Genarris 2.0: A Random Structure Generator for Molecular Crystals

Rithwik Tom[a], Timothy Rose[b], Imanuel Bier[b], Harriet O'Brien[b], Álvaro Vázquez-Mayagoitia[c], Noa Marom[a,b,d,*]

[a]*Department of Physics, Carnegie Mellon University, Pittsburgh, PA, 15213, USA*
[b]*Department of Materials Science and Engineering, Carnegie Mellon University, Pittsburgh, PA, 15213, USA*
[c]*Computational Science Division, Argonne National Lab, Lemont, Illinois 60439, USA*
[d]*Department of Chemistry, Carnegie Mellon University, Pittsburgh, PA, 15213, USA*

Abstract

Genarris is an open source Python package for generating random molecular crystal structures with physical constraints for seeding crystal structure prediction algorithms and training machine learning models. Here we present a new version of the code, containing several major improvements. A MPI-based parallelization scheme has been implemented, which facilitates the seamless sequential execution of user-defined workflows. A new method for estimating the unit cell volume based on the single molecule structure has been developed using a machine-learned model trained on experimental structures. A new algorithm has been implemented for generating crystal structures with molecules occupying special Wyckoff positions. A new hierarchical structure check procedure has been developed to detect unphysical close contacts efficiently and accurately. New intermolecular distance settings have been implemented for strong hydrogen bonds. To demonstrate these new features, we study two specific cases: benzene and glycine. For all polymorphs, the final pools contained the experimental structure.

Keywords: Molecular crystals, Random structure generation, Crystal structure prediction

*Corresponding author.
E-mail address: nmarom@andrew.cmu.edu

NEW VERSION SUMMARY

Manuscript Title: Genarris 2.0: A random structure generator for molecular crystals
Authors: Rithwik Tom, Timothy Rose, Imanuel Bier, Harriet O'Brien, Alvaro Vazquez-Mayagoitia, Noa Marom
Program Title: Genarris 2.0
Journal Reference:
Catalogue identifier:
Licensing provisions: BSD-3 Clause
Programming language: Python, C
Operating system: Linux
Classification: Crystallography
External routines/libraries: Spglib, ASE, pymatgen, SciPy, mpi4py, scikit-learn, PyTorch, FHI-aims.
Nature of problem: Molecular crystal structure prediction.

Solution method: Genarris 2.0 generates molecular crystal structures over the 230 space groups, on general and special Wyckoff positions, using physical constraints. Down-sampling of the generated structures may be performed subsequently, based on molecular crystal packing descriptors and an unsupervised machine learning algorithm. Lastly, ab initio structure relaxation may be performed for the final pool. Depending on the user-defined workflow implemented, Genarris may be used to generate diverse molecular crystal datasets to seed evolutionary algorithms or to train machine learning algorithms or as a standalone crystal structure prediction method.

Restrictions: For crystal structure generation, the molecule of interest must be semi-rigid with no bond rotational degrees of freedom.

Unusual features: Genarris 2.0 is a highly distributed program, making use of MPI for Python to implement bindings of the Message Passing Interface (MPI) and offers the user the ability to design and implement workflows by executing a user-defined list of procedures. Genarris 2.0 implements new features including a machine learning model for estimating the molecular volume in the solid state from the single molecule structure, structure generation in special Wyckoff positions of space groups, hierarchical structure checks including rigorous treatment of non-orthogonal structures, and clustering and down-selection workflows combining first principles simulations with machine learning.

1. Introduction

The properties of molecular crystals depend not only on their constituents but also the relative arrangement of the molecules inside the unit cell. Properties such as the stability [1–3], electronic conductivity [4–8], solubility and bioavailability [9, 10], have all been observed to vary as a function of the molecular crystal solid state form. The molecules comprising these crystals are held together by weak intermolecular interactions [11, 12] and thus can commonly be experimentally synthesized in multiple forms [13, 14]. This phenomenon, known as polymorphism, has been of great importance to pharmaceutical research and for the design of high performance organic electronics [5, 15, 16].

The field of crystal structure prediction (CSP) is devoted to the prediction of the solid state forms of a molecule [17–23]. CSP requires algorithms that can efficiently generate new structures in order to sample the high dimensional configuration space associated with molecular crystals [23, 24]. Random, and quasi-random, sampling of the configuration space has been established as a critical component of CSP workflows within the Cambridge Crystallographic Data Centre (CCDC) CSP blind test [17–22]. Most of the groups that participated in the sixth CSP blind test used a random crystal structure generation method [25–29]. Random crystal structure generation methods identified four of the five, chemically diverse target systems in the sixth blind test, demonstrating their importance for CSP [22].

Random crystal structure generation methods for CSP follow a similar procedure. First a space group [30] is chosen for the new structure. Second, random unit cell parameters commensurate with the space group's crystal system are generated. Third, the molecule positions and orientation of each independent molecule are randomly sampled within the asymmetric unit. Finally, the symmetry operations of the space group are applied to the asymmetric unit generating all molecules in the unit cell. The generated structures are subsequently relaxed using either a system specific force field [31–33] or a fully *ab initio* approach [26, 34, 35]. The success of random structure generation stems from unbiased and diverse sampling covering the potential energy surface, followed by a structural relaxation to the nearest local minima, hopefully converging to all experimentally observed polymorphs [25–27].

Despite their overall similarity, structure generation methods from the sixth blind test differ in subtle ways. Structure parameters may be sampled using either a uniformly random number generator [26, 34], or quasi-random, low discrepancy sequences [25, 27, 28]. Structure generation may be performed over all space groups [34], or using only the most common space groups [24, 27] observed in the Cambridge Structural Database (CSD) [36, 37]. A critical component of the generation procedure is approximating the volume of the molecular crystal before generation. Several methods have been proposed, such as adding up atomic volumes [26], using the morphology of the molecule [25, 38], or relaxing a few handmade structures [27, 29]. It has been demonstrated that random CSP methods may be sensitive to the choice of unit cell volume [25]. Therefore, it is important to use an accurate volume estimation method. Additionally, structures with reasonable densities are typically closer to their respective local minima making structure relaxations more efficient. Lastly, most random crystal structure generation packages are only capable of generating struc-

tures in general Wyckoff positions and rely on the serendipitous generation of structures with molecules occupying special positions. However, analysis of the CSD has shown that molecules with internal symmetry often occupy special Wyckoff positions [39].

In addition to random structure generation, optimization methods, such as Monte Carlo (MC) simulated annealing or parallel tempering and evolutionary algorithms have also demonstrated success in the CSP blind tests [22]. Optimization algorithms use information about the energy of generated structures to make sampling decisions and speed up convergence towards the global minimum. MC Simulated annealing is performed at temperatures that allow configuration exploration [40]. Updates to the geometry are performed by randomly sampling a new position, calculating the potential energy of updated system, and then deciding to accept or reject the update based on the temperature of the simulation. MC Parallel tempering works similarly while also allowing systems at different temperatures to exchange complete configurations [41]. Evolutionary algorithms explore the configuration space by blending or mutating structural genes to generate new structures [42–44]. The genes of structures that are stabler are chosen more often to perform global optimization. Optimization algorithms typically require an initial set of random structures to start.

Here we present a new version of Genarris [34], an open source Python package that performs random structure generation for homomolecular molecular crystals of semi-rigid molecules with no bond rotational degrees of freedom using general and special Wyckoff positions. Genarris 2.0 offers several improvements over the previous version. The parallelization model has been changed from Python multiprocessing to MPI for Python (mpi4py) [45] to enable more efficient utilization of many cores and seamless sequential execution of user-defined workflows. A new machine learning method for volume estimation, based on a topological molecular descriptor, provides accurate volume predictions across a chemically diverse dataset from the CSD. The speed of structure generation has been significantly increased by developing a new hierarchical scheme for intermolecular distance checks. New settings have been implemented to improve structure generation for systems with strong hydrogen bonds. The performance of new the features in Genarris 2.0 is demonstrated for glycine, which contains relatively strong intermolecular hydrogen bonds, and benzene, a symmetric molecule occupying special Wykckoff positions.

2. Code description

Genarris 2.0 is written in Python 3, with the exception of the new structure generation function, Pygenarris, which is written in C and automatically compiled and installed into Genarris 2.0 as a Python library. Genarris only requires standard Python libraries to install on any machine (i.e. numpy, sci-kit learn, mpi4py, spglib, pymatgen,and ASE, PyTorch). Genarris 2.0 is parallelized with MPI, using mpi4py (Sec. 2.1). For energy evaluations and geometry relaxations, Genarris currently interfaces with the electronic structure package FHI-aims [46]. It may be adapted to interface with any other electronic structure, force field, or machine learning package that accepts an MPI communicator as an argument.

The workflow of Genarris 2.0 is depicted in Figure 1. It begins by estimating the crystal unit cell volume. Given the desired number of molecules per unit cell (Z), the estimate is

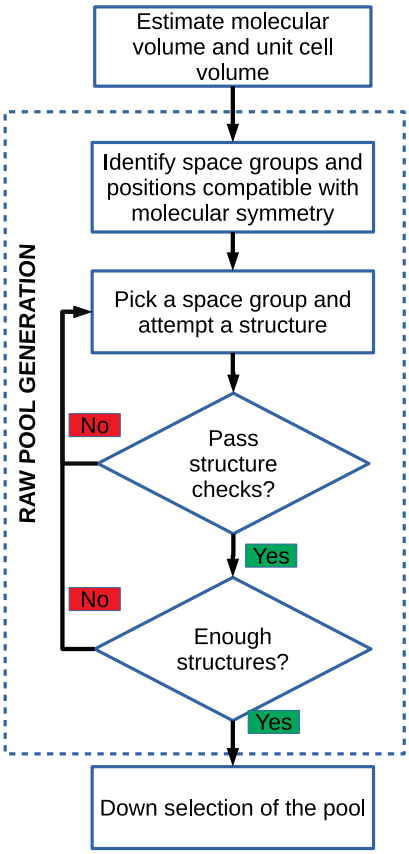

Figure 1: The workflow of Genarris 2.0

obtained by relaxing the single molecule geometry and applying a machine-learned model trained on a dataset of experimental structures from the Cambridge Structural Database (CSD) (Sec. 2.2). Crystal structure generation begins by determining all compatible space groups that have Z molecules in the conventional unit cell, which can occupy general or special Wyckoff positions, and no more than one molecule in the asymmetric unit (Sec. 2.3). Genarris automatically determines if the molecule is sufficiently symmetric to be placed on a special position, within a user-defined tolerance (Sec. 2.3.1). Genarris moves sequentially through this list of space groups, generating a user-defined number of structures per allowed space group and checking them to ensure that no two molecules are unphysically close to each other (Sec. 2.4). If the user-defined maximum number of consecutive failed generation attempts for a space group is reached, Genarris will proceed to the next space group on the list.

Once a "raw" pool of physically reasonable, random structures is generated, a user-defined sequence of energy evaluation, clustering, and selection steps may be performed to produce a smaller curated pool of structures, which can be used, e.g., as an initial population for a genetic algorithm [44, 47]. For clustering, Genarris uses the affinity propagation (AP) machine learning algorithm [48]. Two types of feature vectors are available in Genarris

2.0, the relative coordinate descriptor (RCD) [34] and radial symmetry functions (RSF), implemented in PyTorch, similar to those described in Ref. [49]. Three workflows for down selection have been proposed previously [34]. Here, a new "Robust" workflow is proposed (Sec. 2.5). Lastly, full geometry relaxation may be performed for the final pool of structures.

Genarris 2.0 automatically executes all the procedures in the user-defined procedure list in the order specified. A single input file contains the user defined settings for all desired procedures. This includes the number of cores to be used for each procedure, as different procedures scale differently (see Sec. 2.1). Genarris can infer some parameters from previous sections of the workflow. For example, the output file containing the relaxed geometry of the single molecule becomes the default molecule path of subsequent sections if it exists. The user may reorder the procedures as long as the dependencies are satisfied (e.g., feature vector calculation must be performed prior to clustering). If Genarris is aborted, it will restart from the procedure where it was stopped. If Genarris is stopped during structure generation, it will resume from the last generated space group. For procedures that run FHI-aims geometry relaxations for a batch of structures, Genarris will restart all FHI-aims jobs from their last relaxation step.

2.1. Parallelization

Genarris 2.0 is parallelized using the message passing interface (MPI) paradigm via the mpi4py package. MPI enables immediate cross-platform portability without code changes. The structure generation function in Genarris 2.0 determines the number of allowed space groups for the given molecule, n, and accepts as input the number of structures to generate for each of these space groups. Hence, structure generation and subsequent structure checks (Sec. 2.3) are embarrassingly parallelized over the total number of structures desired, N, with the problem size (maximum number of usable cores) for the generation and structure check procedures equal to N/n.

For clustering (see Sec. 2.5), both the RCD and RSF feature vector calculations are embarrassingly parallelized with problem size N. As explained in Sec. 2.5, the number of cluster exemplars output by the affinity propagation (AP) algorithm generally increases with the value of the *preference* hyperparameter. Therefore, a parallelized version of the standard binary search algorithm has been implemented to output a specified number of clusters C within a tolerance tol. The *preference* range is initially wide ($[-1000, 1000]$). This range is evenly partitioned into R *preference* values, where R is the number of total MPI ranks available. Each rank executes AP with its assigned value of *preference* and reports the number of clusters obtained to the root rank. The root rank sets the *preference* range upper (lower) bound to the *preference* that returned the lowest (highest) number of clusters above (below) the target number of clusters. The root then partitions the updated *preference* range and assigns each rank its new *preference* value. The procedure is repeated until a *preference* value is found which yields $C \pm tol$ clusters. Because there is a small chance that increasing *preference* may cause a lower number of clusters output, fail-safes have been implemented. For example, if the current *preference* range fails to yield a number of clusters within $C \pm tol$, then the *preference* range is widened by a random amount. In addition, the user may have the program output the closest number of clusters to C within a desired number of iterations.

The memory usage is kept manageable by writing and accessing the affinity and distance matrices via memory maps so that each rank does not make a redundant copy.

Genarris currently interfaces with FHI-aims for energy evaluations and geometry relaxations. It may be adapted to interface with any other electronic structure, force field, or machine learning package that accepts an MPI communicator (a group of ranks (cores) that can send/receive messages among themselves) as an argument. The master rank (rank 0 of the global MPI communicator) keeps track of which jobs remain and gives the next job to the first responsive rank. It gives the following job to the next responder, and so on. This yields automatic load-balancing. Only a single, designated rank of a particular sub-communicator (sub-group of ranks) requests a job from the master to limit communication latency. Each group performs a different FHI-aims job enabling massive parallelism.

2.2. Unit cell volume estimation

The solid form volume of a molecule is defined as the volume of the unit cell divided by Z, the number of molecules contained in the unit cell. Accurate prediction of the solid form volume of an input molecule is critical for generating structures with reasonable unit cell volumes. To this end, a machine learned model using a Monte Carlo volume estimation scheme and a topological molecular fingerprint constructed based on atomic neighborhoods was developed. The model was trained on a dataset obtained from the CSD using the Conquest program [50]. A chemically diverse dataset was compiled, containing molecules with 5 to 260 atoms comprising the organic elements, H, C, N, O, all the halogens, F, Cl, Br, and I, as well as B, P, S, Si, Te, and Se. The accuracy of the machine learned model is within the range of polymorph density differences as identified from 2,173 unique, homomolecular polymorph pairs from the CSD.

2.2.1. Dataset construction

The dataset used for training the volume estimation model was obtained from the CSD using the Conquest program [50]. The search was performed over entries of the 2017 version of the CSD for structures of homomolecular organic crystals, characterized at room temperature, under standard pressure, and containing the text phrase 'polymorph'. As described elsewhere [51–53], all polymorphic compounds in the CSD are flagged with the tag 'polymorph'. All duplicate structures were identified using the COMPACK program [54] and removed. This yielded 3,768 individual entries in the dataset and 2,173 unique polymorph pairs, which is similar in size to previous statistical studies of homomolecular polymorphs [52, 53].

The expected variance of the percent difference in the solid form volume of a molecule due to polymorphism was calculated using this dataset. All unique pairs of polymorphs were identified and the percent difference between each polymorph density was calculated. The percent difference of densities is equivalent to that of the solid form molecular volume because the molecular weight remains constant for these systems. The distribution of percent differences is plotted in Figure 2. The distribution has a standard deviation of 2.95% with respect to the solid form volume of the molecule, consistent with numerous previous reports of molecular crystal density estimation [55–58]. This indicates that polymorphs which can

exist under the same temperature and pressure conditions could posses significant volume differences owing to the complex nature of the relatively weak intermolecular interactions that govern the lattice energy of homomolecular crystals. Thus, the distribution presented in Figure 2 places a lower bound on the expected accuracy of estimated solid form molecular volumes.

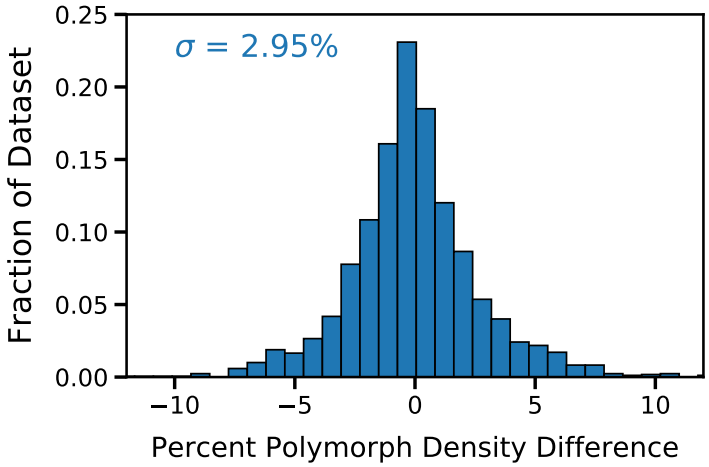

Figure 2: Histogram of the percent difference of polymorph density for 2,173 unique pairs of polymorphs in a dataset obtained from the CSD. The standard deviation of the distribution is displayed in the top left corner.

2.2.2. Monte Carlo volume estimation

Volume estimation is performed by placing a sphere with a van der Waals (vdW) radius [59] at the position of each atom in the molecule [60, 61]. A Monte Carlo method is then used evaluate the volume occupied by the spheres. First, a three-dimensional box encompassing the molecule is defined. Points within the box are sampled randomly and determined if they fall within at least one of the atomic vdW spheres. The ratio between the number of sampled points and the number of points found within a sphere multiplied by the volume of the three-dimensional box is the estimated volume of the molecule. The Monte Carlo volume estimation is deemed to be converged when the estimated volume changes by less than 10^{-3}Å^3 after 10^6 new points are sampled.

The ratio between the experimental molecular solid form volume and the Monte Carlo volume estimate for the polymorph dataset was found to be 1.47, indicating that the Monte Carlo method systematically underestimates the true solid form volume. Using this linear relationship to predict the solid form volume of the molecule achieves a standard deviation of 4.72% error with respect to the dataset (Figure 3). To improve the accuracy of the volume estimation model, specific information about the chemical environment of the atoms in the molecule must be included. To this end, a molecular topological fragment representation has been developed.

2.2.3. Molecular topological fragment model

We present a topological molecular fingerprint representation for predicting solid form molecular volume within the accuracy of polymorph density differences. The representation is based on molecular fragments determined through analysis of the CSD dataset. The fact that the fragments are not predefined enables an unbiased choice of fragments such that they can represent any structural class. The complexity of the model increases with the size and chemical diversity of the dataset making this representation amenable to large datasets as well as datasets comprising a restricted chemical space. Moreover, representation is invariant to permutations of the atom indexing. The molecular topological fragment representation can be used to predict any molecular property of interest with linear and non-linear regression or classification models and can also be used to compute chemical similarity between molecules using metrics such as the Tanimoto coefficient [62]. The Genarris 2.0 source code includes a model construction Python class, enabling users to quickly build topological fingerprints for a training dataset, regularize the model, evaluate the accuracy on a target dataset, and output graphs of predicted values versus target values to asses the performance of the model.

The construction of each molecule's topological fragment representation begins by generating a string representation for every atom in the molecule. The string captures each atom's local environment and is built as follows: First, a graph is constructed with nodes and edges that correspond to the nuclei and bonds of the molecule, respectively. Second, for each atom, the bonds to the nearest neighbors are transformed into a string sorted first by the elements of the terminal nuclei, in alphabetical order, then by the atom itself if it is not terminal, and lastly by the elements of the other nuclei the atom is bonded to, which are also sorted in alphabetical order. This string representation is unique for each distinct atomic environment and is generated and stored for each atom in a given molecule.

In order to train a machine learned volume estimation model, every molecule in the dataset must have a vector representation of the same length. This requirement was satisfied by first finding all unique atomic environments across the entire dataset and then constructing their string representation using the above-described procedure. These strings were then sorted in alphabetical order and used to index a vector representation of each molecule. The value at each index of the vector was equal to the number of times the fragment was present in a given molecule. This representation also ensured that the sum of all elements in the vector was equal to the number of atoms in the molecule. By following the described procedure, a unique vector representation for each distinct molecule in the dataset was constructed. Examples of vector representations of glycine and benzene are shown in Table 1. The representation described here is similar to other fragment based representations used in chemical informatics [63, 64].

Table 1: Example of vector representations constructed for a dataset containing benzene and glycine using the molecular topological fragment model.

Fragment	HC	HCCC	HHCCN	HHHNC	HN	OC	OOCC
Benzene	6	6	0	0	0	0	0
Glycine	2	0	1	1	3	2	1

To construct a predictive model for solid form molecular volumes, the volume predicted by the Monte Carlo method was concatenated to the topological fragment representation vector of each molecule. The coefficients for a linear model were then calculated using Bayesian ridge regression as implemented in scikit-learn [65]. The regularization parameter was optimized using a grid search method and five-fold cross validation. The number of features contained in the model was constrained by removing features that did not occur at least thirty times in the dataset. Thirty was identified as the optimal number using a five-fold cross validation scheme. This left 64 unique molecular fragments in the model.

The distribution of errors obtained using the topological fragment model is displayed in Figure 3. It is shown that the fragment based model significantly reduces the error in the predicted solid state molecular volumes compared to the Monte Carlo volume estimation. Furthermore, the fragment based model achieves an error of similar magnitude to the volume differences between polymorphs found in the CSD. Thus, the topological fragment model developed here achieves an accuracy within the error one could expect from polymorph density differences.

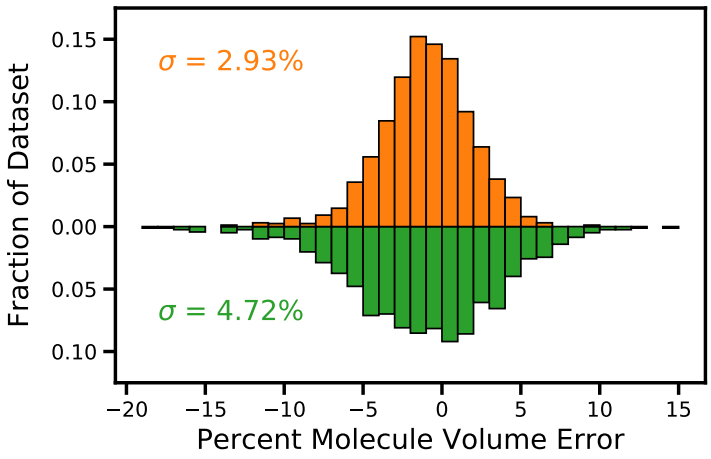

Figure 3: Percent error of the predicted solid form volume for the described dataset from the CSD for a linear model using Monte Carlo volumes and a linear model using Monte Carlo volumes in green and the topological fragment representation of the molecules in orange.

2.3. Structure generation

The generation process begins by identifying all space groups compatible with the number of molecules in the unit cell (Z). Space groups are considered compatible if they have Z molecules in the conventional cell unit cell and no more than one molecule in the asymmetric unit. Genarris 2.0 detects compatible space groups automatically. If the multiplicity of the general position of a space group equals Z, it is deemed compatible regardless of the symmetry of the molecule. If the multiplicity of a special position of a space group equals Z and the molecule has sufficient symmetry to occupy it (within a given numerical tolerance), then the space group is considered compatible. Once the compatible space groups are found,

Genarris 2.0 attempts generation of crystal structures sequentially, starting from the lowest space group number. We note that compatibility does not guarantee successful structure generation in a given space group because generated structures are subjected to additional constraints on the unit cell volume and intermolecular distances, as explained below.

A random volume is drawn from a Gaussian distribution whose mean and standard deviation are the predicted volume and three times the prediction error of our volume estimation method (see Sec. 2.2). The volume is redrawn after a successful generation or after a user-specified number of failed attempts. Subsequently, using this volume, a unit cell of the desired lattice system is constructed randomly as shown in Figure 4. If the attempted position is a general Wyckoff position, then the molecule's orientation is sampled randomly and placed randomly inside the unit cell. The space group symmetry operations are then applied to generate the remaining molecules in the unit cell. Special positions, with the exception of inversion centers, require alignment of the molecule and their treatment is described in Sec. 2.3.1. The attempted structures that pass the intermolecular distance checks, as described in Sec. 2.4, are added to the raw pool. If Genarris is unable to generate a structure within the maximum attempt limit specified by the user, then it proceeds to the next space group.

2.3.1. Generation in special positions of space groups

Special Wyckoff positions are left invariant under at least one symmetry operation of the space group in addition to the identity operation. These symmetry operations define the site symmetry of the special position. For each space group, the International Table of Crystallography [66] lists the special positions whose multiplicity is lower than that of the general position. Only molecules that satisfy the site symmetry of the special position can occupy it. Most molecules do not have higher order symmetries, therefore molecular crystals with molecules occupying special positions are infrequent. According to an analysis of the CSD [39], in 70.1% of the molecular crystals, molecules occupy general positions, and in the remaining structures molecules occupy special positions. Among the special positions, two-fold rotation (2), mirror planes (m), and inversion centers ($\bar{1}$) are the most frequent.

Genarris 2.0 generates molecular crystals with molecules on special positions by checking all possible orientations of the molecule with respect to the symmetry directions of the crystal system [66], as shown in the flowchart in Figure 4. At the start of generation, the program finds all possible molecular axes that may be associated with a symmetry element. For this purpose, first the center of mass of the molecule is shifted to the origin. Then, all atoms of the same element that are farthest from the center of mass are selected. The possible symmetry elements of the molecule would map any of these atoms onto itself or onto another. The axes corresponding to these symmetry elements are obtained by calculating the averages and cross products of the position vectors of the selected atoms. A list of potential molecular axes is constructed. To keep the length of the list minimal, parallel vectors are deleted. Once the potential molecular axes are identified, the code proceeds to check the compatibility of molecule placement at a special position with the specified number of molecules in a unit cell. The molecule's center of mass is placed in a special position, such that one of the molecular axes is oriented along one of the symmetry directions of the crystal system. Then, the

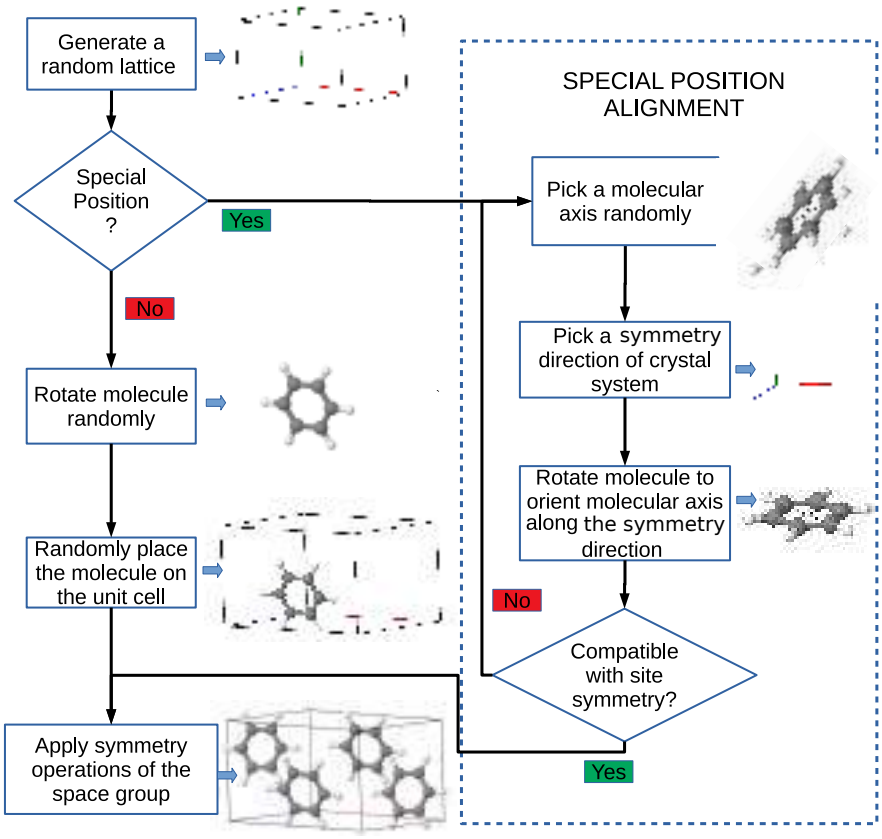

Figure 4: Flowchart of crystal structure generation in Genarris 2.0. Molecules are placed in general Wyckoff positions with a random orientation. In contrast, special positions require specific orientations of the molecule to be compatible with the site symmetry.

symmetry operations of the space group are applied. If the number of molecules that coalesce into one molecule is equal to the order of the site, the special Wyckoff position is regarded as compatible. If not, different molecular axes and symmetry directions are considered. All combinations of molecular symmetry axes and lattice symmetry directions are examined and compatible ones are stored for subsequent generation attempts. Once a molecule is placed in a compatible special position, its geometry is slightly adjusted (within a user-defined tolerance) by averaging over the atomic positions of all the overlapping molecules occupying the same site. Depending on the site symmetry of the special position, the allowed degrees of freedom are randomized. For example, a molecule placed on an inversion center can be freely rotated if the molecule's inversion center coincides with the inversion center of the space group. A molecule with a 2-fold axis of rotation can placed at a suitable Wyckoff position provided this axis coincides with the 2-fold site symmetry axis. The molecule is free to rotate about this axis because the rotated molecule still satisfies the site symmetry of the Wyckoff position. These freedoms of rotation are randomized to promote diversity within special positions.

2.4. Structure checks

Attempted structures are checked to avoid unphysically close intermolecular contacts. Checking the distance between every atom of a molecule and every atom of all neighboring molecules, including its own periodic replicas, has a scaling of $O(N^2)$, where N is the number of atoms in the unit cell. This is found to be the bottleneck of structure generation. To improve the efficiency, Genarris 2.0 performs a series of three hierarchical structure checks. Failed structures are discarded at each stage, such that fewer structures undergo the more rigorous and computationally expensive checks.

The threshold for allowed close contacts between two atoms is called the cutoff distance and is defined based on a specific radius fraction, s_r, of the sum of atomic van der Waals radii [34]. The crystal structure is deemed unphysical and rejected if the distance, d, between two atoms belonging to different molecules is such that

$$d < s_r(r_A + r_B) \tag{1}$$

where r_A and r_B are the van der Waals radii of atom A and atom B, respectively. This ensures the quality of the generated structures. The default value of s_r is 0.85. Based on statistical analysis of structures extracted from the CSD, this is a reasonable setting for all but the strong hydrogen bonds. For these cases, special settings have been implemented in Genarris 2.0, as described in Sec. 2.4.4. For this value of s_r and the target unit cell volume determined as described in Sec. 2.2, random generation of crystal structures may require a large number of attempts (a few thousand to millions) before it passes all three stages of structure checks and is accepted into the pool. Therefore, the new hierarchical structure check procedure is a significant efficiency improvement in Genarris 2.0. The details of each stage are explained below.

2.4.1. Stage I: Fast screening without periodic boundary conditions

For preliminary screening, periodic boundary conditions are completely ignored and only intermolecular distances in the unit cell are evaluated. Because distances are computed using the Euclidean norm, this stage is the fastest. If the centers of mass of a pair of molecules in a cell are much farther than twice the molecule length, defined as the maximum distance between two atoms of a molecule, then those pairs are ignored as these molecules cannot overlap. We find that most of the unphysical structures generated are rejected at this stage. The structures that pass this screening proceed to the second stage of structure checks.

2.4.2. Stage II: Distance checks with periodic replicas

In this stage, the distances of a molecule from other molecules in the unit cell as well as its own periodic images are checked against the cutoff distance. An approximate minimum image convention is implemented for non-orthogonal cells. To accelerate the distance checks, non-orthogonal cells undergo a lattice reduction. Let $\mathbf{a} = [a_x, 0, 0]$, $\mathbf{b} = [b_x, b_y, 0]$, and $\mathbf{c} = [c_x, c_y, c_z]$ be the lattice vectors in a Cartesian coordinate system. It is possible to choose a less oblique lattice which satisfies: $a_x, b_y, c_z > 0$; $|b_x|, |c_x| \leq a_x/2$; and $|c_y| \leq b_y/2$.

The Stage II algorithm is illustrated in Figure 5. First, the atom positions are expressed in fractional coordinates. Then, the distance between two atoms is computed in fractional

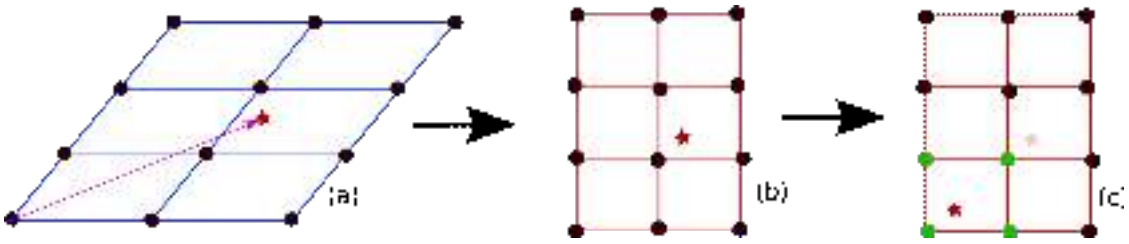

Figure 5: A two dimensional representation of Stage II approximate distance evaluation under periodic boundary conditions. a) An oblique lattice in Cartesian coordinates. The star denotes the point whose distance to the nearest lattice point we need to find. b) Once the lattice is converted from real space into the fractional basis, it is easy to find the box that bounds the point. c) The nearest lattice point is likely to be one of the real space points that map to the green points.

space and translated to the "origin cube", spanned by the vectors $[1,0,0],[0,1,0],[0,0,1]$. Finally, the minimum Cartesian distance of this point from the corners of the origin cube is calculated. For orthogonal cells, the closest point in fractional space necessarily corresponds to the closest point in real Cartesian space. However, for an oblique triclinic lattice a different lattice point may be closer to this point. Therefore, if a non-orthogonal structure passes Stage II, it proceeds to Stage III for a more rigorous check.

2.4.3. Stage III: Rigorous checks for non-orthogonal cells

Complete structure checks require exact evaluation of distances under minimum image convention. For non-orthogonal cells, this problem is a three-dimensional case of the well-studied closest vector problem [67]. If the lattice is translated such that one of the two points coincides with the origin, we need to find the distance d to the nearest lattice point of the position vector $\mathbf{x}$ of the second point. That is,

$$d^2 = \min_{\mathbf{n}} |\mathbf{L}^T\mathbf{n} - \mathbf{x}|^2, \tag{2}$$

where $\mathbf{n} = [n_x, n_y, n_z]$; $n_x, n_y, n_z \in \mathbb{Z}^3$; and $\mathbf{L} = [\mathbf{a}, \mathbf{b}, \mathbf{c}]^T$. This problem is encountered in communication theory, where the received signal over a communication line is decoded by finding the nearest lattice point [68]. One popular approach is the Finck and Pohst sphere decoder [69, 70] method, where the closest lattice point is found using a tree search and the depth of the tree corresponds to the dimension of the problem. Genarris 2.0 uses a version of the sphere decoder to compute the exact distance under minimum image convention for non-orthogonal cells. The distance estimate obtained from Stage II is used as the initial sphere radius for the sphere decoder algorithm. This step is the slowest, but only few non-orthogonal structures that pass Stage I and Stage II reach Stage III. Hence, the overall efficiency is not compromised.

2.4.4. Intermolecular cutoff distances

Choosing appropriate intermolecular cutoff distances is critical for generating physically reasonable structures. In Genarris 2.0, cutoff distances are a function of the elements participating in the intermolecular interaction. For vdW interactions, cutoff distances are im-

plemented using an s_r of 0.85. An s_r of 0.85 was determined to be a physically reasonable value based on our statistical analysis of intermolecular contacts in a data extracted from CSD and presented in Figure 6 as well as an earlier analysis [71]. However, for hydrogen bonds, the intermolecular distance may be considerably shorter than the s_r value used for weaker intermolecular interactions [71, 72]. Hence, new settings for the allowed interatomic distances for hydrogen bonds have been implemented in Genarris 2.0.

Hydrogen bonds among the most important intermolecular interactions in both naturally occurring and artificially engineered molecular crystals [72]. Intermolecular hydrogen bonds are denoted as XH$\cdots$Y where X is the donor, which is covalently bonded to the hydrogen, and Y is the acceptor, which belongs to a different molecule than X. The cutoff distance between H and Y implemented in Genarris 2.0 depends on the identity of atoms X and Y. However, these cutoff distances are applicable to any functional group pair that would participate in an intermolecular hydrogen bond for homomolecular crystals.

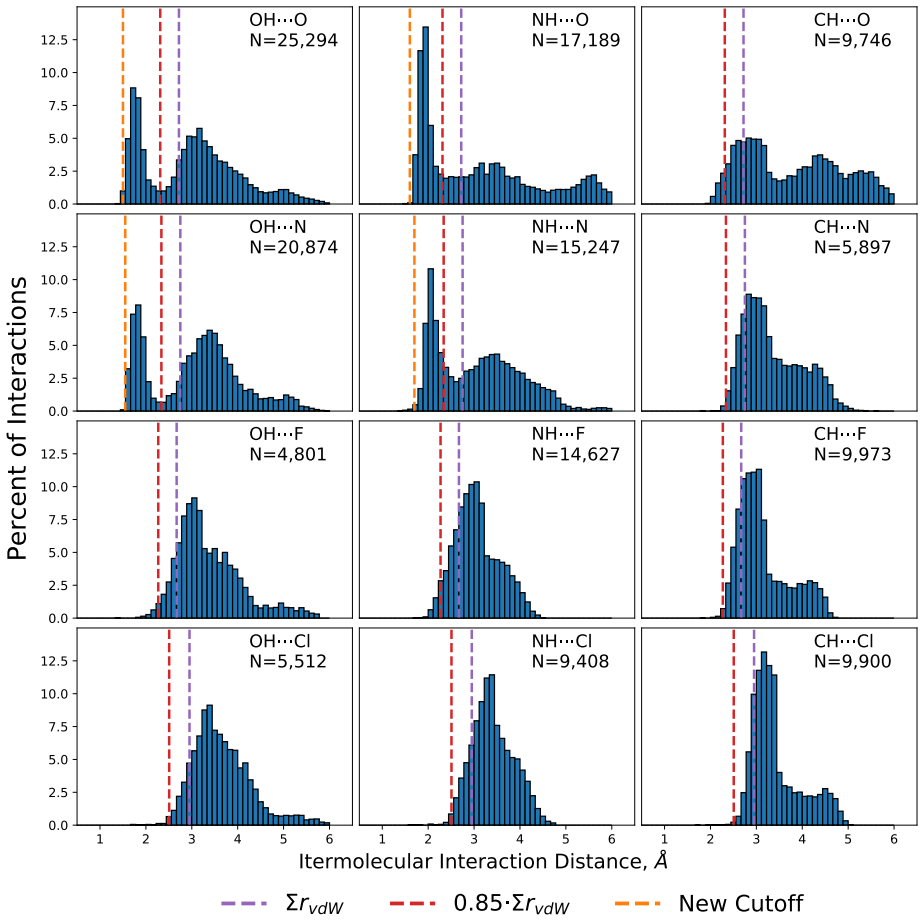

Figure 6: Plots of the number of observations as a function of distance for intermolecular contact distances mined from the CSD and gathered from the IsoStar program. Each histogram is labeled with its hydrogen bond and the number of interactions obtained from IsoStar are labeled N. Drawn on the histograms are vertical lines at the distance corresponding to the sum of the vdW radii, a vdW cutoff of 0.85, and the new hydrogen bond cutoff distances.

Table 2: New cutoff distances implemented in Genarris for intermolecular hydrogen bonds. The cutoff distances are compared to the sum of the van der Waals radii for the intermolecular interactions using the specific radius (s_r) fraction defined in Sec. 2.4.

Contact Type	**Cutoff Distance** ($\mathring{A}^3$)	**Sum of van der Waals radii** ($\mathring{A}^3$)	s_r
OH$\cdots$O	1.5	2.72	0.55
OH$\cdots$N	1.6	2.75	0.58
NH$\cdots$O	1.6	2.72	0.59
NH$\cdots$N	1.7	2.75	0.62

Table 2 displays the newly implemented contact distances for hydrogen bonds, in which oxygen or nitrogen are the donor and acceptor. These values were determined based on the existing literature [73], as well as statistical searches of the CSD using the IsoStar program [74]. The IsoStar program provides distributions of nonbonded, intermolecular distances between pairs of functional groups. The central and contact functional groups were chosen across the available pK_a range [75] for each type of hydrogen bond in order to develop general three body cutoff distances for all relevant hydrogen bonds. The results of the IsoStar searches are shown in Figure 6. For hydrogen bonds involving oxygen and nitrogen as the donor and acceptor, the sum of the vdW radii multiplied by the default s_r value of 0.85 (red dashed lines) exceeds a large number of non-bonded interaction distances, illustrated by the left-most peak of the bimodal distributions. Using the default s_r value, structures with strong hydrogen bonds, such as glycine, would be deemed unphysical and discarded. With the new settings listed in Table 2, they would be considered physically reasonable. For hydrogen bonds involving halogens [76], or those with carbon as the donor atom [77], the default s_r value of 0.85 is still appropriate.

2.5. Clustering and down-selection

Once a "raw" pool of physically reasonable, random structures is generated, Genarris 2.0 offers the option of performing a user-defined sequence of clustering, energy evaluation, and down selection steps in order to form a smaller curated pool of structures. Here, we use the Robust workflow, as shown in Figure 7. The affinity propagation (AP) algorithm [48] is used here for unsupervised clustering because it has been found to perform better than a popular alternative, k-means clustering. [34]. Briefly, the AP algorithm works by iteratively letting each structure update its belief about which structure is its representative example (exemplar) based on its structural similarity (affinity) which is derived from a feature vector describing the structure. The AP algorithm converges on the autonomously determined number of exemplars that accumulate the most"votes" from their cluster constituents. The *preference* hyperparameter value is proportional to each structure's belief that it is an exemplar, and thus increasing *preference* generally increases the number of exemplars produced. The *preference* hyperparameter of AP is automatically tuned by Genarris 2.0 to produce the desired number of clusters within a user-defined tolerance, as described in Sec. 2.1.

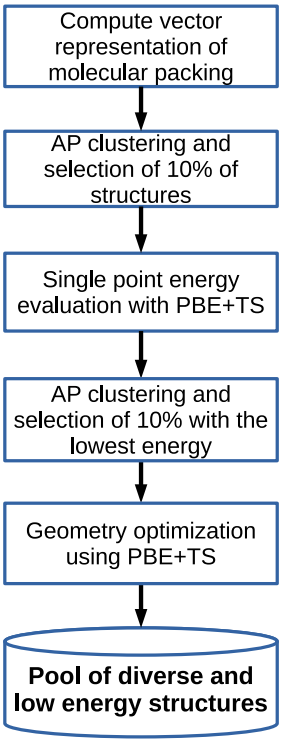

Figure 7: The clustering and down-selection steps of the "Robust" workflow of Genarris 2.0.

For the purpose of clustering via AP, a feature vector describing the molecular packing is calculated for each structure. The relative coordinate descriptor (RCD) [34] and radial symmetry functions (RSF) [49, 78, 79] descriptor are implemented in Genarris 2.0. The default number of clusters for this step is 10% of the the number of structures in the raw pool. For the exemplar of each cluster, a single point energy (SPE) evaluation is performed using FHI-aims with the settings described in Sec. 3 below. Then, AP clustering is performed again with the target number of clusters set to 10% of the reduced pool and the lowest energy structure is selected out of each cluster. Finally, the remaining structures are fully relaxed with FHI-aims as described in Sec. 3. This constitutes the final pool of structures output by Genarris 2.0 using the Robust workflow. The Robust workflow differs from the Diverse workflow of Genarris 1.0 [34] in that it first down-samples based on diversity considerations, which enables evaluating self-consistent single-point DFT energies for a smaller number of structures rather than using the Harris approximation.

3. DFT settings

Genarris 2.0 interfaces with the FHI-aims electronic structure code [46] for geometry relaxation of the single molecule and of the structures in the final pool, as well as for single point energy (SPE) evaluations within the Robust workflow used here. All invocations of FHI-aims in this work used the Perdew-Burke-Ernzerhof (PBE) generalized gradient approximation [80] and the Tkatchenko-Scheffler (TS) pairwise dispersion correction [81] with

lower-level numerical settings, which correspond to the light/tier1 settings of FHI-aims. SPE calculations for crystals were done self-consistently with a $1 \times 1 \times 1$ k-point grid for fast screening. Geometry relaxations of the final pool were performed using a $3 \times 3 \times 3$ k-point grid. Additionally, no constraints were placed on the lattice. Relaxations for the high-pressure $Z = 2$ polymorph of benzene were performed with the pressure set to 25 kbar to reflect the experimental conditions.

4. Case studies

4.1. Benzene

With the chemical formula of C_6H_6, benzene is one of the simplest aromatic hydrocarbons. It is a highly symmetric molecule with a $6/mmm$ point group which allows special positions with 20 different site symmetries. Two known polymorphs of benzene are [82]: a) $Z = 4$ and space group $Pbca$ (61) under ambient pressure and b) $Z = 2$ and space group $P2_1/c$ (14) under high pressure. In both structures, benzene occupies a special position with an inversion center ($\bar{1}$).

Column I in Figures 8 and 9 shows the volume histograms obtained at each step of the Robust workflow for benzene with $Z = 2$ and $Z = 4$, respectively, using RSF-based clustering. The results obtained using RCD-based clustering are provided in the Supplemental Information. The volumes of the experimental structures, $206\mathring{A}^3$ and $490\mathring{A}^3$, respectively, are indicated by solid red lines. The volumes of the experimental structures relaxed with PBE+TS and *lower-level* numerical settings, 180 $\mathring{A}^3$ and 491 $\mathring{A}^3$, respectively, are indicated by solid green lines. The volumes predicted by our machine learned model are indicated by dashed orange lines. For both $Z = 2$ and $Z = 4$, the predicted volumes are closer to the unrelaxed experimental volumes because the volume estimation model was trained on unrelaxed structures from the CSD (see Sec. 2.2). Our prediction for $Z = 4$ is closer to the experimental volume than for $Z = 2$ because the latter forms under pressure of 25 kbar whereas the volume estimation model was trained on structures obtained under ambient pressure. Raw pools of about 6000 structures were generated for both $Z = 2$ and $Z = 4$ with predicted volumes of 243 $\mathring{A}^3$ and 487 $\mathring{A}^3$, and volume standard deviations of 18 $\mathring{A}^3$ and 37 $\mathring{A}^3$, respectively. Figure 8 shows noteworthy density about the experimental volume throughout the workflow progression. The resulting volume distributions are approximately Gaussian until the relaxation step. Panels (IIIc) and (IIId) of Figure 8 show that for $Z = 2$, relaxation under pressure resulted in significant volume contraction, whereas Panels (IIIc) and (IIId) of Figure 9 show that some $Z = 4$ structures expanded beyond the initial volume range.

Column II in Figures 8 and 9 shows the space group distributions obtained at each step of the Robust workflow using RSF-based clustering for benzene with $Z = 2$ and $Z = 4$, respectively.The results obtained using RCD-based clustering are provided in the Supplemental Information. Space groups with general Wyckoff positions are colored in blue and space groups with special Wyckoff positions are colored in orange. Genarris 2.0 attempts to generate a uniform space group distribution. We find that the generated space group

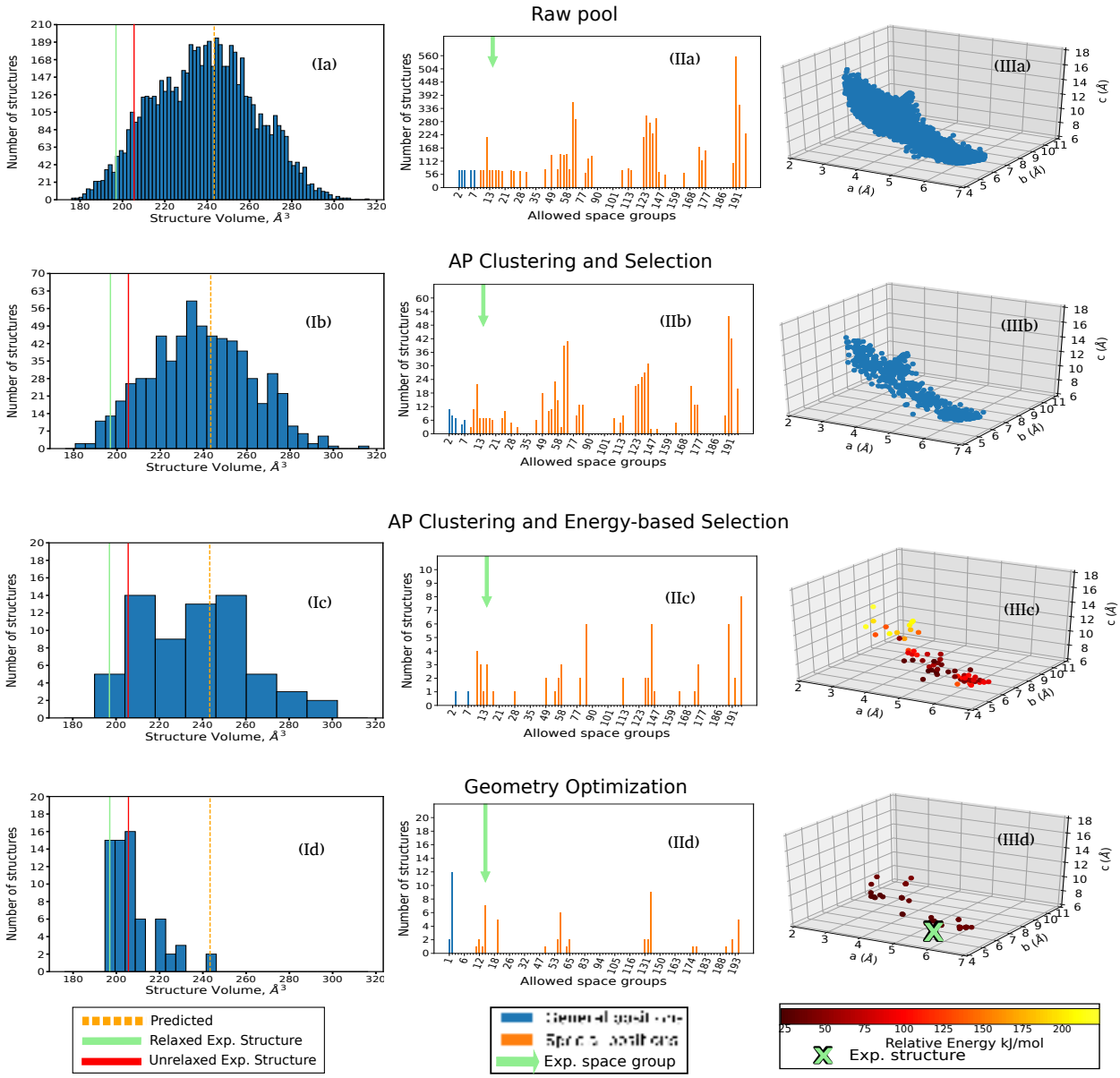

Figure 8: (I) Unit cell volume distributions, (II) space group distributions, and (III) lattice parameter distributions for benzene with $Z = 2$, obtained at each step of the Robust workflow, using RSF-based clustering. On the volume histograms, the solid red line denotes the volume of the experimental structure, observed under pressure of 25 kbar, the solid green line denotes the unit cell volume of the experimental structure after relaxation under pressure of 25 kbar, and the dashed orange line shows the volume predicted by our model, as described in Sec.2.2. On the space group distribution histograms, the green arrow points to the space group of the experimental structure, $P2_1/c$ (14). The green cross in panel (IIId) denotes the (relaxed) experimental structure, which was found in the final pool.

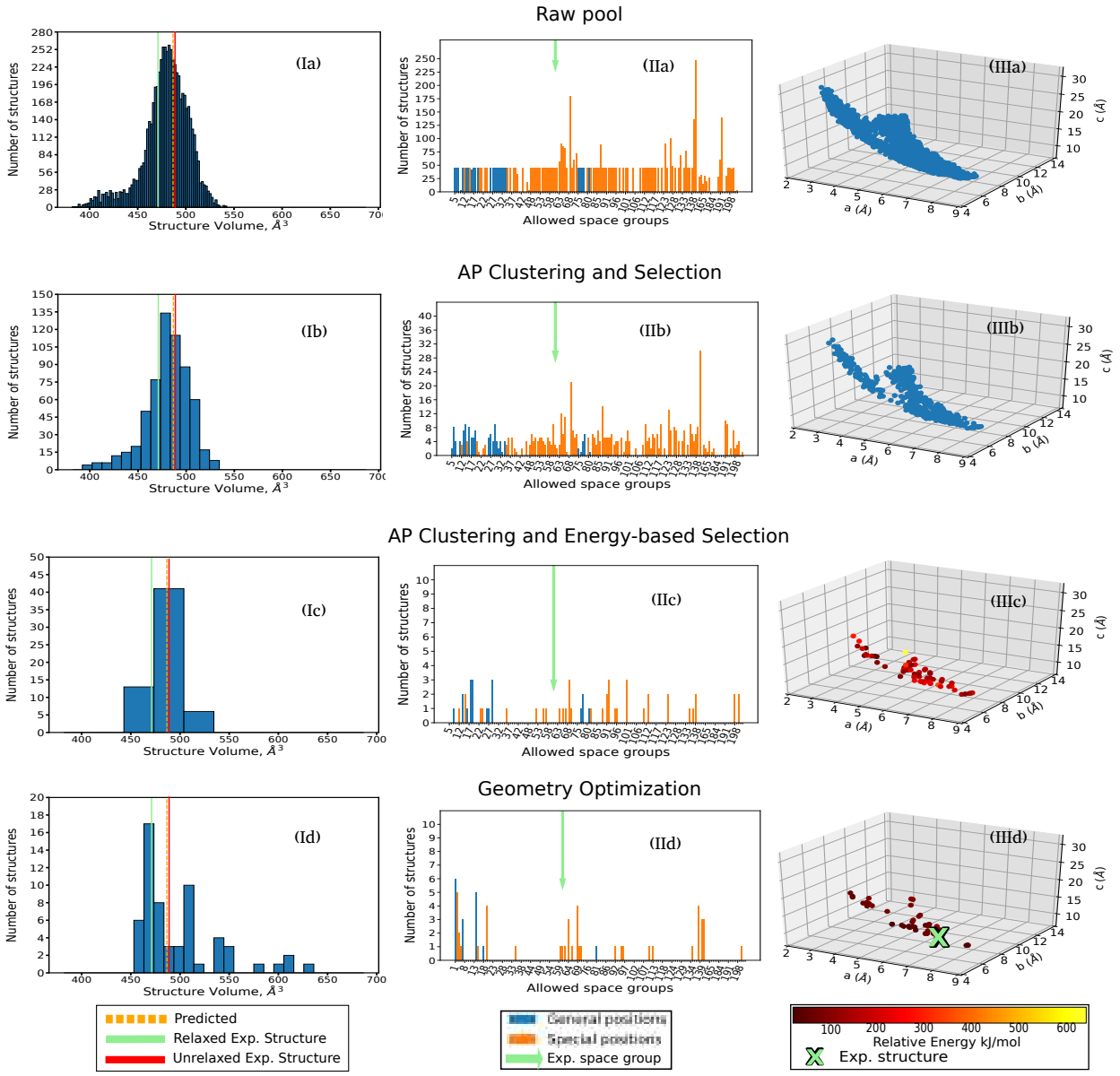

Figure 9: (I) Unit cell volume distributions, (II) space group distributions, and (III) lattice parameter distributions for benzene with $Z = 4$, obtained at each step of the Robust workflow, using RSF-based clustering. On the volume histograms, the solid red line denotes the volume of the experimental structure, the solid green line denotes the unit cell volume of the experimental structure after relaxation, and the dashed orange line shows the volume predicted by our model, as described in Sec.2.2. On the space group distribution histograms, the green arrow points to the space group of the experimental structure, *Pbca* (61). The green cross in panel (IIId) denotes the (relaxed) experimental structure, which was found in the final pool.

distributions are approximately uniform with significant number of structures in the experimental space group for both $Z = 2$ and $Z = 4$. Some space groups may be very difficult or impossible to generate within the given physical constraints. For example, for $Z = 4$, space groups like $P2/m$ (10), $Pmm2$ (25), and $Pmmm$ (47) which have mirror planes are harder to generate as molecules that touch the planes overlap with their own mirror image [39]. In contrast, space groups with glide planes and screw axes are easier to generate because symmetry-equivalent molecules are translated in space. Some structures can have a higher site symmetry on a special position than we attempted to generate, resulting in overpopulation of some space groups. For example, for $Z = 2$, space group $P6/mmm$ (191) has a relatively large occupation as shown in panel (Ib) of Figure 8. Many of these structures were discarded in the subsequent selection steps.

Column III in Figures 8 and 9 shows the lattice parameter distributions obtained at each step of the Robust workflow, using RSF-based clustering for benzene with $Z = 2$ and $Z = 4$, respectively. The results obtained using RCD-based clustering are provided in the Supplemental Information. For the energy-based selection and final relaxation steps, the color scale corresponds to relative energies with respect to the lowest energy structure in the final relaxed pool. As shown in Panels (IIIc) and (IIId) of Figures 8 and 9, the range of relative energies in the relaxed pools for benzene is about $100kJ/mol$ or $1eV$, which is about six times smaller than the range of energies in the unrelaxed pools. The lattice parameter distribution of the raw pools resembles the shape of the surface $|a||b||c| =$ constant (an approximate relation given that benzene is able to assume many lattice types), indicating approximately uniform sampling of the lattice parameter space. Down-selection based on energy tends to filter out very elongated structures whose c parameter is significantly longer than a and b, indicating that these are relatively unstable for benzene. In fact, the experimental unit cells are not elongated. Panels (IIIc) and (IIId) of Figure 8 show that relaxation under pressure resulted in a distribution characterized by a few clusters, suggesting that pressure may have restricted the physically feasible regions. For both $Z = 2$ and $Z = 4$ the experimental structure, indicated by a green X, is found in the final relaxed pools.

4.2. Glycine

Glycine is the simplest proteinogenic amino acid. It is achiral and forms a zwitterion in the solid state. Under ambient conditions, glycine has three common polymorphs: a) α-glycine with $Z = 4$ and space group $P2_1/n$ (14), b) β-glycine with $Z = 2$ and space group $P2_1$ (4), and c) γ-glycine with $Z = 3$ and space group $P3_1$ (144)/ $P3_2$ (145) [83]. The structures belonging to the two space groups are enantiomorphic forms of the chiral γ-glycine crystal. Experimentally, it has been found that the relative thermodynamic stability of the polymorphs at room temperature is $\gamma > \alpha > \beta$ with Gibbs free energy difference (ΔG) of 0.16 kJ/mol between γ-glycine and α-glycine [84]. At temperatures higher than 440 K, α-glycine becomes more stable than γ-glycine. The crystal structure and relative stabilities of the glycine polymorphs have been studied extensively, using different computational methods [85–91].

Glycine is known for its ability to form strong intermolecular hydrogen bonds, owing to which it crystallizes in a relatively dense molecular solid. Column I in Figures 10, 11, and

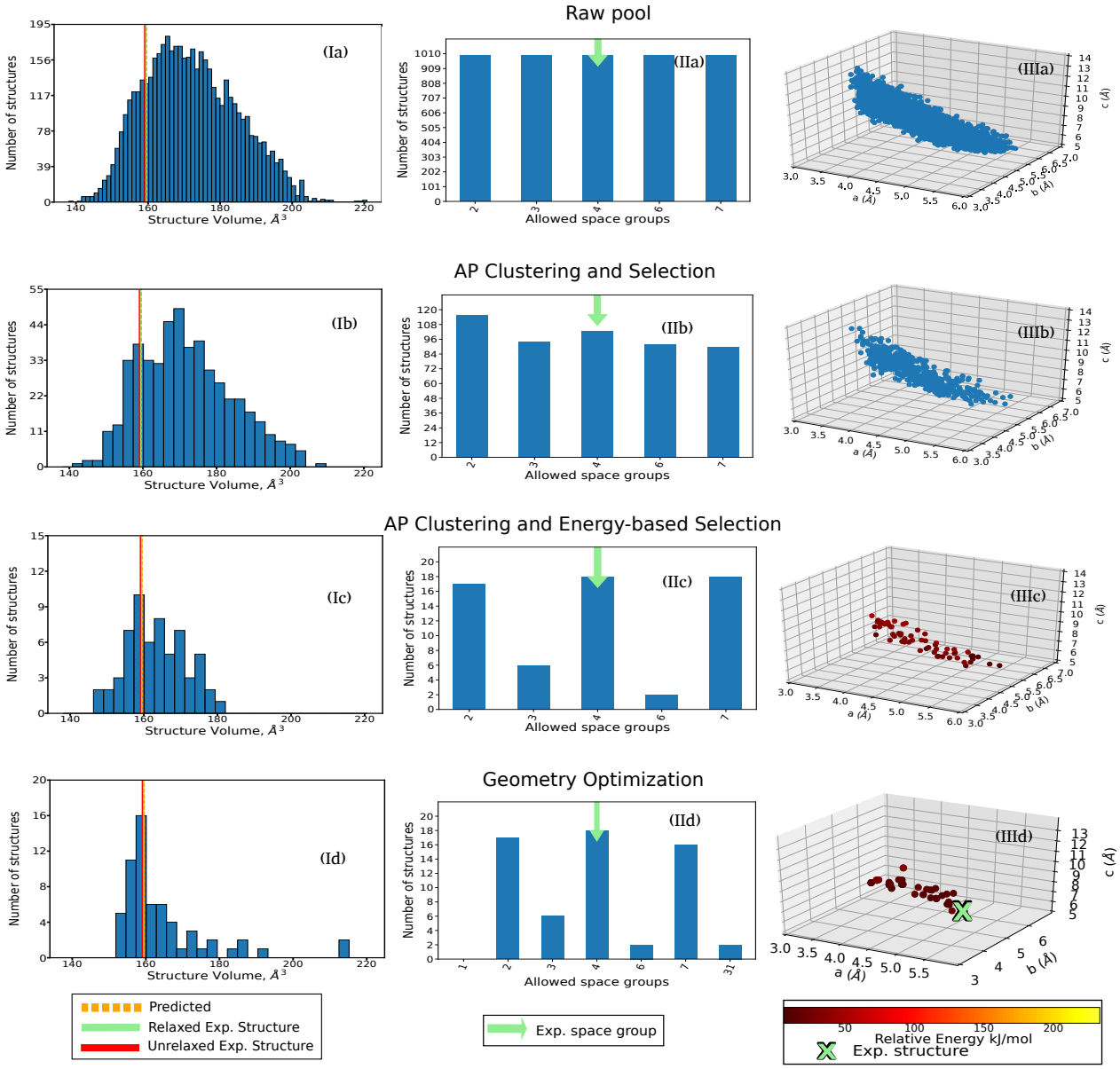

Figure 10: (I) Unit cell volume distributions, (II) space group distributions, and (III) lattice parameter distributions for glycine with $Z = 2$, obtained at each step of the Robust workflow, using RSF-based clustering. On the volume histograms, the solid red line denotes the volume of the experimental structure, the solid green line denotes the unit cell volume of the experimental structure after relaxation, and the dashed orange line shows the volume predicted by our model, as described in Sec.2.2. On the space group distribution histograms, the green arrow points to the space group of the experimental structure, $P2_1$ (4). The green cross in panel (IIId) denotes the (relaxed) experimental structure, which was found in the final pool.

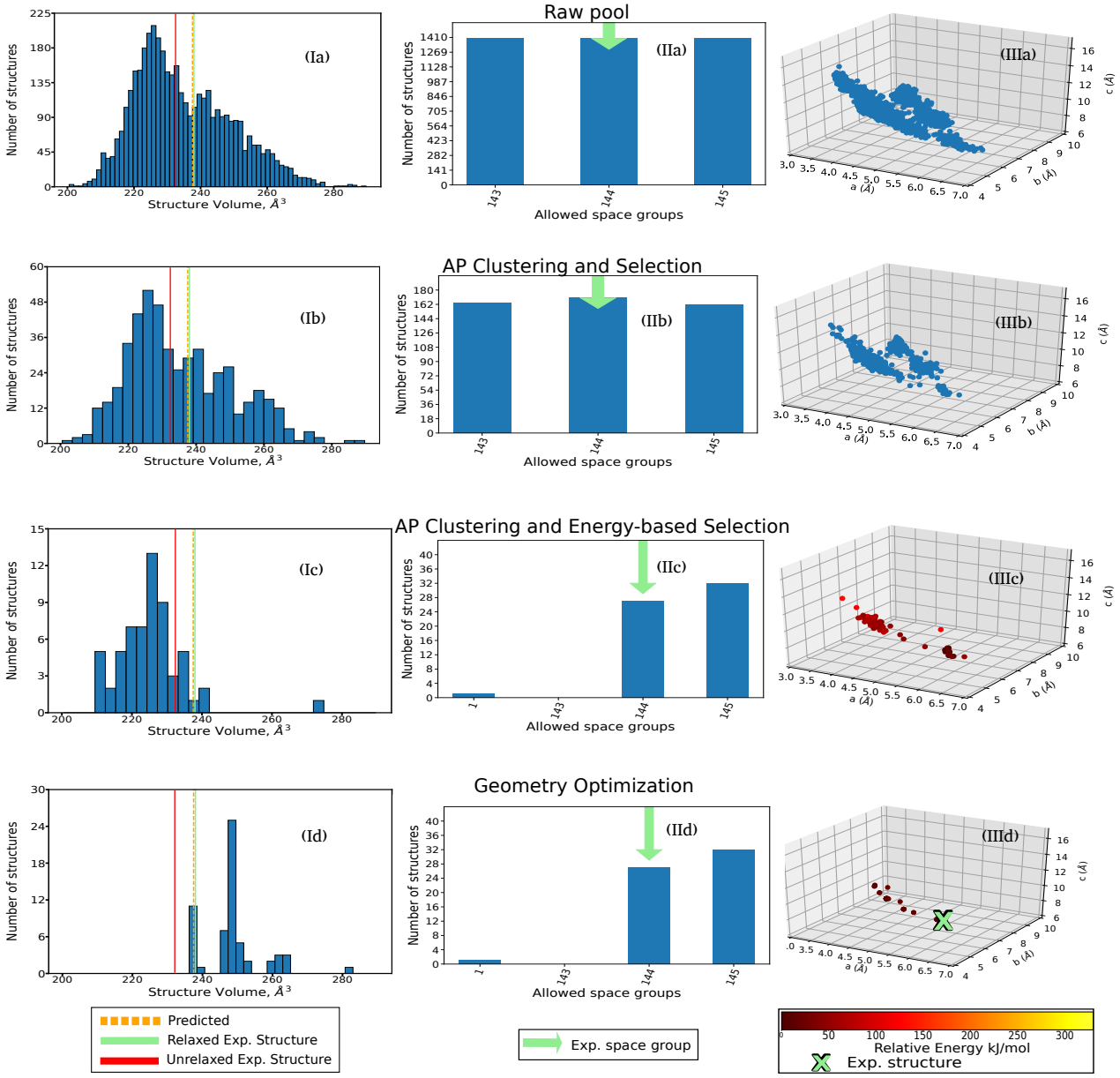

Figure 11: (I) Unit cell volume distributions, (II) space group distributions, and (III) lattice parameter distributions for glycine with $Z = 3$, obtained at each step of the Robust workflow, using RSF-based clustering. On the volume histograms, the solid red line denotes the volume of the experimental structure, the solid green line denotes the unit cell volume of the experimental structure after relaxation, and the dashed orange line shows the volume predicted by our model, as described in Sec.2.2. On the space group distribution histograms, the green arrow points to the space group of the experimental structure, $P3_1$ (144). The green cross in panel (IIId) denotes the (relaxed) experimental structure, which was found in the final pool.

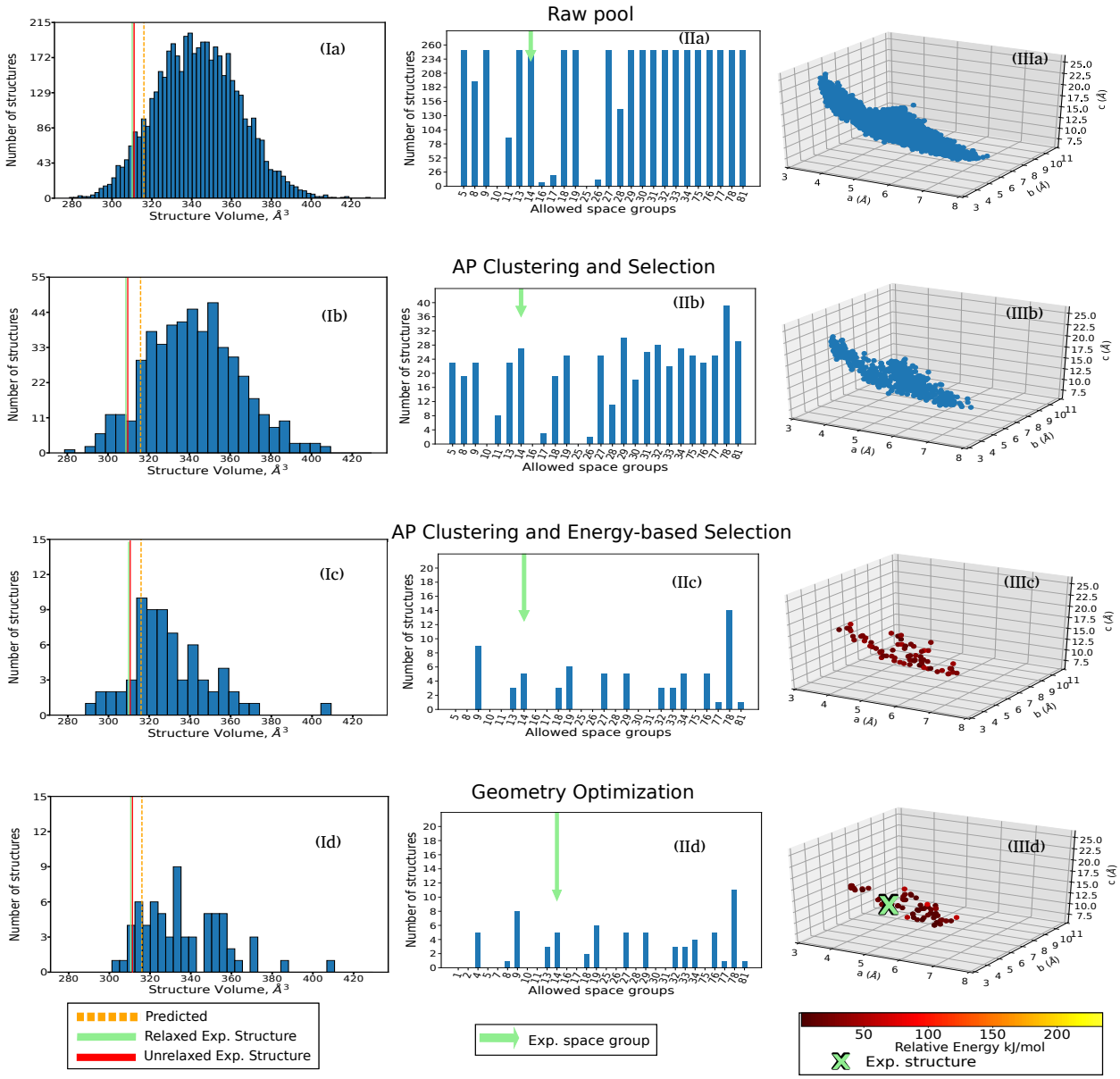

Figure 12: (I) Unit cell volume distributions, (II) space group distributions, and (III) lattice parameter distributions for glycine with $Z = 4$, obtained at each step of the Robust workflow, using RSF-based clustering. On the volume histograms, the solid red line denotes the volume of the experimental structure, the solid green line denotes the unit cell volume of the experimental structure after relaxation, and the dashed orange line shows the volume predicted by our model, as described in Sec.2.2. On the space group distribution histograms, the green arrow points to the space group of the experimental structure, $P2_1/n$ (14). The green cross in panel (IIId) denotes the (relaxed) experimental structure, which was found in the final pool.

12 shows the volume histograms for $Z = 2$, $Z = 3$, and $Z = 4$, respectively, at each step of the Robust workflow, using RSF-based clustering. The results obtained using RCD-based clustering are provided in the Supplemental Information. The relaxed volumes of 158, 238, and 312 $\mathring{A}^3$, for $Z = 2$, $Z = 3$, and $Z = 4$, respectively, indicated by solid green lines, are very close to the experimental volumes of 157, 233, and 310 $\mathring{A}^3$, indicated by solid red lines. The volumes predicted by our machine learned volume estimation model, indicated by dashed orange lines, are close to the experimental values for all polymorphs. About 5000 structures with mean unit cell volume and standard deviation of $(159, 236, 316)$ $\mathring{A}^3$ and $(12, 18, 24)$ $\mathring{A}^3$, respectively, were generated for the $Z = (2, 3, 4)$ polymorphs. The single molecule geometry for structure generation was extracted from the experimental structure for each Z. For $Z = 2$ and $Z = 4$, the mean of the raw pool volume distribution is larger than the predicted volume, whereas for $Z = 3$ the mean is closer to the predicted volume. This is because the $Z = 3$ is easier to generate as two out of the three space groups that are allowed have screw axes. The new settings for hydrogen-bonded systems helped generate dense structures that are close to the predicted volume. Panels (Ic) and (Id) in Figures 10, 11, and 12 show that energy-based selection and the final relaxation favor structures near the experimental volume.

Column II in Figures 10, 11, and 12 shows the space group distribution for each step of the Robust workflow for glycine with $Z = 2$, $Z = 3$, and $Z = 4$, respectively, using RSF-based clustering. The results obtained using RCD-based clustering are provided in the Supplemental Information. The raw pools for all cases show almost uniform space group distribution. For $Z = 4$, space groups $P2/m$ (10) and $Pmm2$ (25) are missing because they contain mirror planes that are hard to generate [39]. There are a significant number of structures in the experimental space group in the raw pool and subsequently selected pools for all cases. Relaxation of the final pool may break existing symmetries or create new ones as there are no constraints imposed. This resulted in additional space groups with a different Z or Z'. For example, space group $Cmc2_1$ (36) and space group $P1$ were created after geometry optimization for glycine with $Z = 2$, as shown in panels (IIc) and (IId) of Figure 10.

Column III in Figures 10, 11, and 12 shows the lattice parameter distributions obtained at each stage of the Robust workflow for glycine with $Z = 2$, 3, and 4, respectively, using RSF-based clustering. The results obtained using RCD-based clustering are provided in the Supplemental Information. For the energy-based selection and final relaxation steps, the color scale corresponds to relative energies with respect to the lowest energy structure in the final relaxed pool. As shown in Panels (IIIc) and (IIId) of Figures 10, 11, and 12, the range of energy in the relaxed pools for glycine is about $100kJ/mol$ or $1eV$ which is an order of magnitude smaller than the range of energies in the unrelaxed pool. For $Z = 2$ and $Z = 4$, the lattice parameter space is well-sampled and diverse regions are obtained upon down-selection. For $Z = 3$, the generated structures are concentrated in distinct regions of the lattice parameter space because there are only three compatible space groups, all of which are in the hexagonal crystal family. The experimental structures of α, β, and γ glycine were found in $Z = 4$, $Z = 2$, and $Z = 3$ runs, respectively.

5. Conclusion

In summary, we have presented a new version of the molecular crystal random structure generator, Genarris, with several new features and demonstrated its application to benzene and glycine. The new MPI parallelization scheme has made Genarris 2.0 significantly faster than the previous version, more portable, and able to scale better on high performance computing architectures. The new machine learning method for volume estimation has been demonstrated to reliably predict the volumes of the polymorphs of benzene and glycine. The somewhat larger deviation from the experimental volume for the high-pressure polymorph of benzene was expected, considering that the model was trained on crystal structures obtained at ambient pressure.

For all polymorphs of benzene and glycine, the new structure generation function has successfully generated structures in the target volume range with approximately uniform space group distributions and has adequately sampled the possible range of lattice parameters. The new capability to generate structures with molecules occupying special Wyckoff positions has proven to be instrumental for benzene. The updated structure check settings for strong hydrogen bonds have been particularly useful for glycine. Thus, Genarris 2.0 is expected to deliver a significantly better performance than the previous version for symmetric molecules and for molecules capable of forming strong hydrogen bonds.

A new Robust workflow has been implemented for clustering and down-selection of the raw pool of random structures to form a small curated population of low-energy structures with diverse crystal packing motifs. The affinity propagation clustering algorithm performs similarly well based on the RSF and RCD descriptors. Although the Robust workflow is intended for producing an initial population for other structure search algorithms (such as genetic algorithms), not as a structure prediction method, the experimental structures of both polymorphs of benzene and of the three forms of glycine were found in the final relaxed pools.

Genarris 2.0 offers the user full flexibility to design and easily implement new workflows by sequentially executing a user-defined list of procedures. For example, to generate datasets for training machine learning models, the user may wish to perform energy evaluations for a larger number of structures from the raw pool. To perform crystal structure prediction, the user may wish to fully relax a larger number of structures and to re-rank them with more accurate methods. Thus, Genarris 2.0 is a useful random structure generator for homomolecular crystals of semi-rigid molecules with no rotatable bonds, which can be applied to generate initial populations for structure search algorithms or to generate datasets for machine learning or as a standalone crystal structure prediction method.

Acknowledgements

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 and of the National Energy Research Scientific Computing Center (NERSC), a U.S. Department of Energy Office of Science User Facility operated under Contract No. DE-AC02-05CH11231. Dr. William Paul Huhn from ALCF is thanked for his help with compiling and importing FHI-aims as a Python library.

References

[1] U. J. Griesser, R. K. Jetti, M. F. Haddow, T. Brehmer, D. C. Apperley, A. King, R. K. Harris, Cryst. Growth Des. 8 (2008) 44–56.
[2] J. Nyman, G. M. Day, CrystEngComm 17 (2015) 5154–5165.
[3] J. Hoja, A. Tkatchenko, Faraday Discuss. 211 (2018) 253–274.
[4] V. Coropceanu, J. Cornil, D. A. da Silva Filho, Y. Olivier, R. Silbey, J.-L. Brédas, Chem. Rev. 107 (2007) 926–952.
[5] O. D. Jurchescu, D. A. Mourey, S. Subramanian, S. R. Parkin, B. M. Vogel, J. E. Anthony, T. N. Jackson, D. J. Gundlach, Phys. Rev. B 80 (2009) 085201.
[6] T. Matsukawa, M. Yoshimura, M. Uchiyama, M. Yamagishi, A. Nakao, Y. Takahashi, J. Takeya, Y. Kitaoka, Y. Mori, T. Sasaki, Jpn. J. Appl. Phys. 49 (2010) 085502.
[7] Y. Diao, K. M. Lenn, W.-Y. Lee, M. A. Blood-Forsythe, J. Xu, Y. Mao, Y. Kim, J. A. Reinspach, S. Park, A. Aspuru-Guzik, et al., J. Am. Chem. Soc. 136 (2014) 17046–17057.
[8] J. Yang, S. De, J. E. Campbell, S. Li, M. Ceriotti, G. M. Day, Chem. Mater. 30 (2018) 4361–4371.
[9] J. Bauer, S. Spanton, R. Henry, J. Quick, W. Dziki, W. Porter, J. Morris, Pharm. Res. 18 (2001) 859–866.
[10] D.-K. Bučar, R. W. Lancaster, J. Bernstein, Angew. Chem. Int. Ed. 54 (2015) 6972–6993.
[11] A. M. Reilly, A. Tkatchenko, Chem. Sci. 6 (2015) 3289–3301.
[12] J. Hermann, R. A. DiStasio Jr, A. Tkatchenko, Chem. Rev. 117 (2017) 4714–4758.
[13] G. P. Stahly, Cryst. Growth Des. 7 (2007) 1007–1026.
[14] A. Y. Lee, D. Erdemir, A. S. Myerson, Annu. Rev. Chem. Biomol. Eng. 2 (2011) 259–280.
[15] A. N. Sokolov, S. Atahan-Evrenk, R. Mondal, H. B. Akkerman, R. S. Sánchez-Carrera, S. Granados-Focil, J. Schrier, S. C. Mannsfeld, A. P. Zoombelt, Z. Bao, et al., Nat. Commun. 2 (2011) 437.
[16] H. Chung, Y. Diao, J. Mater. Chem. C 4 (2016) 3915–3933.
[17] J. P. Lommerse, W. S. Motherwell, H. L. Ammon, J. D. Dunitz, A. Gavezzotti, D. W. Hofmann, F. J. Leusen, W. T. Mooij, S. L. Price, B. Schweizer, et al., Acta Crystallogr. B 56 (2000) 697–714.
[18] W. S. Motherwell, H. L. Ammon, J. D. Dunitz, A. Dzyabchenko, P. Erk, A. Gavezzotti, D. W. Hofmann, F. J. Leusen, J. P. Lommerse, W. T. Mooij, et al., Acta Crystallogr. B 58 (2002) 647–661.
[19] G. Day, W. Motherwell, H. Ammon, S. Boerrigter, R. Della Valle, E. Venuti, A. Dzyabchenko, J. D. Dunitz, B. Schweizer, B. Van Eijck, et al., Acta Crystallogr. B 61 (2005) 511–527.
[20] G. M. Day, T. G. Cooper, A. J. Cruz-Cabeza, K. E. Hejczyk, H. L. Ammon, S. X. Boerrigter, J. S. Tan, R. G. Della Valle, E. Venuti, J. Jose, et al., Acta Crystallogr. B 65 (2009) 107–125.
[21] D. A. Bardwell, C. S. Adjiman, Y. A. Arnautova, E. Bartashevich, S. X. Boerrigter, D. E. Braun, A. J. Cruz-Cabeza, G. M. Day, R. G. Della Valle, G. R. Desiraju, et al., Acta Crystallogr. B 67 (2011) 535–551.
[22] A. M. Reilly, R. I. Cooper, C. S. Adjiman, S. Bhattacharya, A. D. Boese, J. G. Brandenburg, P. J. Bygrave, R. Bylsma, J. E. Campbell, R. Car, et al., Acta Crystallogr. B 72 (2016) 439–459.
[23] S. L. Price, Chem. Soc. Rev. 43 (2014) 2098–2111.
[24] G. M. Day, Crystallogr. Rev. 17 (2011) 3–52.
[25] D. H. Case, J. E. Campbell, P. J. Bygrave, G. M. Day, J. Chem. Theory Comput. 12 (2016) 910–924.
[26] C. J. Pickard, R. Needs, J. Phys. Condens. Matter 23 (2011) 053201.
[27] P. G. Karamertzanis, C. C. Pantelides, J. Comput. Chem. 26 (2005) 304–324.
[28] P. Karamertzanis, C. Pantelides, Mol. Phys. 105 (2007) 273–291.

[29] B. P. Van Eijck, J. Kroon, Acta Crystallogr. B 56 (2000) 535–542.
[30] H. Wonderatschek, U. Müller, International Tables for Crystallography: Volume A1: Symmetry Relations Between Space Groups, Springer, 2004.
[31] W. T. Mooij, B. P. van Eijck, J. Kroon, J. Phys. Chem. A 103 (1999) 9883–9890.
[32] S. L. Price, M. Leslie, G. W. Welch, M. Habgood, L. S. Price, P. G. Karamertzanis, G. M. Day, Phys. Chem. Chem. Phys 12 (2010) 8478–8490.
[33] P. Zhang, G. P. Wood, J. Ma, M. Yang, Y. Liu, G. Sun, Y. A. Jiang, B. C. Hancock, S. Wen, Cryst. Growth Des. 18 (2018) 6891–6900.
[34] X. Li, F. S. Curtis, T. Rose, C. Schober, A. Vazquez-Mayagoitia, K. Reuter, H. Oberhofer, N. Marom, J. Chem. Phys. 148 (2018) 241701.
[35] M. Zilka, D. V. Dudenko, C. E. Hughes, P. A. Williams, S. Sturniolo, W. T. Franks, C. J. Pickard, J. R. Yates, K. D. Harris, S. P. Brown, Phys. Chem. Chem. Phys 19 (2017) 25949–25960.
[36] F. H. Allen, Acta Crystallogr. B 58 (2002) 380–388.
[37] C. R. Groom, I. J. Bruno, M. P. Lightfoot, S. C. Ward, Acta Crystallogr. B 72 (2016) 171–179.
[38] E. Pidcock, W. S. Motherwell, Cryst. Growth Des. 4 (2004) 611–620.
[39] E. Pidcock, W. S. Motherwell, J. C. Cole, Acta Crystallogr. B 59 (2003) 634–640.
[40] E. Paquet, H. L. Viktor, BioMed research international 2015 (2015).
[41] D. J. Earl, M. W. Deem, Physical Chemistry Chemical Physics 7 (2005) 3910–3916.
[42] A. R. Oganov, C. W. Glass, The Journal of chemical physics 124 (2006) 244704.
[43] S. Kim, A. M. Orendt, M. B. Ferraro, J. C. Facelli, Journal of computational chemistry 30 (2009) 1973–1985.
[44] F. Curtis, X. Li, T. Rose, A. Vazquez-Mayagoitia, S. Bhattacharya, L. M. Ghiringhelli, N. Marom, J. Chem. Theory Comput. 14 (2018) 2246–2264.
[45] L. Dalcín, R. Paz, M. Storti, J. DElía, J. Parallel Distrib. Comput. 68 (2008) 655–662.
[46] V. Blum, R. Gehrke, F. Hanke, P. Havu, V. Havu, X. Ren, K. Reuter, M. Scheffler, Comput. Phys. Commun. 180 (2009) 2175–2196.
[47] F. Curtis, T. Rose, N. Marom, Faraday discussions 211 (2018) 61–77.
[48] B. J. Frey, D. Dueck, Science 315 (2007) 972–976.
[49] J. Behler, M. Parrinello, Phys. Rev. Lett. 98 (2007) 146401.
[50] I. J. Bruno, J. C. Cole, P. R. Edgington, M. Kessler, C. F. Macrae, P. McCabe, J. Pearson, R. Taylor, Acta Crystallogr. B 58 (2002) 389–397.
[51] J. Van De Streek, S. Motherwell, Acta Crystallogr. B 61 (2005) 504–510.
[52] A. J. Cruz-Cabeza, S. M. Reutzel-Edens, J. Bernstein, Chem. Soc. Rev. 44 (2015) 8619–8635.
[53] K. Kersten, R. Kaur, A. Matzger, IUCrJ 5 (2018) 124–129.
[54] J. A. Chisholm, S. Motherwell, J. Appl. Crystallogr. 38 (2005) 228–231.
[55] A. Burger, R. Ramberger, Microchim. Acta 72 (1979) 273–316.
[56] H. L. Ammon, S. Mitchell, Propellants Explos. Pyrotech. 23 (1998) 260–265.
[57] D. W. Hofmann, Acta Crystallogr. B 58 (2002) 489–493.
[58] C. Ye, J. M. Shreeve, J. Chem. Eng. Data 53 (2008) 520–524.
[59] A. Bondi, J. Phys. Chem. 68 (1964) 441–451.
[60] A. Gavezzotti, J. Am. Chem. Soc. 105 (1983) 5220–5225.
[61] L. F. Pacios, Comput. Biol. Chem. 18 (1994) 377–385.
[62] G. Maggiora, M. Vogt, D. Stumpfe, J. Bajorath, J. Med. Chem. 57 (2013) 3186–3204.
[63] D. Rogers, M. Hahn, J. Chem. Inf. Model. 50 (2010) 742–754.
[64] K. Hansen, F. Biegler, R. Ramakrishnan, W. Pronobis, O. A. Von Lilienfeld, K.-R. Muller, A. Tkatchenko, J. Phys. Chem. Lett. 6 (2015) 2326–2331.
[65] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, et al., J. Mach. Learn. Res. 12 (2011) 2825–2830.
[66] T. Hahn, U. Shmueli, J. W. Arthur, International tables for crystallography, volume 1, Reidel Dordrecht, 1983.
[67] E. Agrell, T. Eriksson, A. Vardy, K. Zeger, IEEE Trans. Inf. Theory 48 (2002) 2201–2214.

doi:10.1109/TIT.2002.800499.
[68] D. M. Rogers, J. Mol. Graph. Model. 68 (2016) 197–205.
[69] U. Fincke, M. Pohst, Math. Comput. 44 (1985) 463–471.
[70] B. Hassibi, H. Vikalo, IEEE Trans. Signal Process. 53 (2005) 2806–2818.
[71] R. S. Rowland, R. Taylor, J. Phys. Chem. 100 (1996) 7384–7391.
[72] T. Steiner, Angew. Chem. 41 (2002) 48–76.
[73] G. Gilli, P. Gilli, The nature of the hydrogen bond: outline of a comprehensive hydrogen bond theory, volume 23, Oxford University Press, 2009.
[74] I. J. Bruno, J. C. Cole, J. P. Lommerse, R. S. Rowland, R. Taylor, M. L. Verdonk, J. Comput. Aided Mol. Des. 11 (1997) 525–537.
[75] P. Gilli, L. Pretto, V. Bertolasi, G. Gilli, Acc. Chem. Res. 42 (2008) 33–44.
[76] L. Brammer, E. A. Bruton, P. Sherwood, Cryst. Growth Des. 1 (2001) 277–290.
[77] T. Steiner, Crystallogr. Rev. 6 (1996) 1–51.
[78] M. Gastegger, L. Schwiedrzik, M. Bittermann, F. Berzsenyi, P. Marquetand, J. Chem. Phys. 148 (2018) 241709.
[79] A. Khorshidi, A. A. Peterson, Comput. Phys. Commun. 207 (2016) 310–324.
[80] J. P. Perdew, K. Burke, M. Ernzerhof, Phys. Rev. Lett. 77 (1996) 3865.
[81] A. Tkatchenko, M. Scheffler, Phys. Rev. Lett. 102 (2009) 073005.
[82] L. Ciabini, M. Santoro, F. A. Gorelli, R. Bini, V. Schettino, S. Raugei, Nat. Mater. 6 (2007) 39.
[83] E. Boldyreva, T. Drebushchak, E. Shutova, Z. Kristallogr. Cryst. Mater. 218 (2003).
[84] E. Boldyreva, V. Drebushchak, T. Drebushchak, I. Paukov, Y. A. Kovalevskaya, E. Shutova, J. Therm. Anal. Calorim. 73 (2003) 409–418.
[85] J. A. Chisholm, S. Motherwell, P. R. Tulip, S. Parsons, S. J. Clark, Cryst. Growth Des. 5 (2005) 1437–1442.
[86] G. M. Day, W. S. Motherwell, W. Jones, Cryst. Growth Des. 5 (2005) 1023–1033.
[87] N. Marom, R. A. DiStasio Jr, V. Atalla, S. Levchenko, A. M. Reilly, J. R. Chelikowsky, L. Leiserowitz, A. Tkatchenko, Angew. Chem. Int. Ed. 52 (2013) 6629–6632.
[88] Q. Zhu, A. R. Oganov, C. W. Glass, H. T. Stokes, Acta Crystallogr. B 68 (2012) 215–226.
[89] A. M. Lund, G. I. Pagola, A. M. Orendt, M. B. Ferraro, J. C. Facelli, Chem. Phys. Lett 626 (2015) 20–24.
[90] R. Sabatini, E. Küçükbenli, B. Kolb, T. Thonhauser, S. De Gironcoli, J. Phys. Condens. Matter 24 (2012) 424209.
[91] J. Rodríguez, G. Costa, M. da Silva, B. Silva, L. Honorio, P. de Lima-Neto, R. Santos, E. Caetano, H. Alves, V. Freire, Cryst. Growth Des. (2019).
[92] A. Ranganathan, G. Kulkarni, C. Rao, J. Mol. Struct. 656 (2003) 249–263.