`
  - Standard deviation (in degrees) of the random rotation mutations applied.
- `stand_dev_strain = 0.3`
  - Standard deviation of the random strain mutations applied.

`[cell_check_settings]`

This section controls parameters related to the geometric constraints of generated crystal structures. Structures are rejected if they don't pass these constraints.

- `target_volume = 473`
  - The mean target_volume for generated structures
- `volume_upper_ratio = 1.4`
  - The upper ratio of target_volume accepted for generated structures.
- `volume_lower_ratio = .6`
  - The lower ratio of target_volume accepted for generated structures.

1.4.3 Filling the initial pool

An initial pool of structures, prepared by the user using the *Genarris* molecular crystal generation package is required to run GAtor. For more information on generating this initial pool see the *Genarris* user's manual. The path to this initial pool of structures is set in the ui.conf file in `[initial pool]/user_structures_dir`. To start, comment out `run_GA = TRUE` and comment `fill_initial_pool = TRUE`. This will just fill the initial pool without running the GA. Then run the master script

```
python ../src/GAtor_master.py ui.conf &
```

If the initial pool has properly been filled, one should see a nonempty file in `/tmp/num_IP_structs.dat` that contains the number of initial pool structures.

1.4.4 Running the GA

One the initial pool has been filled you may run the GA by uncommenting `run_GA = TRUE`. One can run the code in-shell (not recommended) by running again

```
python ../src/GAtor_master.py ui.conf &
```

Putting the `&` at the end of the script allows the code to run in the background and frees up your terminal. However, this may take quite a while to finish as FHI-aims calculations are being performed. Therefore, it is highly recommended to submit this command to a cluster. An example `submit_to_cluster.sh` is provided in the tutorial folder. Make sure the number of `aims_processes_per_replica` is set in accordance with the number of processes allocated on the cluster.

Your GAtor replica is now running! The next step is to look at the different output files being produced.

1.4.5 Individual and Combined Replica Outputs

The output an individual replica is stored in, e.g.,

```
./tmp/replica_out/fa2f201fe6.out
```

This file records information from the genetic algorithm tasks from each iteration an individual replica and is reset when either a structure is rejected or successfully accepted. This file includes details from selection, crossover, mutation, comparison, and FHI-aims evaluation.

Since in the example ui.conf it was elected to output all FHI-aims geometries to the replica outputs in the configuration file, you may choose to visualize the geometries from the most recent parents, children, and mutations by copy/pasting their FHI-aims geometries from the replica output file into Jmol. The combined output from all successful iterations of all running replicas is stored in.

```
./GAtor.out
```

This combined output file is written to every time any replica starts and finishes an iteration. Feel free to inspect this file as GAtor runs.

1.4.6 GAtor time log

A time log that mainly entails information on the execution of FHI-aims energy evaluations for all replicas is stored in,

```
./GAtor.log
```

The user can refer to this file for inquiring when the latest FHI-aims evaluation for their replica has started and stopped as well as any errors that may have passed (hopefully not...).

1.4.7 Temp Directories for FHI-aims evaluations

The currently-running FHI-aims calculation folders are located in the directory, e.g.

```
./tmp/fa2f201fe6
```

If you change into this directory you will find the control.in, geometry.in, and aims.out files for the currently-running FHI-aims calculation for your replica, as well as a JSON file which includes properties of the currently running structure. The user can inspect aims.out if they wish to know exactly where an FHI-aims evaluation is at.

1.4.8 Structures directory

The database of the entire common pool of the genetic algorithm is located in the the directory, e.g.,

```
./structures/S:6_C:16_N:8/0
```

Each subfolder in this directory corresponds to one structure in the pool, and they are named according to random-indices (if they are an initial pool structure their original name is used). FHI-aims geometries as well as JSON files (which store the structure's geometry and properties) are stored in these files. Feel free to inspect any of these directories.

1.4.9 Energy Hierarchy

An energy hierachy, which ranks structures from the database by their energy is updated in,

```
./tmp/energy_hierarchy_S:6_C:16_N:8_0.dat
```

If you inspect this file, you will see it includes key information from each structure in the collection including their energy ranking, the size of the pool when they were added (initial pool structures have a value of 0, and GA structures indicate the size of the collection when the structure was added), which replica they came from, their structure index, their energy, their unit cell volume and parameters, and their spacegroup. Additionally, for GA-added structures, information about the mutation procedures performed to generate the structure, as well as the indices of the structure's parents, are included. This file is often the simplest one to look at to see if new structures have been added, and where they fall energy-wise in the collection.

1.4.10 Duplicates

An essential part of any genetic algorithm is the identification of duplicate structures as they are inevitably generated in random crossover processes. The database of structures that are deemed as duplicates (and not included in the common pool) are found in,

```
./structures/S:6_C:16_N:8/duplicates
```

Within the GA, GAtor uses *pymatgen's* `StructureMatcher` class to identify duplicate structures within a user-defined energy window. For more information on changing these duplicate tolerances from their default values, see the user manual.

Chapter 2

Full Configuration File Parameters

The configuration file ui.conf (or [user_defined_label].conf) is the only file the user has to modify to control all parameters used in GAtor. Listed below are all the possible parameters for ui.conf, listed under their respective section headings.

`[GAtor_master]`

- `fill_initial_pool =` (optional; Boolean)
  - If present, fills the user-defined initial pool into the common pool of structures before running the GA. The user should omit this keyword if the pool has already been filled and there are just desiring adding another replica to write to the common pool.
- `run_ga =` (optional; Boolean)
  - If present, enables execution of the main GA procedure. See the `parallel_settings` section for details on spawning additional replicas of GAtor.

`[run_settings]`

- `num_molecules =` (required; integer)
  - Number of molecules per unit cell for the current search (must match number of molecules in initial pool structures).
- `orthogonalize_unit_cell =` (optional yet recommended; Boolean, set to TRUE or omit)
  - If `TRUE` will orthogonalize all structures in the initial pool whose lattice vector angles are less than 60 degrees or greater than 120 degrees.
- `end_GA_structures_added =` (optional; integer)

  - A simple way to end the GA by stopping after this many structures have been added by the GA.

- `end_GA_structures_total =` (optional; integer)
  - A simple way to end the GA by stopping after this many structures total structures are in the common pool. This includes the structures added by the GA and the structures in the initial pool.
- `followed_top_structures =` (optional, must be used with `max_iterations_no_change`; integer)
  - Track the top number of structures (as ranked by their energy) to see if they have changed in `max_iterations_no_change`. This is a way of determining convergence.
- `max_iterations_no_change =` (optional, must be used with `followed_top_structures;` integer)
  - If `followed_top_structures` hasn't changed in `max_iterations_no_change`, then stop the GA.
- `verbose=` (optional; Boolean, set to TRUE or omit)
  - If `TRUE`, include for detailed information printed to outputs.
- `output_all_geometries =` (optional; Boolean, set to TRUE or omit)
  - Set to `TRUE` to enable replica output of FHI-aims style geometry whenever a new trial structure is generated or altered.
- `failed_generation_attempts =` (optional; default = 1000)
  - Number of attempts allowed for the structure generation scheme to fail (e.g., failed cell check) before an error is raised.

`[parallel_settings]`

- `parallelization_method =` (optional; default = `serial`)
  - `serial` With this setting only one GA replica which reads and writes to the common pool is spawned. If this setting is used no additional keywords need to be specified in `[parallel_settings]`. If desired, additional simple multiprocessing can be used within the single replica (for parallel python processes such as child creation) by setting `processes_per_replica`. Make sure to not oversubscribe processes of the master node (especially log-in nodes).
  - `subprocess` With this setting the user can spawn several replicas of the GA in the master node (or where GAtor is running) using Python subprocessing. This setting also requires `number_of_replicas` to be set.

- `mpirun` With this setting the user can spread several replicas of the GA across multiple computing cores or nodes using the *mpirun* command. This setting requires additionally setting at least one of the following: `number_of_replicas`, `processes_per_replica`, or `nodes_per_replica`. If only one of these options is specified, GAtor will automatically calculate the others based on the available resources. If more than one of these options is specified, GAtor will check compatibility of the parameters with the system and proceed. Below are a few common scenarios in a sample job which has been submitted to 20 nodes with each node having 20 processes per node.
  - The user specifies `number_of_replicas = 10`. GAtor will allocate 2 nodes and 40 processes total for each of the 10 replicas.
  - The user specifies `number_of_replicas = 40`. GAtor will allocate 10 processes for each of the 40 parallel replicas. This means 2 replicas will be running per node.
  - The user specifies `number_of_replicas = 3`. GAtor will allocate 7 nodes = 140 processes each for 2 replicas, and 6 nodes = 120 processes to 1 replica.
  - The user specifies `processes_per_replica = 10`. GAtor will spawn 40 replicas (with 2 replicas assigned to each node) with 10 processes per replica.
  - The user specifies `processes_per_replica = 30`. GAtor will spawn 10 replicas, each assigned 2 nodes, but each replica only being assigned 30 processes each (used e.g. for memory requirements). User has to specify `additional_arguments` in order for the 30 processes to be evenly distributed across the 2 nodes (e.g. `-rr` for round-robin).
  - The user specifies `processes_per_replica = 6`. GAtor will spawn 60 replicas, assign 3 replicas to each node, and allocate 6 processes to each replica.
  - The user specifies `nodes_per_replica = 4`. GAtor will spawn 5 replicas,and allocate 4 nodes (80 processes) to each replica.
  - The user specifies `processes_per_replica = 20` and `nodes_per_replica = 2`. GAtor will spawn 10 replicas, each allocated 2 nodes with 20 processes total. The user has to specify `additional_arguments` (e.g., -rr for round-robin) in order for the 20 processes to be evenly distributed across the 2 nodes.
  - The user specifies `number_of_replicas = 5` and `nodes_per_replica = 2`. GAtor will spawn 5 replicas, each allocated 2 nodes and 40 processes total.
  - The user specifies `number_of_replicas = 20` and `nodes_per_replica = 1` and `processes_per_replica = 15`. GAtor will spawn 20 replicas, each on 1 node with 15 processes.
- If ValueError is raised when using `mpirun` for a job submitted to, e.g., 20 nodes with 20 processes per node, it is possibly caused by scenarios similar to the following:
  - The user specified `number_of_replicas = 10` and `nodes_per_replica > 2`. GAtor will raise a ValueError for oversubscription of nodes.

    - The user specified `number_of_replicas = 10` and `processes_per_replica > 40`. GAtor will raise a ValueError for oversubscription of processes.
  - `srun` With this setting the user can spread several replicas of the GA across multiple computing cores or nodes using the *srun* command. The same parallelization procedure is used as with the setting `mpirun`. See the description for `mpirun` for parameters requirements and how nodes and processes are distributed to each replica.
  - `mira` Special implementation for ALCF's IBM BG/Q cluster Mira. Required additional parameter: `nodes_per_replica`. Additional Python instances of GAtor will be spawned through subprocess on the front-end nodes. The blocks and corners in the back-end nodes are automatically assigned to each replica. Each replica can be assigned more front-end processes by the
  - `processes_per_replica` parameter.
  - `cetus` Special implementation for ALCF's IBM BG/Q testing cluster Cetus. Required additional parameter: `nodes_per_replica`. See the setting `mira` for further details. Different from the `mira` setting in that by default, blocks of 128 nodes are created, instead of 512.

- `python_command` (optional; default: python)
  - The command used to call Python. This parameter can be set to call an alternative version of Python.

- `number_of_replicas`
  - Required in "subprocess" parallelization mode; optional in "mpirun" and "srun"; ignored in "mira" and "cetus"
  - Number of parallel replicas running the GA.

- `processes_per_replica` (optional)
  - Number of processes allocated to each replica.
  - In "subprocess", "mira" or "cetus" parallelization modes, defaults to 1.
  - In "mpirun" and "srun" modes, defaults to be calculated according to other specified parameters. (See description above about the `mpirun` mode).

- `processes_per_node` (optional)
  - A further constraint on the size of a multiprocessing pool of workers that each replica can spawn. Useful when replicas control more than 1 node to constrain the amount of workers spawned on the main node. The smaller between codeprocesses_per_replica and `processes_per_node` determines the size of the multiprocesses.pool.
  - Honored only in "mpirun" and "srun" parallelization modes.

  - Defaults to the value obtained through mpirun a Python test code on a node.

- `allocated_nodes` (optional)
  - Nodes allocated for this replica. While additional replicas are spawned, this value is set internally to allocate nodes to each replica.
  - Honored only in "mpirun" and "srun" parallelization mode
  - Defaults to the returned value of the function, *parallel_run.get_all_hosts()*.

- `replica_name` (optional; default: "master"):
  - Name of the currently running process.
  - A random replica name is assigned while internally spawning replicas, or when the main GA processes begin with this parameter still being the default "master" (to avoid conflict of names).

- `im_not_master_replica` (optional; Boolean):
  - If present and set to TRUE, suppresses all initialization information printed to time log.

    Here are a few parameters specifically set for the implementation on system using the `srun` command. Note that overcommitting memory resources will lead to job unable to run. To successfully run on system with srun, make sure to allocate the necessary general resources in the submission file.

- `srun_max_runtime`:
  - Maximum run time in seconds before the master process kills the job

- `srun_gator_memory` (optional; default = 2048):
  - Memory (in MB) devoted to the GAtor python processes spwaned in a different node.

- `srun_memory_per_core` (optional; default = 1024):
  - Memory per core (in MB) devoted to additional srun processes (e.g., for FHI-aims calculations).

- `srun_command_file` (optional; default = ./srun_calls.info)
  - The path to the file where each replica sends an `srun` call's command to be picked up by the master thread that spawned all the replicas. This is necessary because `srun` does not allow nested calls.

- `srun_submitted_file` (optional; default = ./srun_submitted.info):
  - The path to the file where the internal job id of srun calls that are picked up by master process and executed is recorded

- `srun_completed_file` (optional; default = ./srun_completed.info)
  - The path where completed commands are sent to notify replicas to pick up results.
- `srun_gres_name` (optional; default = "craynetwork"):
  - Name to the generic resource to that serves as the first field in the argument `-gres` for an srun command. Make sure to configure such resources in the original submission file.
    Here are a few parameters specifically set for the implementation on IBM's BG/Q system with the runjob command:
- `bgq_block_size` (optional):
  - Number of nodes per booted block
  - Defaults to 512 for `mira` mode, 128 for `cetus` mode.
- `runjob_processes_per_node` (optional; default: 16):
  - Number of processes per node. Should be set to the number of cores per node.
- `runjob_block` (optional):
  - For internal distribution of nodes only. The block that is assigned to the replica.
- `runjob_corner` (optional):
  - For internal distribution of nodes only. The corner that is assigned to the replica.
- `runjob_shape` (optional):
  - For internal distribution of nodes only. The shape of the corner that is assigned to the replica.

`[bundled_run_settings]`

- `parallelization_method` (required)
  - Parallelization method to spawn additional replicas.
  - Currently only supporting `mira` and `cetus`
  - `mira` Achieves node distribution for bundled run on ALCF's IBM BG/Q cluster Mira. Required additional parameters for each run: `number_of_blocks`, `nodes_per_replica`
- `run_names` (required)
  - Names of each one of the bundled runs. Given as a list of strings delimited by space. Each run must have a section in the configuration file bearing the same section name, where the additional parameters are stored.

- `bgq_block_size` (optional):
  - Number of nodes per booted block on Mira or Cetus.
  - Defaults to 512 for `mira` mode, 128 for `cetus` mode.
- `runjob_processes_per_node` (optional; default: 16):
  - Number of processes per node. Should be set to the number of cores per node.

Here are the additional parameters that should be included in the section for each of the bundled runs:

`[sample_bundled_run_section]`

- `working_directory` (required)
  - Working directory for this run.
- `config_file_path` (required)
  - Path to the configuration file for this run.
- `number_of_blocks`
  - Number of blocks for this run.
  - Required for `mira` and `cetus` mode.
- `nodes_per_replica`
  - Nodes per replica for this run.
  - With the `mira` and `cetus` mode, this value should divide the block size.

`[FHI-aims]`

- `path_to_aims_executable =` (required; /path/to/aims.x)
  - Path to FHi-aims executable being used for energy evaluations and/or structural relaxations.
- `execute_command =` (required; `mpirun`, `srun`, `runjob`, or `shell`).
  - Command to run the FHI-aims binary. The shell command should be used when calling a serial version of aims via /path/to/aims.x control.in. Note that if `execute_command = shell`, then `additional_arguments` will not be appended to the execute command.
- `additional_arguments=` (optional; not valid if `execute_command = shell`)

  - A Python-evaluable list of strings to append as additional arguments used in the subprocess.Popen call of the FHi-aims binary. For example, set this to ["-rr"] to enable round-robin spawning method in mpirun. Or set this to ["–envs","OMP_NUM_THREADS=4"] to allow the runjob command to alter the environmental variable, OMP_NUM_THREADS. Note that the nodes and processes information are automatically included in the argument list via keywords set in `parallel_settings`.

- `control_in_directory =` (required; control_directory_name)
  - Folder name in current directory that holds the control.in files used within GAtor.
- `control_in_filelist =` (required; control.in.1 control.in2 ...)
  - Folder name in current directory that holds the control.in files used within GAtor for successive steps of the FHI-aims cascade. These can be named arbitrarily in the `control_in_directory`. e.g. perform single point calculations with control.in.1 and full relaxations with control.in.2.
- `monitor_execution =` (optional; Boolean)
  - If present, enables monitoring of the FHI-aims binary call spawned through Python's subprocess.Popen module. The monitoring involves: (1) Confirmation of successful job launch, and (2) prevention of job being hung. A job is given 10 attempts to launch before being determined as failed.
- `absolute_thresholds=` (optional; energy1 energy2 ...)
  - List of highest total energies (in eV) allowed for a structure to to be deemed as acceptable for each level of `control_in_filelist`. Must match length of `control_in_filelist`. For example, if one does not want to allow into the common pool structures with a single point energy higher than -45,000 eV or a fully-relaxed energy higher than -45,575 eV, then `absolute_thresholds= -45000 -45575`.
- `relative_energy_thresholds=` (optional; rel_energy1 rel_energy2 ...)
  - Energy (in eV) allowed for a structure to to be deemed as acceptable for each level of `control_in_filelist`, relative to the current running global minimum. Must match length of `control_in_filelist`. For example, if one does not want to allow into the common pool structures with a single point energy 5 eV higher than minimum energy in the pool or a fully-relaxed energy higher than 3 eV than the minimum energy in the pool, then `relative_thresholds= 5 3`.
- `double_store_last_energy` (optional; Boolean)

  - Enables additional storing of the energy obtained from the last tier of the FHI-aims cascade to the key `[run_settings]/property_to_optimize`. For example, this parameter can be used if the final tier of FHI-aims is first stored as "energy_tier_1_full_relax" but should be further used for fitness evaluation of the structure when the property being optimized is simply named “energy".

- `absolute_success` (optional; Boolean)
  - If set to `TRUE`, requires "Have a nice day" to appear in the FHI-aims output file in order for a job to be determined as successful.

- `save_failed_calc` (optional; Boolean)
  - If set to `TRUE`, entire failed FHi-aims calculation folders will be saved to (./failed_calc).

- `save_successful_calc` (optional; Boolean)
  - If set to `TRUE`, entire successful FHI-aims calculation folders will be saved to (./successful_calc). By default, only necessary information such as a structures energy and geometry are saved from FHI-aims' outputs before the output files are discarded.

- `update_poll_interval =` (required if `monitor_execution = TRUE`; time)
  - Length of time in seconds to sleep between two checks on the FHI-aims output file. An FHI-aims job must output something within the time period of `update_poll_interval` * `update_poll_times`; otherwise, the job is determined to be hung. Must match the length of `control_in_filelist`.

- `update_poll_times =` (required if `monitor_execution = TRUE`; integer)
  - Number of times the FHI-aims output file is polled without new updates before determining that the FHI-aims job has hung. Must match the length of `control_in_filelist`.

`[initial_pool]`

- `user_structures_dir =` (required; /path/to/user_defined_initial_pool)
  - Path to the user-defined initial pool, as generated by *Genarris*.

- `duplicate_check =` (optional; Boolean)
  - If present, will perform a duplicate check on the initial pool of structures by 1) computing cosine similarity between the RDF vectors of pairs structures in the initial pool. If these vectors are determined as similar as defined by `RDF_sym_tol` then 2) pymatgen's structure_matcher function is called.

- `vector_cosdiff_threshold =` (used when `duplicate_check = TRUE` and `vector_for_comparison` is set; default = 0.001)
  - This parameter sets the tolerance for determining if two `vector_for_comparison` vectors from pairs of structures in the initial pool are similar using cosine similarity. If similiar, structure's will further be checked for duplication with pymatgen's structure comparer. If not set by user, only pymatgen's structure comparer will be used, but this takes more time depending on the size of the initial pool.
- `vector_for_comparison =` (string; used when `vector_cosdiff_threshold` is set)
  - Name of vector in initial pool jsons (e.g. a distance, RDF, or fingerprint function) to be compared as a preliminary measure to determine if two structures are similar.
- `scale_vol =` (used when `duplicate_check = TRUE`; Boolean)
  - This option determines whether or not to scale the cell volume of two structures when using pymatgen's structure_matcher. If not set as `TRUE` by user, the volume is not scaled by default.
- `ltol =` (used when `duplicate_check = TRUE`; default = 0.2)
  - This parameter determines the fractional length tolerance of lattice vectors allowed between two duplicate structures using pymatgen's structure_matcher. If not set by user the default value is used.
- `stol =` (used when `duplicate_check = TRUE`; default = 0.3)
  - This option determines the site tolerance allowed between two duplicate structures using pymatgen's structure_matcher. It is defined as the fraction of the average free length per atom. If not set by user the default value is used.
- `angle_tol =` (used when `duplicate_check = TRUE`; default = 3 (degrees))
  - This parameter determines the lattice vector angle tolerance allowed between two duplicate structures using pymatgen's structure_matcher. If not set by user the default value is used.

`[cell_checks]`

- `full_atomic_distance_check` (optional; default= 0.211672 Angstrom)
  - Enforces a minimum distance between all pairs of atoms in the system. The default value 0.211672 Å is the equivalent of 0.4 bohr, which is the minimum distance enforced by FHI-aims.

- `interatomic_distance_check =` (optional; default= 1 Angstrom) If present, enforces a minimum distance for atom pairs from different molecules. This value should usually be set larger than full_atomic_distance_check to enforce a larger distance between atoms from different molecules.

- `COM_distance_check =` (optional) If present, enables the COM distance check, which enforces minimum distance between the center of mass of different molecules.

- `specific_radius_proportion =` (optional)

  - A closeness check for potential structures where each atom is assigned a specific radius (by default, their van der Waals radii). In this check, two atoms from different molecules need to be at least a certain proportion (specified by this parameter, which is often shortened as s_r) of the sum of their specific radii apart. E.g. the van der Waals radius of carbon is 1.70 Å, nitrogen's is 1.55 Å; thus if sr=0.75, then any pair of intermolecular C-N contact must be at least (1.70+1.55)*0.75=2.44 Å apart.

- `target_volume = None` (optional) If present, enables volume checks on generated structures. Enforces the volume of a newly generated structure to be within target_volume*volume_lower_ratio - target_volume*volume_upper_ratio.:wq

- `volume_upper_ratio = 1.2` (optional) The upper ratio that defines the lower bound of the volume of a newly generated structure when times by the target_volume.

- `volume_lower_ratio = 0.8` The lower ratio that defines the lower bound of the volume of a newly generated structure when times by the target_volume.

`[selection]`

- `percent_best_structs_to_select` (optional; default = 100)

  - The user may set this parameter if they wish to bias selection to only a certain percentage of top fitness structures.

- `fitness_function =` (optional; default = `standard`)

  - If the user wishes fitness to be calculated on a linear scale, the default value of `standard` is used, and the user doesn't need to explicitly specify this parameter. However, if the user wishes for fitness to be scaled in an exponential fashion, they may choose to set this paramater to `exponential`.

- `fitness_reversal_probability =` (optional; default = 0.0)

  - The user may set this parameter to be between 0.0 and 1.0 to allow a probability of the fitness function being reversed when selecting parents. This may create better diversity in the pools to allow an occasional unfit structure to be selected.

- `pre_relaxation_comparison` (optional; defaults: `ltol = 0.2`, `stol = 0.4`, `angle_tol = 3`)

  - Determines if a structure is too similar to an existing structure in the collection before passing it on to relaxation.(See `[initial_pool]` or pymatgen's structure comparer for definitions of these parameters).
  - User may also set `scale_vol = TRUE` if they wish to scale the structure's before comparison. By default this keyword is omitted and structures are not scaled.
  - It is recommended to to set `stol` larger than in the `post_relaxation_comparison` section since in this section you are most likely comparing un-relaxed structures to relaxed one.

- `post_relaxation_comparison` (optional; defaults: `ltol = 0.2`, `stol = 0.3`, `angle_tol = 3`)
  - Determines if a structure is too similar to an existing structure in the collection after it has been relaxed (See `[initial_pool]` or pymatgen's structure comparer for definitions of these parameters).
  - User may also set `scale_vol = TRUE` if they wish to scale the structure's before comparison. By default this keyword is omitted and structures are not scaled.
  - It is recommended to to set `stol` smaller than in the `pre_relaxation_comparison` section since in this section you are most likely comparing newly relaxed structures to relaxed structures in the common pool.

`[mutation]`

- `mutation_probability =` (required)
  - This parameter sets the probability of performing mutation on structures which have been crossed over. IF set to, e.g., 0.3, the structure has a 30% chance of undergoing mutation. Can be set $0.0 \leq$ mutation_probability ≤ 1.0.
- `double_mutate_prob =` (optional)
  - A user may set this parameter to allow double mutations on crossover structures. If set, the probability of a structure undergoing double mutation is `double_mutate_prob` * `mutation_probability`
- `stand_dev_trans =` (optional; default = 0.3 A)
  - Sets the standard deviation of the random translation mutations to the COM of the molecules in the cell. The translations are randomly picked from a gaussian distribution of this with.
- `stand_dev_rot =` (optional; default = 5 degrees)
  - Sets the standard deviation of random rotation mutations to the COM of the molecules in the cell (euler angles).

- `stand_dev_strain =` (optional; default = 0.25) This parameter sets the standard deviation of mutations which involve strain (using a generic strain tensor) on the lattice of the child structure. It's a proportional parameter so, e.g., (default $stand\_dev\_strain = 0.25$ so e.g. $Ax_{strain} = Ax + (0.25 * Ax\ )$).
- `enable_symmetry` (optional; Boolean)
  - If set to `TRUE`, allows mutation to preserve the highest level of symmetry in the pre-mutation structure.

`[symmetric_crossover]`

The term, "seed molecules," means the symmetrically independent molecules within a structure. The symmetric crossover module takes the 1st selected structure as standard and conducts crossover that blends/swaps certain features of the 1st structure with/by that of the 2nd.

- `swap_sym_prob` (optional; default = 0.50)
  - The probability of the symmetry operation of the 2nd structure to be applied to the 1st. In this case, the seed molecules of the 1st structure become those closest, in terms of absolute COM coordinates, to the seed molecules of the 2nd structure.
- `swap_sym_tol` (optional; default = 0.01)
  - Tolerance for determining whether the 2nd structure's symmetry operations are compatible with the 1st structure's lattice vectors.
- `blend_lat_prob` (optional; default = 0.50)
  - The probability for the lattice vectors to be blended during crossover. If without blending, the vectors will be taken straight from the 1st selected structure.
- `blend_lat_tol` (optional; default = 0.01)
  - Tolerance for determining whether the blended lattices are compatible with the symmetry operations.
- `blend_lat_cent` (optional; default = 0.50)
  - The center of the Gaussian sampling for the blending parameter, b. Let L_1 be the lattice matrix of the first structure, L_2 be that for the second. Then the blended lattice matrix will be $b \cdot L_2 + (1 - b) \cdot L_1$. Therefore, $b = 0$ takes the unchanged lattice of first structure. $b = 1$ takes the unchanged lattice of second structure.
- `blend_lat_std` (optional; default = 0.25)
  - The standard deviation for the Gaussian sampling of the blending parameter.

- `blend_lat_ext` (optional; Boolean)
  - If set to TRUE, then the blending parameter can be smaller than 0 or greater than 1.
- `blend_mol_COM_prob` (optional; default = 0.50)
  - The probability for the COM of the molecules to be blended during a crossover. During the blending process, each "seed molecule" in the 1st structure will be paired up with a closest neighbor in the 2nd structure, in terms of absolute COM coordinates. If without blending, the absolute COM coordinates will be taken from the 1st selected structure.
- `blend_mol_COM_cent` (optional; default = 0.50)
  - The center of the Gaussian sampling for the blending parameter, b. Let c_1 be the COM of the seed molecule. Let c_2 be the COM of the paired molecule. Then the COM of the seed molecule will be moved to $b \cdot c_2 + (1 - b) \cdot c_1$. Thus, $b = 0$ takes the unchanged COM positions of the seed molecule in the first structure. $b = 1$ takes that of the second structure. If there are multiple seed molecules, b is generated seperately for each blending.
- `blend_mol_COM_std` (optional; default = 0.25)
  - The standard deviation for the Gaussian sampling of the blending parameter.
- `blend_mol_COM_ext` (optional; Boolean)
  - If set to TRUE, then the blending parameter can be smaller than 0 or greater than 1.
- `swap_mol_geo_prob` (optional; default = 0.50)
  - The probability for the molecule geometry of the 2nd structure to be swapped into the 1st. The final orientation will be selected from 20 random orientations that have the least coordinate residual from the original geometry in the 1st structure.
- `swap_mol_geo_tol` (optional; default = 3.0)
  - The tolerance on coordinate residual in determining whether two molecule conformations are the same. If yes, then the geometry will not be swapped. If all pairs of molecules are the same, then this operation will be ruled invalid.
- `swap_mol_geo_orient_attempts` (optiona; default = 100)
  - Number of attempts to randomly orient the swapped geometry. The final orientation is selected to be the orientation that has the least coordinate difference with the original molecule.

- `blend_mol_orien_prob` (optional; default = 0.50)
  - The probability for the orientation of the molecules to be blended during a crossover. During the blending process, each "seed molecule" in the 1st structure will be paired up with a closest neighbor in the 2nd structure, in terms of absolute COM coordinates.
  - If the paired up molecule has different geometry than the seed molecule (see parameter `blend_mol_orien_tol`), then blind blending will be pursued. a number of random rotations (`blend_mol_orien_orient_attempts`) will be applied and the new orientation with the minimum average coordinate difference from the two original molecules will be selected. The average is weighted by the blending parameter b (the coordinate difference from the paired molecule gets weighted as b, while that from the seed molecule gets $1 - b$).
  - The blind blending has a probability specified by `blend_mol_orien_ref_prob` to allow exploration of reflection after applying random rotations. If the exploration is pursued, half of the random rotations will be followed by a mirror reflection across z axis.
  - If the paired up molecule has the same geometry as the seed molecule, then the blending will be based on the calculated mapping information from one to the other. The mapping information gives whether or not a mirror reflection is involved, and a rotation in terms of an axis and an angle in degrees. If the mapping does not involve a mirror reflection, then a portion (b) of the rotation will be applied to the seed molecule as the final rotation.
  - If the mapping involves a mirror reflection, then a mirror reflection is applied with a probability given by `blend_mol_orien_ref_prob` . If the mirror reflection is not applied, then blind blending will be pursued. If it is applied, first the orientation of the reflected molecule that has the smallest coordinate differences from the original will be found (with `blend_mol_orien_orient_attempt` random rotation attempts). Then the mapping information will be recalculated. A portion (b) of the rotation will be applied to the reflected and readjusted molecule geometry as the final orientation. Note that it is likely for the mapping calculation to yield a larger tolerance and thus consider the molecules to be different after reflection is applied. In that case, blind blending will be pursued.
- `blend_mol_orien_cent` (optional; default = 0.50)
  - The center of the Gaussian sampling for the blending parameter, b. See description above for parameter `blend_mol_orien_prob` for how b is used.
- `blend_mol_orien_std` (optional; default = 0.25)
  - The standard deviation for the Gaussian sampling of the blending parameter, b.
- `blend_mol_orien_ext` (optional; Boolean)

  - If set to TRUE, then the blending parameter b can be smaller than 0 or greater than 1.

- `blend_mol_orien_tol` (optional; default = 3.0)
  - The coordinate difference tolerance in determining whether the seed molecule has the same geometry as the paired molecule. See description above for parameter `blend_mol_orien_prob` for how this affects the procedure.

- `blend_mol_orien_ref_prob` (optional; default = 0.5)
  - The probability for mirror reflection to be allowed in the final orientation. See description above for parameter `blend_mol_orien_prob` for the usage of this parameter.

- `blend_mol_orien_orient_attempts` (optional; default = 100)
  - The number of attempts to randomly orient a molecule. See description above for parameter `blend_mol_orien_prob` for the usage of this parameter.

- `allow_no_crossover` (optional; Boolean)
  - If set to TRUE, then a crossover attempt that did not invoke any of the above listed operations will be allowed. Otherwise, a `while` loop will be used until 1 attempt uses any of the operations above.

References

[1] J. Bernstein, *Polymorphism in molecular crystals*, Vol. 14 (Oxford University Press, 2002).

[2] G. M. Day, W. D. S Motherwell, and W. Jones, "A strategy for predicting the crystal structures of flexible molecules: the polymorphism of phenobarbital", Phys Chem Chem Phys **9**, 1693–704 (2007).

[3] A. M. Reilly, and A. Tkatchenko, "Role of dispersion interactions in the polymorphism and entropic stabilization of the aspirin crystal", Phys Rev Lett **113**, 055701 (2014).

[4] D. P. Elder, J. E. Patterson, and R. Holm, "The solid-state continuum: A perspective on the interrelationships between different solid-state forms in drug substance and drug product", Journal of Pharmacy and Pharmacology **67**, 757–772 (2015).

[5] C. Reese, and Z. Bao, "Organic single-crystal field-effect transistors", Materials Today **10**, 20–27 (2007).

[6] T. Hasegawa, and J. Takeya, "Organic field-effect transistors using single crystals", Science and Technology of Advanced Materials **10**, 024314 (2009).

[7] S. Bergantin, and M. Moret, "Rubrene polymorphs and derivatives: The effect of chemical modification on the crystal structure", Crystal Growth and Design **12**, 6035–6041 (2012).

[8] P. Cudazzo, M. Gatti, and A. Rubio, "Excitons in molecular crystals from first-principles many-body perturbation theory: picene versus pentacene", Phys. Rev. B **86**, 195307 (2012).

[9] P. Cudazzo, F. Sottile, A. Rubio, and M. Gatti, "Exciton dispersion in molecular solids", J Phys Condens Matter **27**, 113204 (2015).

[10] N. Panina, F. J. J. Leusen, F. F. B. J. Janssen, P. Verwer, H. Meekes, E. Vlieg, and G. Deroover, "Crystal structure prediction of organic pigments: quinacridone as an example", J Appl Crystallogr **40**, 105–114 (2007).

[11] M. Fitzgerald, M. G. Gardiner, D. Armitt, G. W. Dicinoski, and C. Wall, "Confirmation of the molecular structure of tetramethylene diperoxide dicarbamide (tmdd) and its sensitiveness properties", J Phys Chem A **119**, 905–10 (2015).

[12] S. L. Price, "Why don't we find more polymorphs?", Acta Crystallographica Section B: Structural Science, Crystal Engineering and Materials **69**, 313–328 (2013).

[13] S. L. Price, D. E. Braun, and S. M. Reutzel-Edens, "Can computed crystal energy landscapes help understand pharmaceutical solids?", Chem. Commun. **52**, 7065–7077 (2016).

[14] B. L. Lichterman, "Aspirin: the story of a wonder drug", BMJ: British Medical Journal **329**, 1408 (2004).

[15] P. Vishweshwar, J. A. McMahon, M. Oliveira, M. L. Peterson, and M. J. Zaworotko, "The predictably elusive form ii of aspirin", Journal of the American Chemical Society **127**, 16802–16803 (2005).

[16] C. Ouvrard, and S. L. Price, "Toward crystal structure prediction for conformationally flexible molecules: the headaches illustrated by aspirin", Crystal Growth & Design **4**, 1119–1127 (2004).

[17] S. Varughese, M. Kiran, K. A. Solanko, A. D. Bond, U. Ramamurty, and G. R. Desiraju, "Interaction anisotropy and shear instability of aspirin polymorphs established by nanoindentation", Chemical Science **2**, 2236–2242 (2011).

[18] A. G. Shtukenberg, C. Hu, Q. Zhu, M. U. Schmidt, W. Xu, M. Tan, and B. Kahr, "The third ambient aspirin polymorph", Crystal Growth & Design (2017).

[19] J. Aaltonen, M. Allesø, S. Mirza, V. Koradia, K. C. Gordon, and J. Rantanen, "Solid form screening–a review", European Journal of Pharmaceutics and Biopharmaceutics **71**, 23–37 (2009).

[20] X. Wang, T. Garcia, S. Monaco, B. Schatschneider, and N. Marom, "Effect of crystal packing on the excitonic properties of rubrene polymorphs", CrystEngComm **18**, 7353–7362 (2016).

[21] R. Tseng, R. Chan, V. Tung, and Y. Yang, "Anisotropy in organic single-crystal photovoltaic characteristics", Adv. Mater. **20**, 435–438 (2008).

[22] R. Pfattner, M. Mas-Torrent, I. Bilotti, A. Brillante, S. Milita, F. Liscio, F. Biscarini, T. Marszalek, J. Ulanski, A. Nosal, M. Gazicki-Lipman, M. Leufgen, G. Schmidt, L. W. Molenkamp, V. Laukhin, J. Veciana, and C. Rovira, "High-performance single crystal organic field-effect transistors based on two dithiophene-tetrathiafulvalene (dt-ttf) polymorphs", Adv. Mater. **22**, 4198–4203 (2010).

[23] M. Wang, J. Li, G. Zhao, Q. Wu, Y. Huang, W. Hu, X. Gao, H. Li, and D. Zhu, "High-performance organic field-effect transistors based on single and large-area aligned crystalline microribbons of 6,13-dichloropentacene", Adv. Mater. **25**, 2229–2233 (2013).

[24] Y. Li, D. Ji, J. Liu, Y. Yao, X. Fu, W. Zhu, C. Xu, H. Dong, J. Li, and W. Hu, "Quick fabrication of large-area organic semiconductor single crystal arrays with a rapid annealing self-solution-shearing method.", Sci. Rep. **5**, 13195 (2015).

[25] V. C. Sundar, J. Zaumseil, V. Podzorov, E. Menard, R. L. Willett, T. Someya, M. E. Gershenson, and J. A. Rogers, "Elastomeric transistor stamps: reversible probing of charge transport in organic crystals", Science **303**, 1644–1646 (2004).

[26] I. N. Hulea, S. Fratini, H. Xie, C. L. Mulder, N. N. Iossad, G. Rastelli, S. Ciuchi, and A. F. Morpurgo, "Tunable fröhlich polarons in organic single-crystal transistors", Nat. Mater. **5**, 982–986 (2006).

[27] V. R. Hathwar, M. Sist, M. R. V. Jørgensen, A. H. Mamakhel, X. Wang, C. M. Hoffmann, K. Sugimoto, J. Overgaard, and B. B. Iversen, "Quantitative analysis of intermolecular interactions in orthorhombic rubrene", IUCrJ **2**, 563–574 (2015).

[28] J. Maddox, "Crystals from first principles", Nature **335**, 201–201 (1988).

[29] M. a. Neumann, J. van de Streek, F. P. a. Fabbiani, P. Hidber, and O. Grassmann, "Combined crystal structure prediction and high-pressure crystallization in rational pharmaceutical polymorph screening", Nature Communications **6**, 7793 (2015).

[30] A. G. Shtukenberg, Q. Zhu, D. J. Carter, L. Vogt, J. Hoja, E. Schneider, H. Song, B. Pokroy, I. Polishchuk, A. Tkatchenko, A. R. Oganov, A. L. Rohl, M. E. Tuckermanejk, and B. Kahral, "Powder diffraction and crystal structure prediction identify four new coumarin polymorphs", Chemical Science (2017).

[31] A. Gavezzotti, "Are crystal structures predictable?", Accounts of chemical research **27**, 309–314 (1994).

[32] W. L. Jorgensen, and J. Tirado-Rives, "Monte carlo vs molecular dynamics for conformational sampling", The Journal of Physical Chemistry **100**, 14508–14513 (1996).

[33] M. A. Neumann, "Tailor-made force fields for crystal-structure prediction", The Journal of Physical Chemistry B **112**, 9810–9829 (2008).

[34] J. Wang, R. M. Wolf, J. W. Caldwell, P. A. Kollman, and D. A. Case, "Development and testing of a general amber force field", Journal of computational chemistry **25**, 1157–1174 (2004).

[35] K. Vanommeslaeghe, E. Hatcher, C. Acharya, S. Kundu, S. Zhong, J. Shim, E. Darian, O. Guvench, P. Lopes, I. Vorobyov, et al., "Charmm general force field: a force field for drug-like molecules compatible with the charmm all-atom additive biological force fields", Journal of computational chemistry **31**, 671–690 (2010).

[36] A. M. Reilly, R. I. Cooper, C. S. Adjiman, S. Bhattacharya, A. D. Boese, J. G. Brandenburg, P. J. Bygrave, R. Bylsma, J. E. Campbell, R. Car, D. H. Case, R. Chadha, J. C. Cole, K. Cosburn, H. M. Cuppen, F. Curtis, G. M. Day, R. A. DiStasio, Jr, A. Dzyabchenko, B. P. van Eijck, D. M. Elking, J. A. van den Ende, J. C. Facelli, M. B. Ferraro, L. Fusti-Molnar, C. A. Gatsiou, T. S. Gee, R. de Gelder, L. M. Ghiringhelli, H. Goto, S. Grimme, R. Guo, D. W. M. Hofmann, J. Hoja, R. K. Hylton, L. Iuzzolino, W. Jankiewicz, D. T. de Jong, J. Kendrick, N. J. J. de Klerk, H. Y. Ko, L. N. Kuleshova, X. Li, S. Lohani, F. J. J. Leusen, A. M. Lund, J. Lv, Y. Ma, N. Marom, A. E. Masunov, P. McCabe, D. P. McMahon, H. Meekes, M. P. Metz, A. J. Misquitta, S. Mohamed, B. Monserrat, R. J. Needs, M. A. Neumann, J. Nyman, S. Obata, H. Oberhofer, A. R. Oganov, A. M. Orendt, G. I. Pagola, C. C. Pantelides, C. J. Pickard, R. Podeszwa, L. S. Price, S. L. Price, A. Pulido, M. G. Read, K. Reuter, E.

Schneider, C. Schober, G. P. Shields, P. Singh, I. J. Sugden, K. Szalewicz, C. R. Taylor, A. Tkatchenko, M. E. Tuckerman, F. Vacarro, M. Vasileiadis, A. Vazquez-Mayagoitia, L. Vogt, Y. Wang, R. E. Watson, G. A. de Wijs, J. Yang, Q. Zhu, and C. R. Groom, "Report on the sixth blind test of organic crystal structure prediction methods", Acta Crystallogr B Struct Sci Cryst Eng Mater **72**, 439–59 (2016).

[37] T.-Q. Yu, and M. E. Tuckerman, "Temperature-accelerated method for exploring polymorphism in molecular crystals based on free energy", Physical review letters **107**, 015701 (2011).

[38] E. Schneider, L. Vogt, and M. E. Tuckerman, "Exploring polymorphism of benzene and naphthalene with free energy based enhanced molecular dynamics", Acta Crystallographica Section B: Structural Science, Crystal Engineering and Materials **72**, 542–550 (2016).

[39] R. L. Akkermans, N. A. Spenley, and S. H. Robertson, "Monte carlo methods in materials studio", Molecular Simulation **39**, 1153–1164 (2013).

[40] Y. Wang, J. Lv, L. Zhu, and Y. Ma, "Calypso: a method for crystal structure prediction", Computer Physics Communications **183**, 2063–2070 (2012).

[41] D. J. Wales, and J. P. Doye, "Global optimization by basin-hopping and the lowest energy structures of lennard-jones clusters containing up to 110 atoms", The Journal of Physical Chemistry A **101**, 5111–5116 (1997).

[42] C. J. Pickard, and R. Needs, "Ab initio random structure searching", Journal of Physics: Condensed Matter **23**, 053201 (2011).

[43] D. H. Case, J. E. Campbell, P. J. Bygrave, and G. M. Day, "Convergence properties of crystal structure prediction by quasi-random sampling", Journal of chemical theory and computation **12**, 910–924 (2016).

[44] R. L. Johnston, "Evolving better nanoparticles: genetic algorithms for optimising cluster geometries", Dalton Transactions, 4193–4207 (2003).

[45] M. Sierka, "Synergy between theory and experiment in structure resolution of low-dimensional oxides", Progress in Surface Science **85**, 398–434 (2010).

[46] S. Heiles, and R. L. Johnston, "Global optimization of clusters using electronic structure methods", International Journal of Quantum Chemistry **113**, 2091–2109 (2013).

[47] N. Marom, R. A. DiStasio, V. Atalla, S. Levchenko, A. M. Reilly, J. R. Chelikowsky, L. Leiserowitz, and A. Tkatchenko, "Many-body dispersion interactions in molecular crystal polymorphism", Angew. Chem. Int. Ed. **52**, 6629–6632 (2013).

[48] A. J. Cruz-Cabeza, S. M. Reutzel-Edens, and J. Bernstein, "Facts and fictions about polymorphism", Chemical Society Reviews **44**, 8619–8635 (2015).

[49] G. J. Beran, "A new era for ab initio molecular crystal lattice energy prediction", Angewandte Chemie International Edition **54**, 396–398 (2015).

[50] G. J. Beran, "Modeling polymorphic molecular crystals with electronic structure theory", Chem. Rev. **116**, 5567–5613 (2016).

[51] M. Dion, H. Rydberg, E. Schröder, D. C. Langreth, and B. I. Lundqvist, "Van der waals density functional for general geometries", Physical review letters **92**, 246401 (2004).

[52] K. Lee, É. D. Murray, L. Kong, B. I. Lundqvist, and D. C. Langreth, "Higher-accuracy van der waals density functional", Physical Review B **82**, 081101 (2010).

[53] O. A. Vydrov, and T. Van Voorhis, "Nonlocal van der waals density functional made simple", Physical review letters **103**, 063004 (2009).

[54] R. Peverati, and D. G. Truhlar, "M11-l: a local density functional that provides improved accuracy for electronic structure calculations in chemistry and physics", The Journal of Physical Chemistry Letters **3**, 117–124 (2011).

[55] R. Peverati, and D. G. Truhlar, "An improved and broadly accurate local approximation to the exchange–correlation density functional: the mn12-l functional for electronic structure calculations in chemistry and physics", Physical Chemistry Chemical Physics **14**, 13171–13174 (2012).

[56] Y. Zhao, and D. G. Truhlar, "The m06 suite of density functionals for main group thermochemistry, thermochemical kinetics, noncovalent interactions, excited states, and transition elements: two new functionals and systematic testing of four m06-class functionals and 12 other functionals", Theoretical Chemistry Accounts: Theory, Computation, and Modeling (Theoretica Chimica Acta) **120**, 215–241 (2008).

[57] O. A. Vydrov, and T. Van Voorhis, "Nonlocal van der waals density functional: the simpler the better", The Journal of chemical physics **133**, 244103 (2010).

[58] O. A. Vydrov, and T. Van Voorhis, "Dispersion interactions from a local polarizability model", Physical Review A **81**, 062708 (2010).

[59] K. Berland, V. R. Cooper, K. Lee, E. Schröder, T. Thonhauser, P. Hyldgaard, and B. I. Lundqvist, "Van der waals forces in density functional theory: a review of the vdw-df method", Reports on Progress in Physics **78**, 066501 (2015).

[60] T. Thonhauser, S. Zuluaga, C. Arter, K. Berland, E. Schröder, and P. Hyldgaard, "Spin signature of nonlocal correlation binding in metal-organic frameworks", Physical review letters **115**, 136402 (2015).

[61] H. Peng, Z.-H. Yang, J. P. Perdew, and J. Sun, "Versatile van der waals density functional based on a meta-generalized gradient approximation", Physical Review X **6**, 041005 (2016).

[62] J. Sun, A. Ruzsinszky, and J. P. Perdew, "Strongly constrained and appropriately normed semilocal density functional", Physical review letters **115**, 036402 (2015).

[63] K. E. Riley, M. Pitonák, P. Jurecka, and P. Hobza, "Stabilization and structure calculations for noncovalent interactions in extended molecular systems based on wave function and density functional theories", Chemical Reviews **110**, 5023–5063 (2010).

[64] S. Grimme, "Semiempirical gga-type density functional constructed with a long-range dispersion correction", Journal of computational chemistry **27**, 1787–1799 (2006).

[65] S. Grimme, J. Antony, S. Ehrlich, and H. Krieg, "A consistent and accurate ab initio parametrization of density functional dispersion correction (dft-d) for the 94 elements h-pu", The Journal of chemical physics **132**, 154104 (2010).

[66] E. R. Johnson, and A. D. Becke, "A post-hartree–fock model of intermolecular interactions", The Journal of chemical physics **123**, 024101 (2005).

[67] A. Otero-De-La-Roza, and E. R. Johnson, "A benchmark for non-covalent interactions in solids", The Journal of chemical physics **137**, 054103 (2012).

[68] P. Jurečka, J. Černỳ, P. Hobza, and D. R. Salahub, "Density functional theory augmented with an empirical dispersion term. interaction energies and geometries of 80 noncovalent complexes compared with ab initio quantum mechanics calculations", Journal of computational chemistry **28**, 555–569 (2007).

[69] Q. Wu, and W. Yang, "Empirical correction to density functional theory for van der waals interactions", The Journal of chemical physics **116**, 515–524 (2002).

[70] X. Wu, M. Vargas, S. Nayak, V. Lotrich, and G. Scoles, "Towards extending the applicability of density functional theory to weakly bound systems", The Journal of Chemical Physics **115**, 8748–8757 (2001).

[71] S. N. Steinmann, and C. Corminboeuf, "A generalized-gradient approximation exchange hole model for dispersion coefficients", The Journal of chemical physics **134**, 044117 (2011).

[72] S. N. Steinmann, and C. Corminboeuf, "Comprehensive benchmarking of a density-dependent dispersion correction", Journal of chemical theory and computation **7**, 3567–3577 (2011).

[73] A. Tkatchenko, and M. Scheffler, "Accurate molecular van der waals interactions from ground-state electron density and free-atom reference data", Phys. Rev. Lett. **102**, 073005 (2009).

[74] J. Brandenburg, J. Bates, J. Sun, and J. Perdew, "Benchmark tests of a strongly constrained semilocal functional with a long-range dispersion correction", Physical Review B **94**, 115144 (2016).

[75] R. A. DiStasio, O. A. von Lilienfeld, and A. Tkatchenko, "Collective many-body van der waals interactions in molecular systems", Proceedings of the National Academy of Sciences **109**, 14791–14795 (2012).

[76] A. Tkatchenko, R. A. DiStasio, R. Car, and M. Scheffler, "Accurate and efficient method for many-body van der waals interactions", Phys. Rev. Lett. **108**, 236402 (2012).

[77] A. Ambrosetti, A. M. Reilly, R. A. DiStasio, and A. Tkatchenko, "Long-range correlation energy calculated from coupled atomic response functions", J. Chem. Phys. **140**, 18A508 (2014).

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

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

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

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

[82] D. A. Bardwell, C. S. Adjiman, Y. A. Arnautova, E. Bartashevich, S. X. M. Boerrigter, D. E. Braun, A. J. Cruz-Cabeza, G. M. Day, R. G. Della Valle, G. R. Desiraju, B. P. van Eijck, J. C. Facelli, M. B. Ferraro, D. Grillo, M. Habgood, D. W. M. Hofmann, F. Hofmann, K. V. J. Jose, P. G. Karamertzanis, A. V. Kazantsev, J. Kendrick, L. N. Kuleshova, F. J. J. Leusen, A. V. Maleev, A. J. Misquitta, S. Mohamed, R. J. Needs, M. A. Neumann, D. Nikylov, A. M. Orendt, R. Pal, C. C. Pantelides, C. J. Pickard, L. S. Price, S. L. Price, H. A.

Scheraga, J. van de Streek, T. S. Thakur, S. Tiwari, E. Venuti, and I. K. Zhitkov, "Towards crystal structure prediction of complex organic compounds–a report on the fifth blind test", Acta Cryst. B **67**, 535–551 (2011).

[83] W. Kohn, and L. J. Sham, "Self-consistent equations including exchange and correlation effects", Physical review **140**, A1133 (1965).

[84] M. Gell-Mann, and K. A. Brueckner, "Correlation energy of an electron gas at high density", Physical Review **106**, 364 (1957).

[85] R. A. Coldwell-Horsfall, and A. A. Maradudin, "Zero-point energy of an electron lattice", Journal of Mathematical Physics **1**, 395–404 (1960).

[86] W. Carr Jr, "Energy, specific heat, and magnetic properties of the low-density electron gas", Physical Review **122**, 1437 (1961).

[87] D. M. Ceperley, and B. Alder, "Ground state of the electron gas by a stochastic method", Physical Review Letters **45**, 566 (1980).

[88] S. H. Vosko, L. Wilk, and M. Nusair, "Accurate spin-dependent electron liquid correlation energies for local spin density calculations: a critical analysis", Canadian Journal of physics **58**, 1200–1211 (1980).

[89] J. P. Perdew, and A. Zunger, "Self-interaction correction to density-functional approximations for many-electron systems", Physical Review B **23**, 5048 (1981).

[90] J. P. Perdew, and Y. Wang, "Pair-distribution function and its coupling-constant average for the spin-polarized electron gas", Physical Review B **46**, 12947 (1992).

[91] M. Ernzerhof, J. P. Perdew, and K. Burke, "Coupling-constant dependence of atomization energies", International Journal of Quantum Chemistry **64**, 285–295 (1997).

[92] A. D. Becke, "Density-functional thermochemistry. i. the effect of the exchange-only gradient correction", The Journal of chemical physics **96**, 2155–2160 (1992).

[93] A. D. Becke, "Density-functional thermochemistry. ii. the effect of the perdew–wang generalized-gradient correlation correction", The Journal of chemical physics **97**, 9173–9177 (1992).

[94] C. Fiolhais, F. Nogueira, and M. A. Marques, *A primer in density functional theory*, Vol. 620 (Springer Science & Business Media, 2003).

[95] B. Winkler, V. Milman, B. Hennion, M. Payne, M.-H. Lee, and J. Lin, "Ab initio total energy study of brucite, diaspore and hypothetical hydrous wadsleyite", Physics and Chemistry of Minerals **22**, 461–467 (1995).

[96] D. C. Langreth, and J. P. Perdew, "Theory of nonuniform electronic systems. i. analysis of the gradient approximation and a generalization that works", Physical Review B **21**, 5469 (1980).

[97] J. P. Perdew, "Accurate density functional for the energy: real-space cutoff of the gradient expansion for the exchange hole", Physical Review Letters **55**, 1665 (1985).

[98] J. P. Perdew, K. Burke, and M. Ernzerhof, "Generalized gradient approximation made simple", Phys. Rev. Lett. **77**, 3865–3868 (1996).

[99] J. P. Perdew, K. Burke, and M. Ernzerhof, "Generalized gradient approximation made simple [phys. rev. lett. 77, 3865 (1996)]", Phys. Rev. Lett. **78**, 1396–1396 (1997).

[100] J. P. Perdew, "Density-functional approximation for the correlation energy of the inhomogeneous electron gas", Physical Review B **33**, 8822 (1986).

[101] A. D. Becke, "Density-functional exchange-energy approximation with correct asymptotic behavior", Physical review A **38**, 3098 (1988).

[102] C. Lee, W. Yang, and R. G. Parr, "Development of the colle-salvetti correlation-energy formula into a functional of the electron density", Physical review B **37**, 785 (1988).

[103] J. Sun, M. Marsman, G. I. Csonka, A. Ruzsinszky, P. Hao, Y.-S. Kim, G. Kresse, and J. P. Perdew, "Self-consistent meta-generalized gradient approximation within the projector-augmented-wave method", Physical Review B **84**, 035117 (2011).

[104] A. D. Becke, "Density functional thermochemistry. iii. the role of exact exchange", J. Chem. Phys **98**, 5648–5652 (1993).

[105] L. A. Curtiss, K. Raghavachari, G. W. Trucks, and J. A. Pople, "Gaussian-2 theory for molecular energies of first-and second-row compounds", The Journal of chemical physics **94**, 7221–7230 (1991).

[106] J. P. Perdew, M. Ernzerhof, and K. Burke, "Rationale for mixing exact exchange with density functional approximations", J. Chem. Phys. **105**, 9982–9985 (1996).

[107] C. Adamo, and V. Barone, "Toward reliable density functional methods without adjustable parameters: the pbe0 model", J. Chem. Phys. **110**, 6158–6170 (1999).

[108] J. Paier, R. Hirschl, M. Marsman, and G. Kresse, "The perdew–burke–ernzerhof exchange-correlation functional applied to the g2-1 test set using a plane-wave basis set", The Journal of chemical physics **122**, 234102 (2005).

[109] S. Kristyán, and P. Pulay, "Can (semi) local density functional theory account for the london dispersion forces?", Chemical physics letters **229**, 175–180 (1994).

[110] J. Pérez-Jordá, and A. D. Becke, "A density-functional study of van der waals forces: rare gas diatomics", Chemical physics letters **233**, 134–137 (1995).

[111] G. A. DiLabio, and A. Otero-de-la-Roza, "Noncovalent interactions in density-functional theory", Reviews in Computational Chemistry **29** (2014).

[112] S. Grimme, A. Hansen, J. G. Brandenburg, and C. Bannwarth, "Dispersion-corrected mean-field electronic structure methods", Chemical reviews **116**, 5105–5154 (2016).

[113] G. Román-Pérez, and J. M. Soler, "Efficient implementation of a van der waals density functional: application to double-wall carbon nanotubes", Physical review letters **103**, 096102 (2009).

[114] J. Klimeš, D. R. Bowler, and A. Michaelides, "Van der waals density functionals applied to solids", Physical Review B **83**, 195131 (2011).

[115] Y. Zhang, and W. Yang, "Comment on "generalized gradient approximation made simple"", Physical Review Letters **80**, 890 (1998).

[116] E. D. Murray, K. Lee, and D. C. Langreth, "Investigation of exchange energy density functional accuracy for interacting molecules", Journal of Chemical Theory and Computation **5**, 2754–2762 (2009).

[117] O. A. Vydrov, and T. Van Voorhis, "Improving the accuracy of the nonlocal van der waals density functional with minimal empiricism", The Journal of chemical physics **130**, 104105 (2009).

[118] Y. Zhao, N. E. Schultz, and D. G. Truhlara, "Exchange-correlation functional with broad accuracy for metallic and nonmetallic compounds, kinetics, and noncovalent interactions", The Journal of Chemical Physics **123**, 161103 (2005).

[119] Y. Zhao, N. E. Schultz, and D. G. Truhlar, "Design of density functionals by combining the method of constraint satisfaction with parametrization for thermochemistry, thermochemical kinetics, and noncovalent interactions", Journal of Chemical Theory and Computation **2**, 364–382 (2006).

[120] Y. Zhao, and D. G. Truhlar, "Density functional for spectroscopy: no long-range self-interaction error, good performance for rydberg and charge-transfer states, and better performance on average than b3lyp for ground states", The Journal of Physical Chemistry A **110**, 13126–13130 (2006).

[121] Y. Zhao, and D. G. Truhlar, "A new local density functional for main-group thermochemistry, transition metal bonding, thermochemical kinetics, and noncovalent interactions", The Journal of chemical physics **125**, 194101 (2006).

[122] Y. Zhao, and D. G. Truhlar, "Exploring the limit of accuracy of the global hybrid meta density functional for main-group thermochemistry, kinetics, and noncovalent interactions", Journal of Chemical Theory and Computation **4**, 1849–1868 (2008).

[123] R. Peverati, and D. G. Truhlar, "Improving the accuracy of hybrid meta-gga density functionals by range separation", The Journal of Physical Chemistry Letters **2**, 2810–2817 (2011).

[124] R. Peverati, and D. G. Truhlar, "Screened-exchange density functionals with broad accuracy for chemistry and solid-state physics", Physical Chemistry Chemical Physics **14**, 16187–16191 (2012).

[125] C.-O. Almbladh, and U. von Barth, "Exact results for the charge and spin densities, exchange-correlation potentials, and density-functional eigenvalues", Physical Review B **31**, 3231 (1985).

[126] L. Goerigk, and S. Grimme, "Efficient and accurate double-hybrid-meta-gga density functionals: evaluation with the extended gmtkn30 database for general main group thermochemistry, kinetics, and noncovalent interactions", Journal of Chemical Theory and Computation **7**, 291–309 (2010).

[127] P. W. Atkins, and R. S. Friedman, *Molecular quantum mechanics* (Oxford university press, 2011).

[128] A. D. Becke, and E. R. Johnson, "A density-functional model of the dispersion interaction", The Journal of chemical physics **123**, 154101 (2005).

[129] E. R. Johnson, and A. D. Becke, "A post-hartree-fock model of intermolecular interactions: inclusion of higher-order corrections", The Journal of chemical physics **124**, 174104 (2006).

[130] S. Grimme, "Accurate description of van der waals complexes by density functional theory including empirical corrections", Journal of computational chemistry **25**, 1463–1473 (2004).

[131] S. Grimme, S. Ehrlich, and L. Goerigk, "Effect of the damping function in dispersion corrected density functional theory", Journal of computational chemistry **32**, 1456–1465 (2011).

[132] X. Chu, and A. Dalgarno, "Linear response time-dependent density functional theory for van der waals coefficients", The Journal of chemical physics **121**, 4083–4088 (2004).

[133] F. L. Hirshfeld, "Bonded-atom fragments for describing molecular charge densities", Theoretical Chemistry Accounts: Theory, Computation, and Modeling (Theoretica Chimica Acta) **44**, 129–138 (1977).

[134] R. F. Nalewajski, and R. G. Parr, "Information theory, atoms in molecules, and molecular similarity", Proceedings of the National Academy of Sciences **97**, 8879–8882 (2000).

[135] A. Bondi, "Van der waals volumes and radii", The Journal of physical chemistry **68**, 441–451 (1964).

[136] P. Jurečka, J. Šponer, J. Černỳ, and P. Hobza, "Benchmark database of accurate (mp2 and ccsd (t) complete basis set limit) interaction energies of small model complexes, dna base pairs, and amino acid pairs", Physical Chemistry Chemical Physics **8**, 1985–1993 (2006).

[137] J. Hoja, A. M. Reilly, and A. Tkatchenko, "First-principles modeling of molecular crystals: structures and stabilities, temperature and pressure", Wiley Interdisciplinary Reviews: Computational Molecular Science **7** (2017).

[138] S. Kim, A. M. Orendt, M. B. Ferraro, and J. C. Facelli, "Crystal structure prediction of flexible molecules using parallel genetic algorithms with a standard force field", J Comput Chem **30**, 1973–85 (2009).

[139] A. Donchev, "Many-body effects of dispersion interaction", The Journal of chemical physics **125**, 074713 (2006).

[140] J. F. Dobson, J. Wang, B. P. Dinte, K. McLennan, and H. M. Le, "Soft cohesive forces", International journal of quantum chemistry **101**, 579–598 (2005).

[141] J. F. Dobson, "Validity comparison between asymptotic dispersion energy formalisms for nanomaterials", Journal of Computational and Theoretical Nanoscience **6**, 960–971 (2009).

[142] J. Contreras-García, E. R. Johnson, S. Keinan, R. Chaudret, J.-P. Piquemal, D. N. Beratan, and W. Yang, "Nciplot: a program for plotting noncovalent interaction regions", Journal of chemical theory and computation **7**, 625–632 (2011).