

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 and then employing a cluster-based fitness function. Evolutionary niching promotes uniform sampling of the potential energy surface by evolving several sub-populations, which helps overcome initial pool biases and selection biases (genetic drift). The various settings offered by GAtor increase the likelihood of locating numerous 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 crystals 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 of 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 prop-

erties,^{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.^{13,25–27} Once considered unachievable,²⁸ 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,^{29–32} 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^{33–44} and pairwise methods that add the leading order C^6/R^6 dispersion term to the inter-nuclear energy.^{45–56} Notably, the recently developed many-body dispersion (MBD) method^{57–59} accurately describes the structure, energetics, dielectric properties, and mechanical properties of molecular crystals^{3,14,29,60–65} 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 affect the relative stabilities of certain molecular crystal polymorphs and may be further included.^{3,62,66–68}

Approaches to configuration space exploration in CSP include molecular dynamics,^{69,70} Monte Carlo methods,^{25,71} particle swarm optimization,⁷² and (quasi)-random searches.^{73,74} Genetic algorithms (GAs) are a versatile class of optimization algorithms inspired by the

evolutionary principle of survival of the fittest.^{75–77} 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 genesⁱ 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 till “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, 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^{78–90} and clusters.^{75–77,91–100} Advantageously, the GA fitness function may be based on any property of interest, not necessarily the energy.^{75,101–106} 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 usⁱⁱ 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

ⁱThe 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.

ⁱⁱIn the sixth blind test we used a preliminary version of GAtor.

relies on a separate package, Genarris, reported elsewhere¹⁰⁷ 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.^{97,98,106} 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.^{108,109}

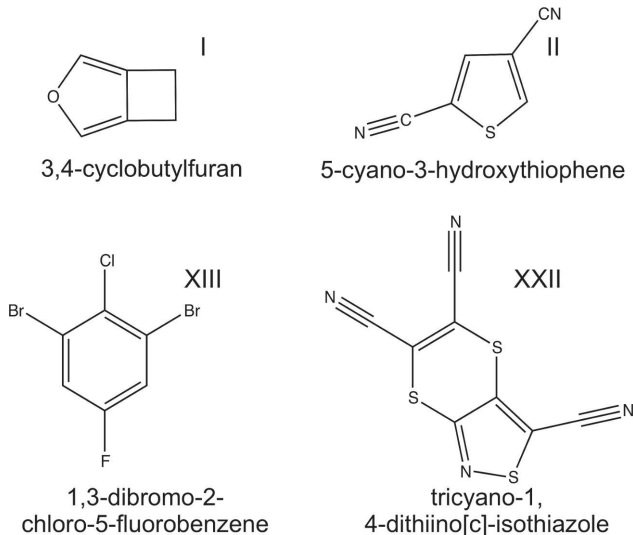


Figure 1: Two-dimensional molecular diagrams of four past blind test targets, Target I,²⁰ Target II,²⁰ Target XIII,²² and Target XXII.²⁴

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 I²⁰), 5-cyano-3-hydroxythiophene (Target II²⁰), 1,3-dibromo-2-chloro-5-fluorobenzene (Target XIII²²), and tricyano-1,4-dithiino[c]-isothiazole (Target XXII²⁴) shown in Fig. 1. Finally, Section 5 provides concluding remarks and best practices.

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. Hierarchical screening approaches have become a common practice in CSP.²⁴ In 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)^{110,111} is used with the pairwise Tkatchenko-Scheffler (TS) dispersion-correction⁵⁵ with *lower-level* numerical settings, which correspond to the light numerical settings and tier 1 basis sets of FHI-aims.¹⁰⁸ During local optimization, the space group symmetry is allowed to vary. 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 relativity¹⁰⁸ 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.¹⁰⁸ Next, single point energy (SPE) evaluations are performed using PBE with the MBD method⁵⁷⁻⁵⁹ for the best structures as ranked by PBE+TS. The final re-ranking is performed using the hybrid functional PBE0^{112,113} 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.^{61,62} For some molecular crystals the correct polymorph ranking is reproduced only when using PBE0+MBD.^{14,29} 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,⁶⁸ are not further included in the present study.

3 Code Description

GAtor is written in Python and uses the spglib¹¹⁴ crystal symmetries library, sci-kit learn¹¹⁵ machine learning package, and pymatgen¹¹⁶ library for materials analysis. GAtor is available for download from www.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,^{108,109} and may be modified to interface with other electronic structure and molecular dynamics packages.

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.^{97,98,106} Additional multiprocessing may be utilized within each replica for child generation. GAtor has been tested on up to 16,384 Blue Gene/Q nodes (262,144 cores) at the Argonne Leadership Computing Facility.

Two classes of breeding operators are implemented in GAtor, crossover and mutation, described in detail in Sections 3.4-3.5. Crossover operators generate 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)

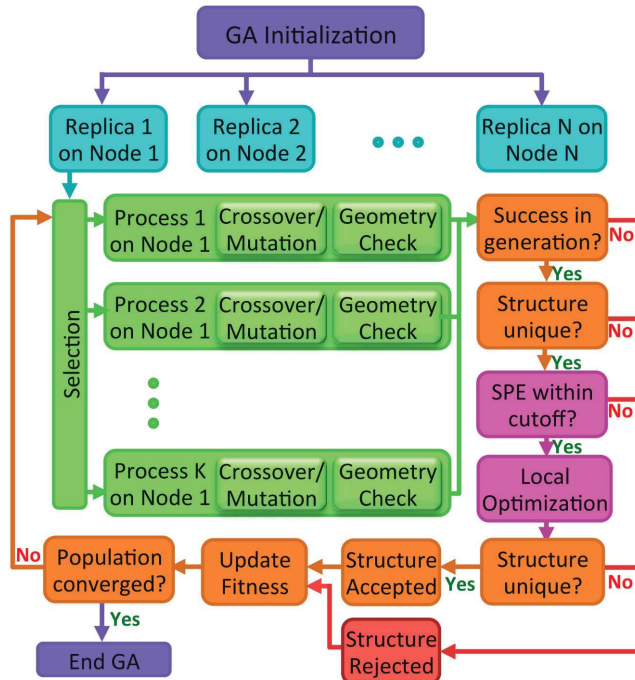


Figure 2: An example workflow of Gator on a high performance computing cluster. In the diagram, N independent GA replicas run on N computing nodes, with K core processing units per node. Single point energy (SPE) evaluations and local optimizations are performed using FHI-aims.

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 package¹⁰⁷ using the diverse workflow. Genarris generates

random symmetric crystal structures in the 230 crystallographic space groups and 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. 107. 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 the starting fitness values of the initial pool structures, as described below, before performing selection.

3.2 Fitness Evaluation

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

In energy-based fitness, the total energy E_i of the i th 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,

$$f_i = \frac{\epsilon_i}{\sum_i \epsilon_i} \quad 0 \leq f \leq 1 \quad (1)$$

$$\epsilon_i = \frac{E_{\max} - E_i}{E_{\max} - E_{\min}} \quad (2)$$

where ϵ_i is the i th 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.^{97,98,106} 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 biases towards or against certain packing motifs over time, a phenomenon known as genetic drift. Genetic drift can result from biases in the initial pool¹⁰⁷ 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.^{117–119} 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 that 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.¹¹⁵ Currently, GAtor offers the use of radial distribution function (RDF) vectors of interatomic distances for user-defined species, relative coordinate descriptor

(RCD) vectors,¹⁰⁷ or a simple lattice parameter based descriptor, L , given by:

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

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-means¹²⁴ and Affinity Propagation (AP)¹²⁵ clustering, 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 matrix, rather than defining the number of clusters *a priori*. This has the advantage of resolving small, structurally distinct clusters.¹⁰⁷ Once the common population has been clustered into niches, a fitness sharing scheme¹¹⁷ is applied such that a structure’s scaled fitness, f'_i , is given by

$$f'_i = \frac{f_i}{m_i} \quad (4)$$

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 evo-

lutionary algorithm. Refs. 126- 127 use fingerprint functions based on inter-atomic distances to prevent too dissimilar structures from mating. Recently, Ref. 99 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 technique¹²⁸ 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 $f_i > r$.^{97,98,106}

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

3.4 Crossover

Crossover is a breeding 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,^{78,79,84,85,94,96–99,106,131,132} 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.^{87,127,133,134} 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 geometryⁱⁱⁱ, 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 an inertial reference frame aligned with each molecule’s principal axes of rotation. When generating a child structure, the molecules in

ⁱⁱⁱThe 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 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 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.

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 pymatgen¹¹⁶ package, which utilizes the spglib crystal symmetries library.¹¹⁴ When generating a child structure, the genes of the parents are combined strategically to preserve one parent’s space group as detailed below.

First, the asymmetric unit and corresponding space group operations are deduced for

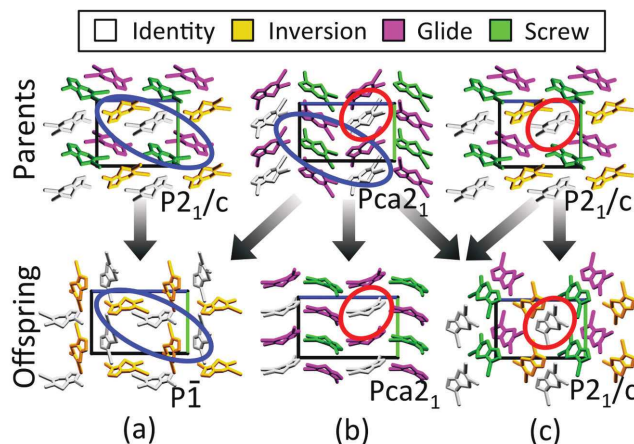


Figure 3: Examples of (a) standard crossover and (b-c) symmetric crossover applied to selected parent structures of Target XXII. The colors of the molecules correspond to the symmetry operations applied to the asymmetric unit of each structure, shown in white. The structures shown are projected along the $\vec{a}$ lattice vector and the $\vec{b}$, and $\vec{c}$ lattice vectors are highlighted in green and blue, respectively.

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, which does not constrain the symmetry of the child structure.

Examples of symmetric crossover are shown in Fig. 3, panels (b) and (c). The participating 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 lattice vectors of both parents 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 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.⁸⁵

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. 78, 87, and 85, the strain tensor is represented using the symmetric Voigt strain matrix ϵ ,

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

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

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

The components of ϵ_{ij} are chosen to produce different modes of strain. To apply 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 strain. 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 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 ($\alpha = \gamma = 90^\circ; \beta \neq 90^\circ$) to triclinic ($\alpha \neq \beta \neq \gamma \neq 90^\circ$). 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.

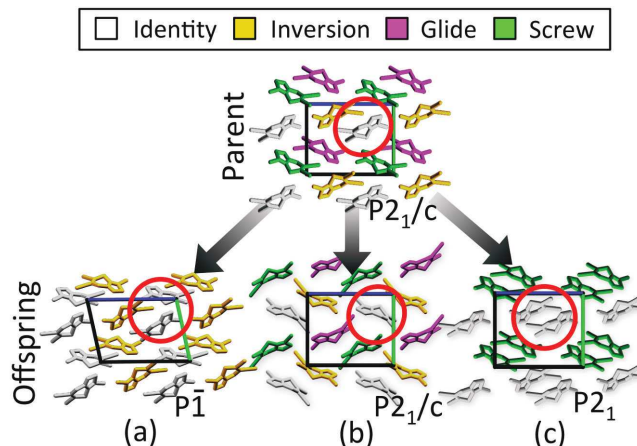


Figure 4: Examples of (a) random strain, (b) rotation, and (c) translation mutations applied to a $P2_1/c$ structure of Target XXII. The colors of the molecules correspond to the symmetry operations applied to the asymmetric unit of each structure, shown in white and circled in red. The structures shown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

3.5.2 Molecular Rotations

Rotation mutations change the orientations of the molecules in the selected parent structure. Different random rotations may 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 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 certain molecules within the unit cell. They are either applied randomly to the COM (in Cartesian coordinates) of randomly selected molecules, or in a random direction in the basis of the each molecule's inertial reference frame, constructed from each molecule's principal axes of rotation. 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 in the basis of its inertial reference frame. In this case, 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 in the parent 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.

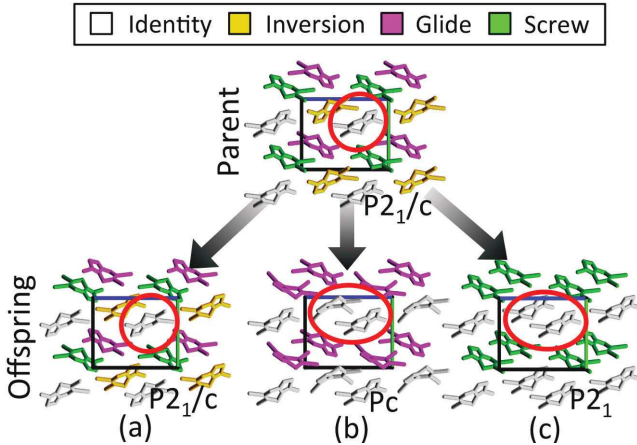


Figure 5: Examples of (a) permutation, (b) permutation-rotation, and (c) permutation-reflection mutations applied to a $P2_1/c$ structure of Target XXII. The colors of the molecules correspond to the symmetry operations applied to the asymmetric unit of each structure, shown in white and circled in red. The structures shown are projected along the $\vec{a}$ lattice vector and the $\vec{b}$, and $\vec{c}$ lattice vectors are highlighted in 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 apply a random rotation 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 crossover and 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:

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

where r_A and r_B are the vdW radii of the atoms A and B , respectively, and s_{r} is a user-defined parameter 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 reduction¹²⁰⁻¹²³ 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 parameter tolerances in pymatgen’s StructureMatcher class.¹¹⁶ 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 parameter tolerance settings. If the candi-

date 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. Here, an iteration is defined as when a structure has passed all rejection criteria and is added to the common pool. 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 Genarris¹⁰⁷ to create a starting population of diverse, high-quality structures.¹⁰⁷ The distribution of space groups for each initial pool is provided in the supporting information. 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 run on 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 ($\text{C}_8\text{S}_3\text{N}_4$) 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.¹³⁵ 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

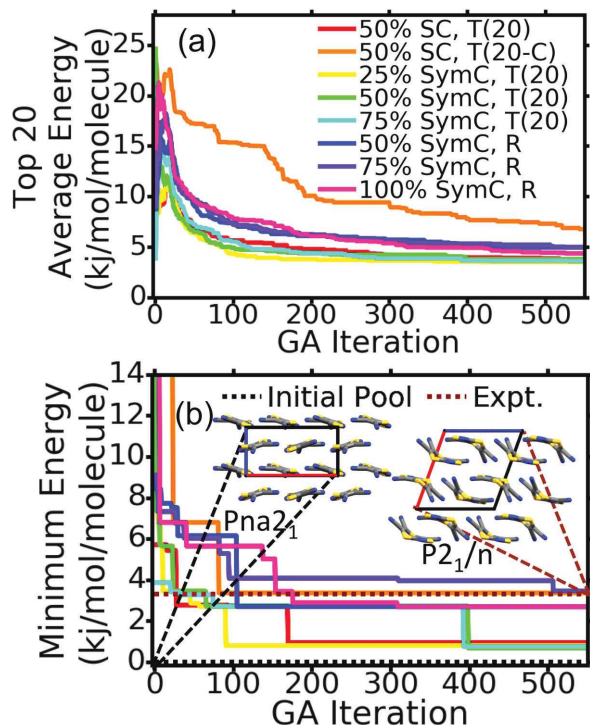


Figure 6: (a) The average energy of the top 20 Target XXII structures as a function of GA iteration and (b) the global minimum structure generated as a function of GA iteration, shown for different GA runs. S, N, and C atoms are colored in yellow, blue, and grey, respectively. The structures shown are projected along the $\vec{b}$ lattice vector and the $\vec{a}$, and $\vec{c}$ lattice vectors are highlighted in red and blue, respectively.

blind test,²⁴ and ranked as the most stable structure by 4 groups.

GATOR was run with a variety of GA settings using the same initial pool. In principle, a GA should be run numerous times to determine how a particular group of settings perform. Because GATOR is a first principles algorithm a more practical approach is adopted, where different GA settings are used in several runs and then the structures produced from all runs are combined for postprocessing. For Target XXII, the initial pool contained 100 structures in a variety of space groups. All runs were stopped when the number of structures added to the common population from the GA reached 550 structures. Fig. 6 shows an analysis of the various GA runs. Here, a GA iteration corresponds to when a single structure has passed all rejection criteria and has been added to the common population. The shorthand notation used for the different GA runs is as follows: standard crossover (SC), symmetric crossover

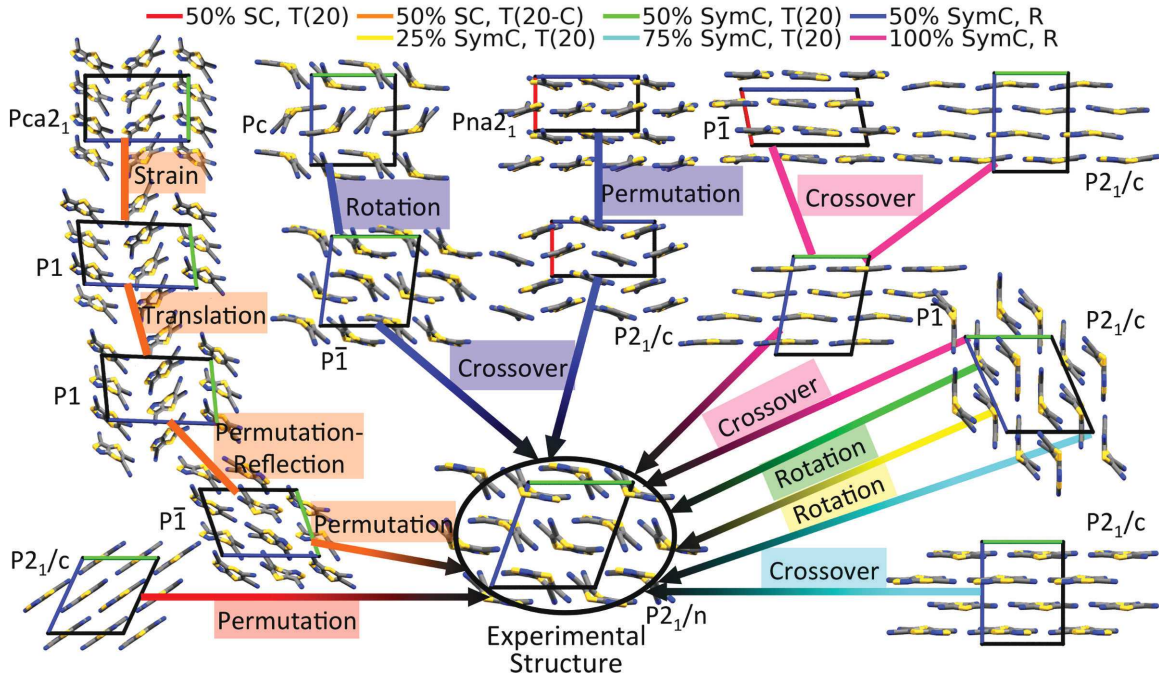


Figure 7: The different evolutionary routes which generated the experimental structure of Target XXII for different runs of the GA. The $\vec{a}$, $\vec{b}$, and $\vec{c}$ crystallographic lattice vectors are displayed in red, green, and blue, respectively.

(SymC), tournament selection (T), and roulette wheel selection (R). The percentage (e.g. 75%) indicates the crossover probability, with the remaining percentage (e.g. 25%) indicating mutation probability. For runs that used tournament selection, the tournament size is shown in parentheses. Cluster-based fitness is denoted by a C after the selection type. Here, Affinity Propagation¹²⁵ 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 seven 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 structures pre-screened with a Harris Approximation,¹³⁶ as detailed in Ref. 107. 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 selected GA runs. Each route starts from an initial pool structure and details the various breeding operations (followed by local optimization), which ultimately generate the experimental structure. The variety of evolutionary routes and paths highlights the flexibility and randomness of the GA. In particular, the run 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 generate the $P2_1/n$ experimental structure. Overall, the combination of symmetric crossover and mutation was highly effective for Target XXII.

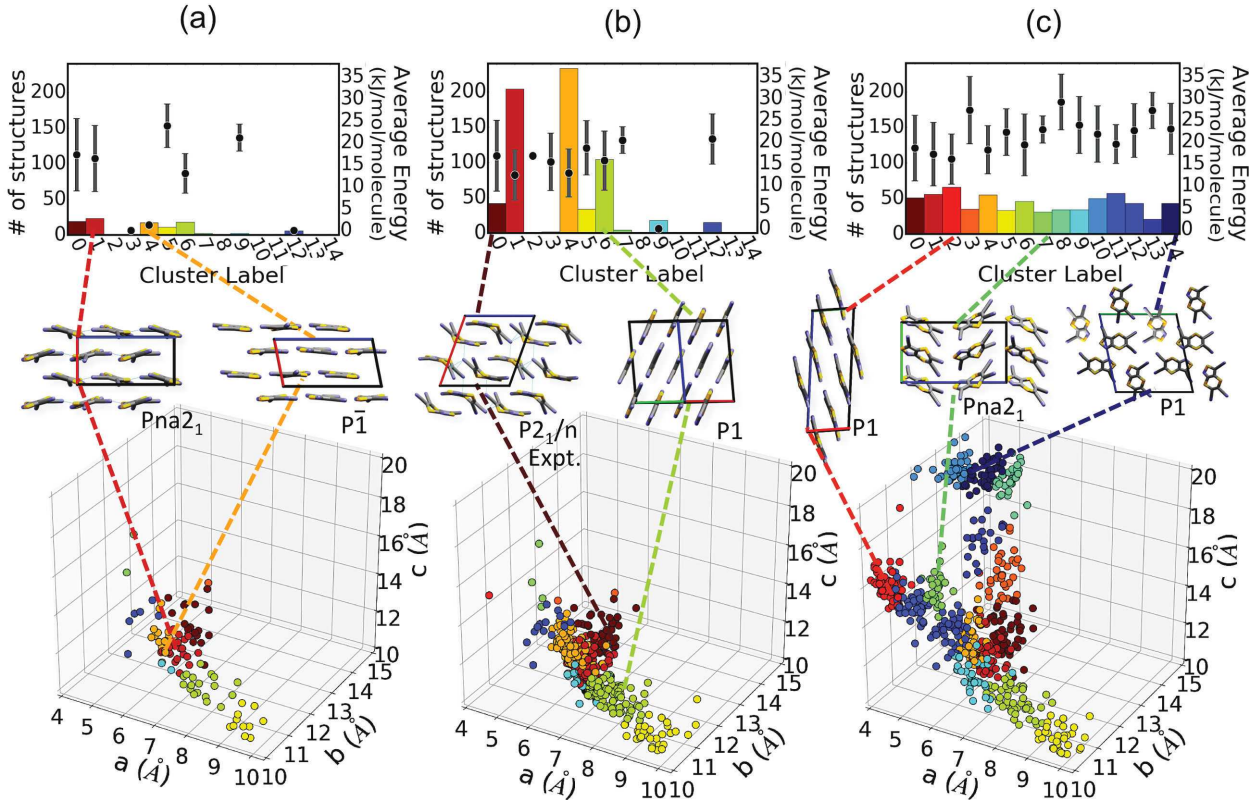


Figure 8: A comparison of the clusters and structural motifs found in (a) the initial pool, (b) the common population evolved using energy-based fitness, and (c) the common population evolved with cluster-based fitness. The average energy for each cluster is plotted using black circles and the standard deviation of energies for each cluster is depicted in grey. For the crystal structures shown, the $\vec{a}$, $\vec{b}$, and $\vec{c}$ crystallographic lattice vectors are displayed in red, green, and blue, respectively.

A detailed comparison between the runs that used tournament selection and 50% percent standard crossover, with and without cluster-based fitness, is shown in Fig. 8. The final structures produced from the cluster-based fitness run, including the initial pool, formed 15 clusters, using Affinity Propagation 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

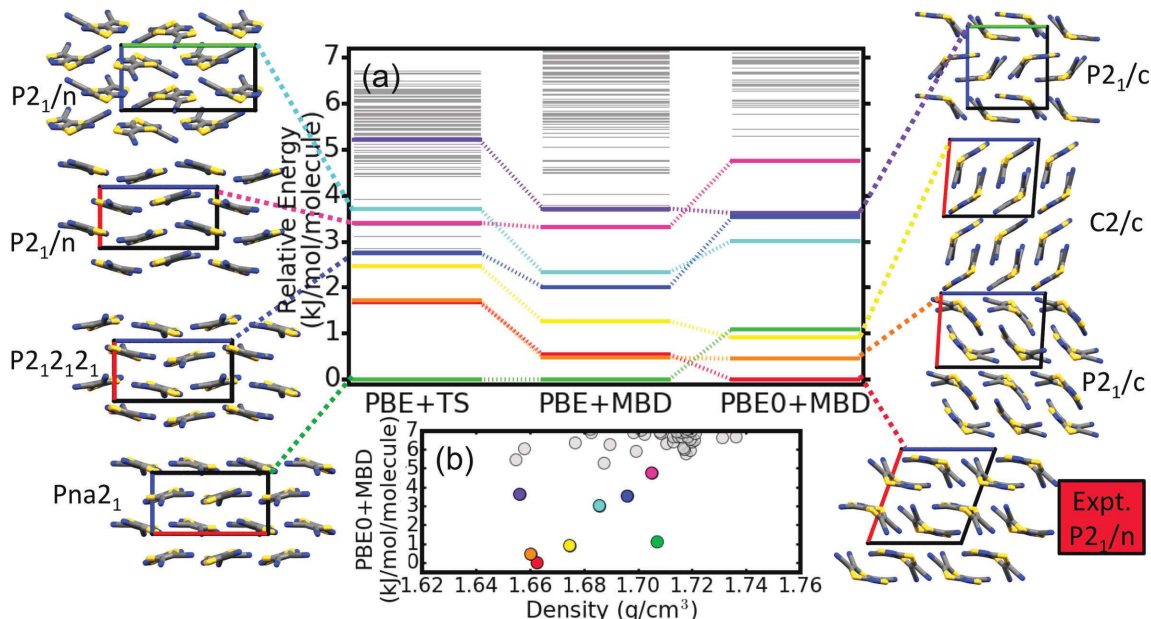


Figure 9: (a) The relative total energies as obtained by different dispersion-inclusive DFT methods and (b) the PBE0+MBD energy versus density of putative crystal structures of Target XXII. The top 8 predicted structures, as ranked by PBE0+MBD, are shown in color. The $\vec{a}$, $\vec{b}$, and $\vec{c}$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

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 evolutionary 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 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 in addition to the experimental structure. 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.¹⁰⁷ 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.^{20,137} 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 of GA iteration is shown along with the lowest-energy structure

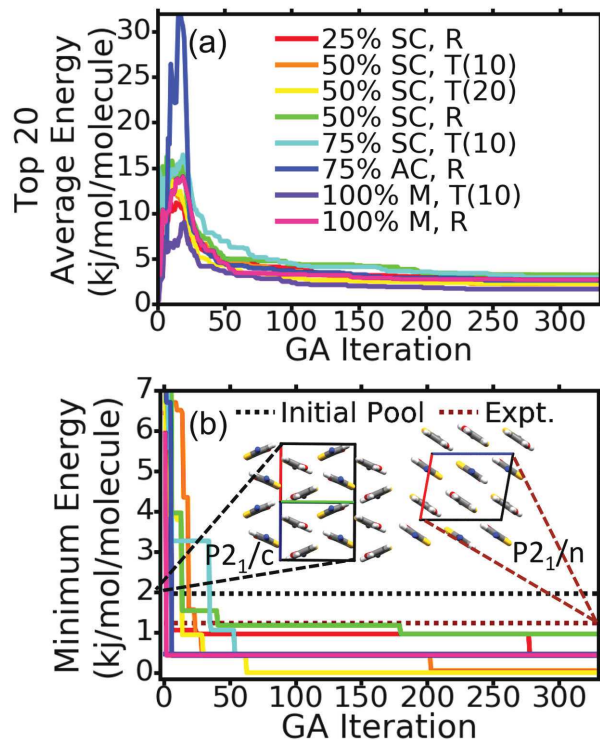


Figure 10: (a) The average energy of the top 20 Target II structures as a function of GA iteration and (b) the global minimum structure generated as a function of GA iteration, shown for different GA runs. S, N, O, C, and H atoms are colored in yellow, blue, red, grey, and white, respectively. The $\vec{a}$,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

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. Strain mutations were particularly effective at generating new low-energy structures for this target.

All structures produced by the different 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 rel-

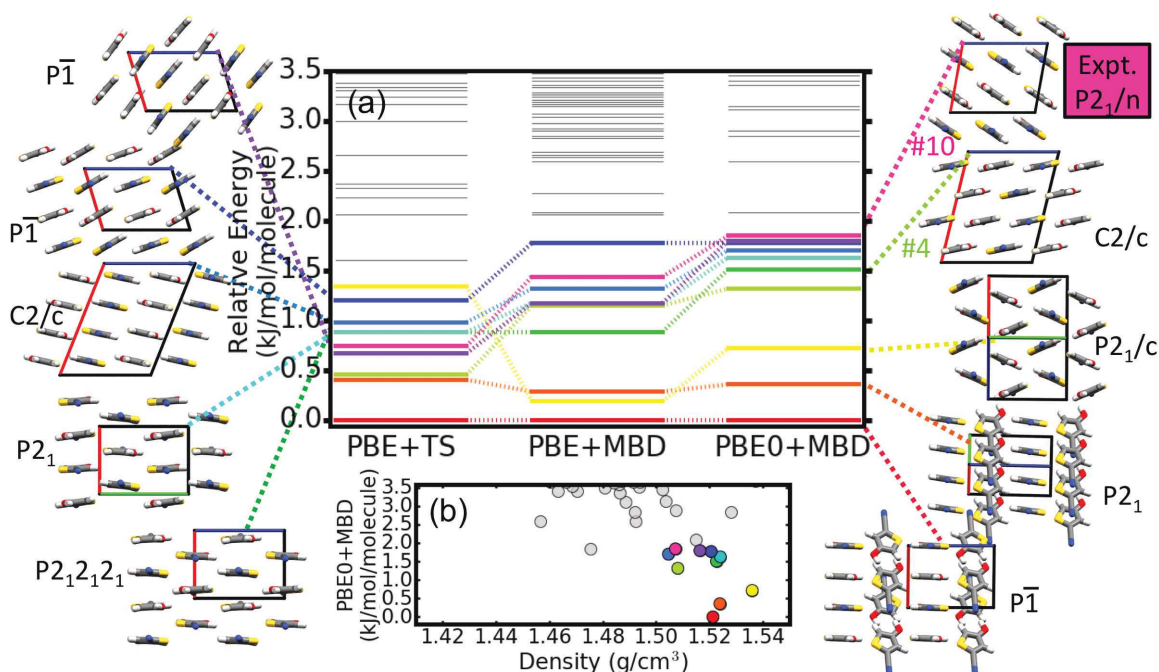


Figure 11: (a) The relative total energies as obtained by different dispersion-inclusive DFT methods and (b) the PBE0+MBD energy versus density of putative crystal structures of Target II. The top 10 predicted structures, as ranked by PBE0+MBD, are shown in color. The $\vec{a}$, $\vec{b}$, and $\vec{c}$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

ative 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 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 ($O-H \cdots N$) that are stacked on top of one another as shown in Fig. S1 in the supporting information. The prediction of nearly energetically degenerate crystal structures consisting of the same sheet stacked in different ways is a common phenomena.^{12,138,139} 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.^{12,140}

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. 141 later performed an additional CSP study on Target II using a tailor-made force field¹⁴² 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 zig-zag structure for the first time, ranking it as the global minimum structure when re-ranked using DFT with a pairwise dispersion correction.¹⁴³ 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. 141 and ranked as #3. Ref. 140 computed the relative stability of the $P2_1$ scaffold structure and the experimental structure using the B86bPBE density functional^{110,144} combined with the exchange-hole dipole moment (XDM)^{145,146} 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 reported in any previous CSP studies of Target II. It has a higher computed density than the experimental form, as shown in panel (b) of Fig. 11. A discussion comparing the packing motifs of the experimental structure and the #1 PBE0+MBD scaffold structure is provided in the supporting information. The #1 PBE0+MBD scaffold structure would not have been found without Gator’s ability to generate crystal structures with $Z' > 1$ through the various crossover and mutation operators. As emphasized in Ref. 147, 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.^{147–149} 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.

4.3 Target XIII

Target XIII ($\text{C}_6\text{H}_2\text{Br}_2\text{ClF}$) 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.^{150–154} 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. 155.

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. 12. The initial pool for all runs contained 48 structures. The various GA 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, generated 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, consistently 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.

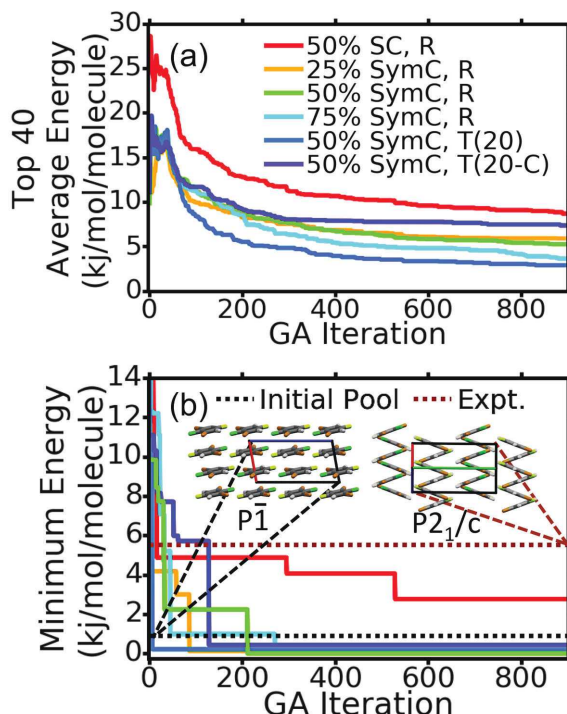


Figure 12: (a) The average energy of the top 40 Target XIII structures and (b) the global minimum structure produced as a function of GA iteration for different GA runs. C, H, Br, Cl, and F atoms are colored in grey, white, brown, green, and yellow, respectively. The $\vec{a}$, $\vec{b}$, and $\vec{c}$ crystallographic lattice vectors are displayed in red, green, and blue, respectively.

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. 13 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 local optimization with PBE+TS and *higher-level* numerical settings, the experimental structure is ranked as #1. It is consistently predicted as the most stable crystal structure by PBE+MBD

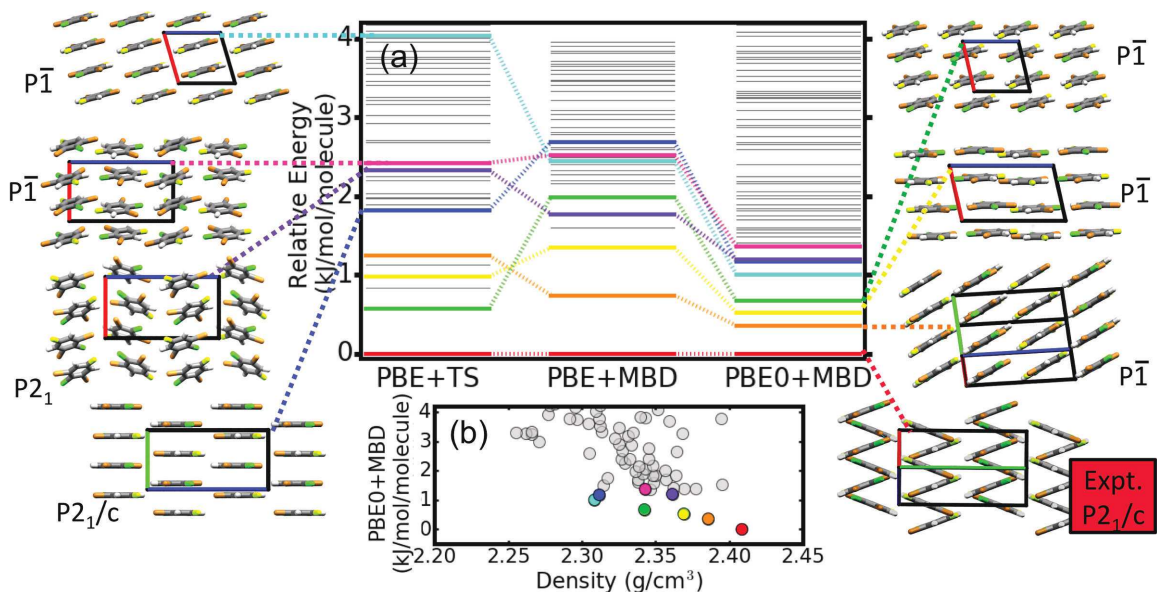


Figure 13: (a) The relative total energies as obtained by different dispersion-inclusive DFT methods and (b) the PBE0+MBD energy versus density of putative crystal structures of Target XIII. The top 8 predicted structures, as ranked by PBE0+MBD, are shown in color. The $\vec{a}$, $\vec{b}$, and $\vec{c}$ crystallographic lattice vectors are displayed in red, green, and blue, respectively.

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. 13 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 gener-

ated 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 ($\text{C}_6\text{H}_6\text{O}$) was selected from the second blind test.^{20,137} 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. 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

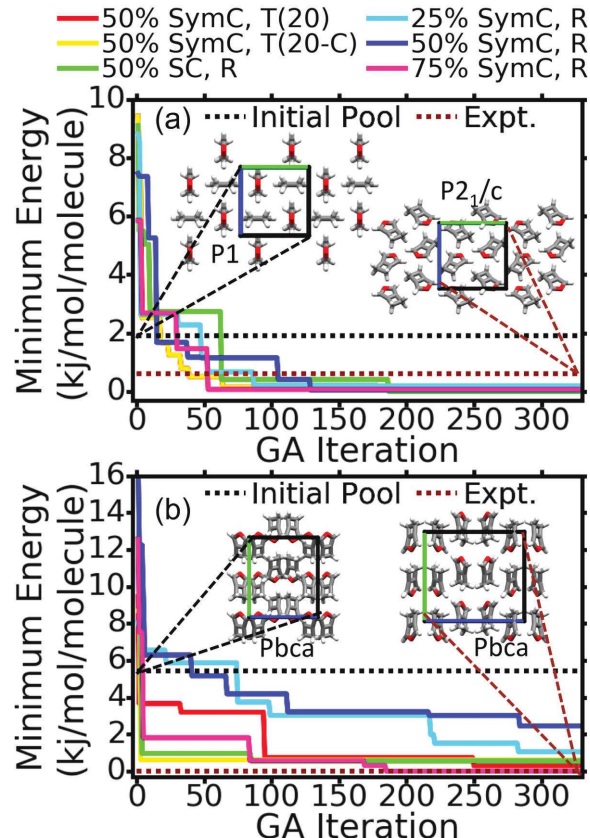


Figure 14: The global minimum structure produced by the GA runs as a function of GA iteration for runs that used (a) $Z=4$ and (b) $Z=8$. C, H, and O atoms are colored in grey, white, and red, respectively. The structures shown are projected along the $\vec{a}$ lattice vector and the $\vec{b}$, and $\vec{c}$ lattice vectors are highlighted in green and blue, respectively.

PBE+TS with *lower-level* numerical settings, is shown in Fig. 14, 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

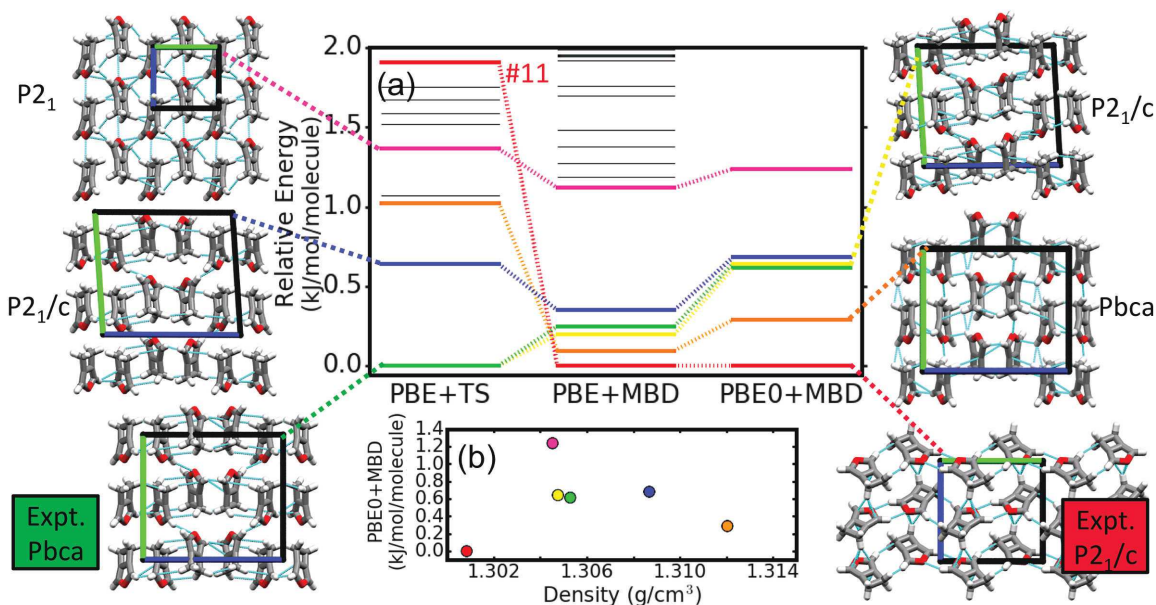


Figure 15: (a) The relative total energies as obtained by different dispersion-inclusive DFT methods and (b) the PBE0+MBD energy versus density of selected crystal structures of Target I. The top 6 predicted structures, as ranked by PBE+MBD, are highlighted in color. Intermolecular contacts less than the sum of vdW radii are shown in cyan. The structures shown are projected along the $\vec{a}$ lattice vector and the $\vec{b}$, and $\vec{c}$ lattice vectors are highlighted in green and blue, respectively.

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. 15. 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, shown in green, is ranked as #1 with PBE+TS, #4 with PBE+MBD, and #3 when re-ranked with PBE0+MBD. With all energy methods 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 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. 15 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^{156–158} 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. 68 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.^{159,160} 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. 140 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 settings, including energy-based and cluster-based fitness (evolutionary niching), roulette wheel and tournament selection, symmetric and standard crossover, different mu-

tation 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 generating important low-energy $Z=4$ structures. 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 structures in all space groups appropriate for the molecule. Such an initial pool may be generated by Genarris¹⁰⁷ 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 may 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 more 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 Mercury¹⁶¹ software. The distribution of space groups for the different initial pools used is also included. For each target, CIF files and total energies of the structures used

for re-ranking are provided. 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: Oxford, England, 2002; Vol. 14.
- (2) Day, G. M.; Motherwell, W. D. S.; 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. *J. Pharm. Pharmacol.* **2015**, *67*, 757–772.
- (5) Reese, C.; Bao, Z. Organic Single-Crystal Field-Effect Transistors. *Mater. Today* **2007**, *10*, 20–27.
- (6) Hasegawa, T.; Takeya, J. Organic Field-Effect Transistors Using Single Crystals. *Sci. Technol. Adv. Mater.* **2009**, *10*, 024314.
- (7) Bergantin, S.; Moret, M. Rubrene Polymorphs and Derivatives: The Effect of Chemical Modification on the Crystal Structure. *Cryst. Growth Des.* **2012**, *12*, 6035–6041.
- (8) Cudazzo, P.; Gatti, M.; Rubio, A. Excitons in Molecular Crystals from First-Principles Many-Body Perturbation Theory: 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 Crystallogr., Sect. B: Struct. Sci., Cryst. Eng. Mater.* **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., Sect. 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. *J. Am. Chem. Soc.* **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. 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. A Test of Crystal Structure Prediction of Small Organic Molecules. *Acta Crystallogr., Sect. B: Struct. Sci.* **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 Crystallogr., Sect. B: Struct. Sci.* **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 Crystallogr., Sect. B: Struct. Sci.* **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 Crystallogr., Sect. B: Struct. Sci.* **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 Crystallogr., Sect. B: Struct. Sci.* **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, R. A., Jr; Dzyabchenko, A.; van Eijck, B. P.; Elking, D. M.; van den Ende, J. A.; Facelli, J. C.; Ferraro, M. B.; Fusti-Molnar, L.; Gatsiou, C. A.; Gee, T. S.; de Gelder, R.; Ghiringhelli, L. M.; Goto, H.; Grimme, S.; Guo, R.; Hofmann, D. W. M.; Hoja, J.; Hylton, R. K.; Iuzzolino, L.; Jankiewicz, W.; de Jong, D. T.; Kendrick, J.; de Klerk, N. J. J.; Ko, H. Y.; Kuleshova, L. N.; Li, X.; Lohani, S.;

- Leusen, F. J. J.; Lund, A. M.; Lv, J.; Ma, Y.; Marom, N.; Masunov, A. E.; McCabe, P.; McMahon, D. P.; Meekes, H.; Metz, M. P.; Misquitta, A. J.; Mohamed, S.; Monserrat, B.; Needs, R. J.; Neumann, M. A.; Nyman, J.; Obata, S.; Oberhofer, H.; Oganov, A. R.; Orendt, A. M.; Pagola, G. I.; Pantelides, C. C.; Pickard, C. J.; Podeszwa, R.; Price, L. S.; Price, S. L.; Pulido, A.; Read, M. G.; Reuter, K.; Schneider, E.; Schober, C.; Shields, G. P.; Singh, P.; Sugden, I. J.; Szalewicz, K.; Taylor, C. R.; Tkatchenko, A.; Tuckerman, M. E.; Vacarro, F.; Vasileiadis, M.; Vazquez-Mayagoitia, A.; Vogt, L.; Wang, Y.; Watson, R. E.; de Wijs, G. A.; Yang, J.; Zhu, Q.; Groom, C. R. Report on the Sixth Blind Test of Organic Crystal Structure Prediction Methods. *Acta Crystallogr., Sect. B: Struct. Sci., Cryst. Eng. Mater.* **2016**, *72*, 439–59.
- (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. *Nat. Commun.* **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.; Tuckerman, M. E.; Kahral, B. Powder Diffraction and Crystal Structure Prediction Identify Four New Coumarin Polymorphs. *Chem. Sci.* **2017**, *8*, 4926–4940.
- (27) Meredig, B.; Wolverton, C. A Hybrid Computational–Experimental Approach for Automated Crystal Structure Solution. *Nat. Mater.* **2013**, *12*, 123–127.
- (28) Gavezzotti, A. Are Crystal Structures Predictable? *Acc. Chem. Res.* **1994**, *27*, 309–314.
- (29) 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.

- (30) Cruz-Cabeza, A. J.; Reutzel-Edens, S. M.; Bernstein, J. Facts and Fictions About Polymorphism. *Chem. Soc. Rev.* **2015**, *44*, 8619–8635.
- (31) Beran, G. J. A New Era for Ab Initio Molecular Crystal Lattice Energy Prediction. *Angew. Chem. Int. Ed.* **2015**, *54*, 396–398.
- (32) Beran, G. J. Modeling Polymorphic Molecular Crystals with Electronic Structure Theory. *Chem. Rev.* **2016**, *116*, 5567–5613.
- (33) Dion, M.; Rydberg, H.; Schröder, E.; Langreth, D. C.; Lundqvist, B. I. Van der Waals Density Functional for General Geometries. *Phys. Rev. Lett.* **2004**, *92*, 246401.
- (34) Lee, K.; Murray, É. D.; Kong, L.; Lundqvist, B. I.; Langreth, D. C. Higher-Accuracy van der Waals Density Functional. *Phys. Rev. B* **2010**, *82*, 081101.
- (35) Vydrov, O. A.; Van Voorhis, T. Nonlocal van der Waals Density Functional Made Simple. *Phys. Rev. Lett.* **2009**, *103*, 063004.
- (36) Peverati, R.; Truhlar, D. G. M11-L: A Local Density Functional That Provides Improved Accuracy for Electronic Structure Calculations in Chemistry and Physics. *J. Phys. Chem. Lett.* **2011**, *3*, 117–124.
- (37) 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. *Phys. Chem. Chem. Phys.* **2012**, *14*, 13171–13174.
- (38) 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. *Theor. Chem. Acc.* **2008**, *120*, 215–241.

- (39) Vydrov, O. A.; Van Voorhis, T. Nonlocal van der Waals Density Functional: The Simpler the Better. *J. Chem. Phys.* **2010**, *133*, 244103.
- (40) Vydrov, O. A.; Van Voorhis, T. Dispersion Interactions from a Local Polarizability Model. *Phys. Rev. A* **2010**, *81*, 062708.
- (41) 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. *Rep. Prog. Phys.* **2015**, *78*, 066501.
- (42) Thonhauser, T.; Zuluaga, S.; Arter, C.; Berland, K.; Schröder, E.; Hyldgaard, P. Spin Signature of Nonlocal Correlation Binding in Metal-Organic Frameworks. *Phys. Rev. Lett.* **2015**, *115*, 136402.
- (43) Peng, H.; Yang, Z.-H.; Perdew, J. P.; Sun, J. Versatile van der Waals Density Functional Based on a Meta-Generalized Gradient Approximation. *Phys. Rev. X* **2016**, *6*, 041005.
- (44) Sun, J.; Ruzsinszky, A.; Perdew, J. P. Strongly Constrained and Appropriately Normed Semilocal Density Functional. *Phys. Rev. Lett.* **2015**, *115*, 036402.
- (45) 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. *Chem. Rev.* **2010**, *110*, 5023–5063.
- (46) Grimme, S. Semiempirical GGA-Type Density Functional Constructed with a Long-Range Dispersion Correction. *J. Comput. Chem.* **2006**, *27*, 1787–1799.
- (47) 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. *J. Chem. Phys.* **2010**, *132*, 154104.

- (48) Johnson, E. R.; Becke, A. D. A Post-Hartree–Fock Model of Intermolecular Interactions. *J. Chem. Phys.* **2005**, *123*, 024101.
- (49) Otero-de-la-Roza, A.; Johnson, E. R. A Benchmark for Non-Covalent Interactions in Solids. *J. Chem. Phys.* **2012**, *137*, 054103.
- (50) Jurečka, P.; Černý, 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. *J. Comput. Chem.* **2007**, *28*, 555–569.
- (51) Wu, Q.; Yang, W. Empirical Correction to Density Functional Theory for van der Waals Interactions. *J. Chem. Phys.* **2002**, *116*, 515–524.
- (52) Wu, X.; Vargas, M.; Nayak, S.; Lotrich, V.; Scoles, G. Towards Extending the Applicability of Density Functional Theory to Weakly Bound Systems. *J. Chem. Phys.* **2001**, *115*, 8748–8757.
- (53) Steinmann, S. N.; Corminboeuf, C. A Generalized-Gradient Approximation Exchange Hole Model for Dispersion Coefficients. *J. Chem. Phys.* **2011**, *134*, 044117.
- (54) Steinmann, S. N.; Corminboeuf, C. Comprehensive Benchmarking of a Density-Dependent Dispersion Correction. *J. Chem. Theory Comput.* **2011**, *7*, 3567–3577.
- (55) 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.
- (56) Brandenburg, J.; Bates, J.; Sun, J.; Perdew, J. Benchmark Tests of a Strongly Constrained Semilocal Functional with a Long-Range Dispersion Correction. *Phys. Rev. B* **2016**, *94*, 115144.

- (57) DiStasio, R. A.; von Lilienfeld, O. A.; Tkatchenko, A. Collective Many-Body van der Waals Interactions in Molecular Systems. *Proc. Natl. Acad. Sci. U. S. A.* **2012**, *109*, 14791–14795.
- (58) 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.
- (59) 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.
- (60) Schatschneider, B.; Liang, J.-J.; Reilly, A. M.; Marom, N.; Zhang, G.-X.; Tkatchenko, A. Electrodynamic Response and Stability of Molecular Crystals. *Phys. Rev. B* **2013**, *87*, 060104.
- (61) 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.
- (62) Reilly, A. M.; Tkatchenko, A. Seamless and Accurate Modeling of Organic Molecular Materials. *J. Phys. Chem. Lett.* **2013**, *4*, 1028–1033.
- (63) Tkatchenko, A. Current Understanding of van der Waals Effects in Realistic Materials. *Adv. Funct. Mater.* **2015**, *25*, 2054–2061.
- (64) Hermann, J.; DiStasio Jr, R. A.; Tkatchenko, A. First-Principles Models for van der Waals Interactions in Molecules and Materials: Concepts, Theory, and Applications. *Chem. Rev.* **2017**, *117*, 4714–4758.
- (65) Flores-Huerta, A. G.; Tkatchenko, A.; Galván, M. Nature of Hydrogen Bonds and S...S Interactions in the L-Cystine Crystal. *J. Phys. Chem. A* **2016**, *120*, 4223–4230.

- (66) Hoja, J.; Reilly, A. M.; Tkatchenko, A. First-Principles Modeling of Molecular Crystals: Structures and Stabilities, Temperature and Pressure. *WIREs Comput. Mol. Sci.* **2017**, *7*, e1294.
- (67) Rossi, M.; Gasparotto, P.; Ceriotti, M. Anharmonic and Quantum Fluctuations in Molecular Crystals: A First-Principles Study of the Stability of Paracetamol. *Phys. Rev. Lett.* **2016**, *117*, 115702.
- (68) Nyman, J.; Day, G. M. Static and Lattice Vibrational Energy Differences Between Polymorphs. *CrystEngComm* **2015**, *17*, 5154–5165.
- (69) Yu, T. Q.; Tuckerman, M. E. Temperature-Accelerated Method for Exploring Polymorphism in Molecular Crystals Based on Free Energy. *Phys. Rev. Lett.* **2011**, *107*, 015701.
- (70) Schneider, E.; Vogt, L.; Tuckerman, M. E. Exploring Polymorphism of Benzene and Naphthalene with Free Energy Based Enhanced Molecular Dynamics. *Acta Crystallogr., Sect. B: Struct. Sci., Cryst. Eng. Mater.* **2016**, *72*, 542–550.
- (71) Akkermans, R. L.; Spenley, N. A.; Robertson, S. H. Monte Carlo Methods in Materials Studio. *Mol. Simul.* **2013**, *39*, 1153–1164.
- (72) Wang, Y.; Lv, J.; Zhu, L.; Ma, Y. CALYPSO: A Method for Crystal Structure Prediction. *Comput. Phys. Commun.* **2012**, *183*, 2063–2070.
- (73) Pickard, C. J.; Needs, R. Ab Initio Random Structure Searching. *J. Phys.: Condens. Matter* **2011**, *23*, 053201.
- (74) Case, D. H.; Campbell, J. E.; Bygrave, P. J.; Day, G. M. Convergence Properties of Crystal Structure Prediction by Quasi-Random Sampling. *J. Chem. Theory Comput.* **2016**, *12*, 910–924.

- (75) Johnston, R. L. Evolving Better Nanoparticles: Genetic Algorithms for Optimising Cluster Geometries. *Dalton Trans.* **2003**, 4193–4207.
- (76) Sierka, M. Synergy Between Theory and Experiment in Structure Resolution of Low-Dimensional Oxides. *Prog. Surf. Sci.* **2010**, *85*, 398–434.
- (77) Heiles, S.; Johnston, R. L. Global Optimization of Clusters Using Electronic Structure Methods. *Int. J. Quantum Chem.* **2013**, *113*, 2091–2109.
- (78) Oganov, A. R.; Glass, C. W. Crystal Structure Prediction Using Ab Initio Evolutionary Techniques: Principles and Applications. *J. Chem. Phys.* **2006**, *124*, 244704.
- (79) Glass, C. W.; Oganov, A. R.; Hansen, N. USPEX—Evolutionary Crystal Structure Prediction. *Comput. Phys. Commun.* **2006**, *175*, 713–720.
- (80) 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.
- (81) Trimarchi, G.; Zunger, A. Global Space-Group Optimization Problem: Finding the Stablest Crystal Structure Without Constraints. *Phys. Rev. B* **2007**, *75*, 104113.
- (82) Wu, S.; Umemoto, K.; Ji, M.; Wang, C. Z.; Ho, K. M.; Wentzcovitch, R. M. Identification of Post-Pyrite Phase Transitions in SiO₂ by a Genetic Algorithm. *Phys. Rev. B* **2011**, *83*, 184102.
- (83) Woodley, S.; Battle, P.; Gale, J.; Catlow, C. A. The Prediction of Inorganic Crystal Structures Using a Genetic Algorithm and Energy Minimisation. *Phys. Chem. Chem. Phys.* **1999**, *1*, 2535–2542.
- (84) Trimarchi, G.; Zunger, A. Finding the Lowest-Energy Crystal Structure Starting from Randomly Selected Lattice Vectors and Atomic Positions: First-Principles Evolution-

- ary Study of the Au–Pd, Cd–Pt, Al–Sc, Cu–Pd, Pd–Ti, and Ir–N Binary Systems. *J. Phys.: Condens. Matter* **2008**, *20*, 295212.
- (85) Lonie, D. C.; Zurek, E. XtalOpt: An Open-Source Evolutionary Algorithm for Crystal Structure Prediction. *Comput. Phys. Commun.* **2011**, *182*, 372–387.
- (86) 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. *Phys. Rev. Lett.* **2002**, *88*, 255506.
- (87) 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., Sect. B: Struct. Sci.* **2012**, *68*, 215–26.
- (88) 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. *Chem. Phys. Lett.* **2015**, *626*, 20–24.
- (89) Avery, P.; Falls, Z.; Zurek, E. XtalOpt Version R10: An Open–Source Evolutionary Algorithm for Crystal Structure Prediction. *Comput. Phys. Commun.* **2017**, *217*, 210–211.
- (90) Falls, Z.; Lonie, D. C.; Avery, P.; Shamp, A.; Zurek, E. XtalOpt Version R9: An Open-Source Evolutionary Algorithm for Crystal Structure Prediction. *Comput. Phys. Commun.* **2016**, *199*, 178–179.
- (91) Morris, J.; Deaven, D.; Ho, K.; Wang, C.; Pan, B.; Wacker, J.; Turner, D. Genetic Algorithm Optimization of Atomic Clusters. *IMA Vol. Math. Its Appl.* **1999**, *111*, 167–176.
- (92) 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. *J. Chem. Theory Comput.* **2005**, *1*, 566–580.
- (93) Marques, J. M. C.; Pereira, F. B. An Evolutionary Algorithm for Global Minimum Search of Binary Atomic Clusters. *Chem. Phys. Lett.* **2010**, *485*, 211–216.
- (94) Hartke, B. Global Cluster Geometry Optimization by a Phenotype Algorithm with Niches: Location of Elusive Minima, and Low-Order Scaling with Cluster Size. *J. Comput. Chem.* **1999**, *20*, 1752–1759.
- (95) Catlow, C. R. A.; Bromley, S. T.; Hamad, S.; Mora-Fonz, M.; Sokol, A. A.; Woodley, S. M. Modelling Nano-Clusters and Nucleation. *Phys. Chem. Chem. Phys.* **2010**, *12*, 786–811.
- (96) 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. *Phys. Rev. A* **2004**, *69*, 053202.
- (97) 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. *Phys. Rev. Lett.* **2013**, *111*, 135501.
- (98) 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 J. Phys.* **2014**, *16*, 123016.
- (99) Jørgensen, M. S.; Groves, M. N.; Hammer, B. Combining Evolutionary Algorithms with Clustering Toward Rational Global Structure Optimization at the Atomic Scale. *J. Chem. Theory Comput.* **2017**, *13*, 1486–1493.
- (100) Tipton, W. W.; Hennig, R. G. A Grand Canonical Genetic Algorithm for the Pre-

- diction of Multi-Component Phase Diagrams and Testing of Empirical Potentials. *J. Phys.: Condens. Matter* **2013**, *25*, 495401.
- (101) O’Boyle, N. M.; Campbell, C. M.; Hutchison, G. R. Computational Design and Selection of Optimal Organic Photovoltaic Materials. *J. Phys. Chem. C* **2011**, *115*, 16200–16210.
- (102) Jain, A.; Castelli, I. E.; Hautier, G.; Bailey, D. H.; Jacobsen, K. W. Performance of Genetic Algorithms in Search for Water Splitting Perovskites. *J. Mater. Sci.* **2013**, *48*, 6519–6534.
- (103) 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. *Phys. Rev. Lett.* **2012**, *108*, 027401.
- (104) Zhang, L.; Luo, J. W.; Saraiva, A.; Koiller, B.; Zunger, A. Genetic Design of Enhanced Valley Splitting Towards a Spin Qubit in Silicon. *Nat. Commun.* **2013**, *4*, 2396.
- (105) 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. *Nat. Mater.* **2010**, *9*, 418–422.
- (106) 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. *Phys. Rev. B* **2015**, *91*, 241115.
- (107) 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. *J. Chem. Phys.* In Press.
- (108) Blum, V.; Gehrke, R.; Hanke, F.; Havu, P.; Havu, V.; Ren, X.; Reuter, K.; Scheff-

- fler, M. Ab Initio Molecular Simulations with Numeric Atom-Centered Orbitals. *Comput. Phys. Commun.* **2009**, *180*, 2175–2196.
- (109) Havu, V.; Blum, V.; Havu, P.; Scheffler, M. Efficient O (N) Integration for All-Electron Electronic Structure Calculation Using Numeric Basis Functions. *J. Comput. Phys.* **2009**, *228*, 8367–8379.
- (110) Perdew, J. P.; Burke, K.; Ernzerhof, M. Generalized Gradient Approximation Made Simple. *Phys. Rev. Lett.* **1996**, *77*, 3865–3868.
- (111) Perdew, J. P.; Burke, K.; Ernzerhof, M. Generalized Gradient Approximation Made Simple. *Phys. Rev. Lett.* **1997**, *78*, 1396–1396.
- (112) Perdew, J. P.; Ernzerhof, M.; Burke, K. Rationale for Mixing Exact Exchange with Density Functional Approximations. *J. Chem. Phys.* **1996**, *105*, 9982–9985.
- (113) Adamo, C.; Barone, V. Toward Reliable Density Functional Methods Without Adjustable Parameters: The PBE0 Model. *J. Chem. Phys.* **1999**, *110*, 6158–6170.
- (114) Togo, A. Spglib. <https://atztogo.github.io/spglib/>.
- (115) 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. *J. Mach. Learn. Res.* **2011**, *12*, 2825–2830.
- (116) 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. *Comput. Mater. Sci.* **2013**, *68*, 314–319.
- (117) Sareni, B.; Krahenbuhl, L. Fitness Sharing and Niching Methods Revisited. *IEEE Trans. Evol. Comput.* **1998**, *2*, 97–106.

- (118) Shir, O. M. *Handbook of Natural Computing*; Springer: Berlin, Heidelberg, Germany, 2012; pp 1035–1069.
- (119) Preuss, M. *Multimodal Optimization by Means of Evolutionary Algorithms*; Springer: New York, NY, 2015.
- (120) Niggli, P. *Krystallographische Und Strukturtheoretische Grundbegriffe*; Akademische Verlagsgesellschaft: Leipzig, Germany, 1928.
- (121) Gruber, B. The Relationship Between Reduced Cells in a General Bravais Lattice. *Acta Crystallogr., Sect. A: Cryst. Phys., Diffr., Theor. Gen. Crystallogr.* **1973**, *29*, 433–440.
- (122) Křivý, I.; Gruber, B. A Unified Algorithm for Determining the Reduced (Niggli) Cell. *Acta Crystallogr., Sect. A: Cryst. Phys., Diffr., Theor. Gen. Crystallogr.* **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 Crystallogr., Sect. A: Found. Crystallogr.* **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 Trans. Pattern Anal. Mach. Intell.* **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. *Comput. Phys. Commun.* **2010**, *181*, 1623–1632.

- (127) Lyakhov, A. O.; Oganov, A. R.; Stokes, H. T.; Zhu, Q. New Developments in Evolutionary Structure Prediction Algorithm USPEX. *Comput. Phys. Commun.* **2013**, *184*, 1172–1182.
- (128) Goldberg, D. E. *Genetic Algorithms in Search, Optimization and Machine Learning*; Addison-Wesley: Reading, MA, 1989.
- (129) Blickle, T.; Thiele, L. A Comparison of Selection Schemes Used in Evolutionary Algorithms. *Evol. Comput.* **1996**, *4*, 361–394.
- (130) Deaven, D. M.; Ho, K. M. Molecular Geometry Optimization with a Genetic Algorithm. *Phys. Rev. Lett.* **1995**, *75*, 288.
- (131) Frolov, V. A.; Reuter, K. Robustness of cut and Splice Genetic Algorithms in the Structural Optimization of Atomic Clusters. *Chem. Phys. Lett.* **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. *Phys. Chem. Chem. Phys.* **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) Avery, P.; Zurek, E. RandSpg: An Open-Source Program for Generating Atomistic Crystal Structures with Specific Spacegroups. *Comput. Phys. Commun.* **2017**, *213*, 208–216.
- (135) 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.
- (136) Harris, J. Simplified Method for Calculating the Energy of Weakly Interacting Fragments. *Phys. Rev. B* **1985**, *31*, 1770.

- (137) 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 Crystallogr., Sect. B: Struct. Sci.* **1999**, *55*, 963–974.
- (138) Braun, D. E.; Tocher, D. A.; Price, S. L.; Griesser, U. J. The Complexity of Hydration of Phloroglucinol: A Comprehensive Structural and Thermodynamic Characterization. *J. Phys. Chem. B* **2012**, *116*, 3961–3972.
- (139) 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. *Mol. Pharmaceutics* **2016**, *13*, 1012–1029.
- (140) 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. *J. Chem. Theory Comput.* **2017**, *13*, 441–450.
- (141) Chan, H. S.; Kendrick, J.; Leusen, F. J. J. Predictability of the Polymorphs of Small Organic Compounds: Crystal Structure Predictions of Four Benchmark Blind Test Molecules. *Phys. Chem. Chem. Phys.* **2011**, *13*, 20361–20370.
- (142) Neumann, M. A. Tailor-Made Force Fields for Crystal-Structure Prediction. *J. Phys. Chem. B* **2008**, *112*, 9810–9829.
- (143) Grimme, S. Density Functional Theory with London Dispersion Corrections. *WIREs Comput. Mol. Sci.* **2011**, *1*, 211–228.
- (144) Becke, A. On the Large-Gradient Behavior of the Density Functional Exchange Energy. *J. Chem. Phys.* **1986**, *85*, 7184–7187.
- (145) Becke, A. D.; Johnson, E. R. Exchange-Hole Dipole Moment and the Dispersion Interaction Revisited. *J. Chem. Phys.* **2007**, *127*, 154108.

- (146) Otero-de-la-Roza, A.; Johnson, E. R. Van der Waals Interactions in Solids Using the Exchange-Hole Dipole Moment Model. *J. Chem. Phys.* **2012**, *136*, 174109.
- (147) Steed, K. M.; Steed, J. W. Packing Problems: High Z Crystal Structures and Their Relationship to Cocrystals, Inclusion Compounds, and Polymorphism. *Chem. Rev.* **2015**, *115*, 2895–2933.
- (148) 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 Crystallogr., Sect. B: Struct. Sci.* **2010**, *66*, 472–481.
- (149) Taylor, R.; Cole, J. C.; Groom, C. R. Molecular Interactions in Crystal Structures with $Z' > 1$. *Cryst. Growth Des.* **2016**, *16*, 2988–3001.
- (150) 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.
- (151) Cavallo, G.; Metrangolo, P.; Milani, R.; Pilati, T.; Priimagi, A.; Resnati, G.; Terraneo, G. The Halogen Bond. *Chem. Rev.* **2016**, *116*, 2478–2601.
- (152) Kozuch, S.; Martin, J. M. L. Halogen Bonds: Benchmarks and Theoretical Analysis. *J. Chem. Theory Comput.* **2013**, *9*, 1918–31.
- (153) Řezáč, J.; Riley, K. E.; Hobza, P. Benchmark Calculations of Noncovalent Interactions of Halogenated Molecules. *J. Chem. Theory Comput.* **2012**, *8*, 4285–92.
- (154) 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.

- (155) Misquitta, A. J.; Welch, G. W.; Stone, A. J.; Price, S. L. A First Principles Prediction of the Crystal Structure of C₆Br₂ClFH₂. *Chem. Phys. Lett.* **2008**, *456*, 105–109.
- (156) Mooij, W. T.; Leusen, F. J. J. Multipoles Versus Charges in the 1999 Crystal Structure Prediction Test. *Phys. Chem. Chem. Phys.* **2001**, *3*, 5063–5066.
- (157) 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. *Cryst. Growth Des.* **2004**, *4*, 1327–1340.
- (158) Asmadi, A.; Neumann, M. A.; Kendrick, J.; Girard, P.; Perrin, M. A.; Leusen, F. J. J. Revisiting the Blind Tests in Crystal Structure Prediction: Accurate Energy Ranking of Molecular Crystals. *J. Phys. Chem. B* **2009**, *113*, 16303–16313.
- (159) Wales, D. J.; Salamon, P. Observation Time Scale, Free-Energy Landscapes, and Molecular Symmetry. *Proc. Natl. Acad. Sci. U. S. A.* **2014**, *111*, 617–622.
- (160) Carr, J. M.; Mazauric, D.; Cazals, F.; Wales, D. J. Energy Landscapes and Persistent Minima. *J. Chem. Phys.* **2016**, *144*, 054109.
- (161) 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. *J. Appl. Crystallogr.* **2008**, *41*, 466–470.

Graphical TOC Entry

