RSC Publishing Faraday Discussions

Evolutionary Niching in the GAtor Genetic Algorithm for Molecular Crystal Structure Prediction

Journal:	*Faraday Discussions*
Manuscript ID	FD-ART-03-2018-000067.R1
Article Type:	Paper
Date Submitted by the Author:	31-Mar-2018
Complete List of Authors:	Curtis, Farren ; Carnegie Mellon University Ros, Timothy; Carnegie Mellon University Marom, Noa; Carnegie Mellon University

Journal Name

ARTICLE TYPE

Cite this: DOI: 10.1039/xxxxxxxxxx

Evolutionary Niching in the GAtor Genetic Algorithm for Molecular Crystal Structure Prediction†

Farren Curtis,[ab] Timothy Rose,[a] and Noa Marom*[abc]

Received Date
Accepted Date

DOI: 10.1039/xxxxxxxxxx

www.rsc.org/journalname

The goal of molecular crystal structure prediction (CSP) is to find all plausible polymorphs for a given molecule. This requires performing global optimization over a high dimensional search space. Genetic algorithms (GAs) perform global optimization by starting from an initial population of structures and generating new candidate structures by breeding the fittest structures in the population. Typically, the fitness function is based on relative lattice energies, such that structures with lower energies have a higher probability of being selected for mating. GAs may be adapted to perform multi-modal optimization by using evolutionary niching methods that support the formation of several stable subpopulations and suppress the over-sampling of densely populated regions. Evolutionary niching is implemented in the GAtor molecular crystal structure prediction code by using techniques from machine learning to dynamically cluster the population into niches of structural similarity. A cluster-based fitness function is constructed such that structures in less populated clusters have a higher probability of being selected for breeding. Here, the effects of evolutionary niching are investigated for the crystal structure prediction of 1,3-dibromo-2-chloro-5-fluorobenzene. Using the cluster-based fitness function increases the success rate of generating the experimental structure and additional low-energy structures with similar packing motifs.

1 Introduction

Molecular crystals have numerous applications in pharmaceuticals, organic electronics, pigments, and explosives.[1–11] Hence, there has been increasing interest in reliable crystal structure prediction (CSP) methods. CSP methods computationally explore various crystal packing arrangements of a given organic compound, aiming to generate any experimentally determined crystal structure(s) in addition to any potential low-energy polymorphs that may be possible to synthesize. CSP methods are increasingly being used to complement experimental investigations of molecular crystal polymorphs.[12–16]

A robust CSP method must be able to produce the global minimum structure in addition to other structures that are close in energy, as molecular crystal polymorphs are typically within a few kJ/mol.[17–20] The global optimization method used needs to efficiently search an enormous configuration space with many local minima that are very close in energy. Genetic algorithms (GAs) are a versatile class of global optimization methods inspired by the evolutionary principle of survival of the fittest.[21–23] GAs are suitable for organic CSP because they can handle systems with complex multidimensional search spaces, including those with many extrema. A GA starts from an initial population of locally optimized structures and proceeds by repeatedly mating the fittest structures in the population through crossover and mutation operators until a convergence criterion is reached. Typically the GA fitness function is based on the relative energy of structures in the population. In this scenario, breeding operators drive the evolutionary search by exploiting the structural motifs associated with lower total energies. GAs have been used extensively for CSP of crystalline solids[24–36] and clusters.[21–23,37–46]

One drawback of GAs is that they may be prone to 'genetic drift', meaning that the algorithm over-samples certain regions of the potential energy surface associated with the fittest structures in the population, while under-sampling other regions. The origin of genetic drift may be related to the nature of the potential energy surface, with some stable crystal packing motifs found in wide basins, whereas others are found in isolated funnels that are rarely sampled.[47] Systematic biases of the total energy method towards or against particular packing motifs may also contribute to genetic drift.[48] Another issue of GAs is that the composition of

[a] Department of Materials Science and Engineering, Carnegie Mellon University, Pittsburgh, Pennsylvania 15213, USA. E-mail: nmarom@andrew.cmu.edu
[b] Department of Physics, Carnegie Mellon University, Pittsburgh, Pennsylvania 15213, USA.
[c] Department of Chemistry, Carnegie Mellon University, Pittsburgh, Pennsylvania 15213, USA.

† Electronic Supplementary Information (ESI) available. See DOI: 10.1039/b000000x/

the initial population may bias the outcome of the search.[44]

To overcome these issues, GAs can be adapted to perform multimodal optimization by incorporating evolutionary niching methods. These methods aim to increase diversity in the population and converge several promising solutions simultaneously. In order to identify niches, i.e. clusters comprised of structures with similar geometric features, unsupervised clustering techniques from machine learning may be employed to influence the GA search strategy as new data is accumulated.[45,47] Clustering can help identify the structural motifs that are over or under represented in the population. This information can be used to modify the fitness function and/or selection strategy.

Recently, we have presented the GAtor genetic algorithm for molecular crystal structure prediction from first principles.[47] GAtor offers the option to perform evolutionary niching by dynamically clustering the population by structural similarity and then employing a cluster-based fitness function. In Ref. 47, we have demonstrated the effectiveness of evolutionary niching for generating the experimental structure of tricyano-1,4-dithiino[c]-isothiazole (Target XXII from the sixth CSP blind test[49] organized by the Cambridge Crystallographic Data Centre (CCDC)). Therein, clustering was performed with respect to a descriptor based on the lattice parameters. Here, we further investigate the effect of using cluster-based fitness functions based on different molecular crystal descriptors versus a traditional energy-based fitness function. As an example, we have chosen 1,3-dibromo-2-chloro-5-fluorobenzene (Target XIII from the fourth CCDC blind test[50]), illustrated in Fig. 1, which contains several halogen elements, namely Cl, F, and Br. The theoretical description of halogen bonds is challenging because it requires an accurate treatment of both electrostatic and dispersion interactions.[51–54] In Ref. 47, the experimental structure of Target XIII was rarely generated when using a traditional energy-based fitness function, making this molecule a good test case for exploring the effect of incorporating evolutionary niching via a modified cluster-based fitness function. In the following, we show that the use of evolutionary niching provides uniform sampling of the potential energy surface by evolving several subpopulations simultaneously. This helps find the experimental structure of Target XIII and several additional low-energy structures that are not otherwise generated in control runs that employ an energy-based fitness function.

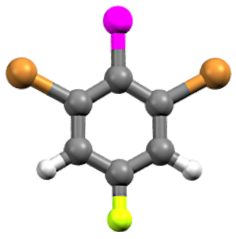

Fig. 1 The molecular diagram of 1,3-dibromo-2-chloro-5-fluorobenzene (Target XIII)[50]. C, H, Br, Cl, and F atoms are colored in grey, white, brown, pink, and yellow, respectively.

2 Methodology

To study the effect of evolutionary niching, a set of GA runs were conducted starting from two initial pools as described in Section 2.1. Section 2.2 describes the DFT settings used for energy evaluations and local optimizations performed in GAtor. For each initial pool, the GA was run four times with the settings described in Section 2.3. One run was used as a control, which employed a traditional energy-based fitness function. The other three runs used evolutionary niching by employing a cluster-based fitness function with three different molecular crystal descriptors, described in Section 2.4. The effect of using a cluster-based fitness function on the final population of structures produced was analyzed for each run and descriptor.

2.1 Initial Population

Two initial pools consisting of 45 molecular crystal structures of Target XIII (Z=4) were prepared using the Random and Diverse workflows of the molecular crystal structure generation package Genarris,[55] previously generated as reported in Ref. 55. Briefly, a raw pool of 5,000 structures was generated in all space groups compatible with Z=4. For the random initial pool, the final structures were randomly selected from the raw pool. For the diverse initial pool, fragment-based density functional theory (DFT) and clustering techniques from machine learning were used for the selection of the final structures. For a more detailed explanation of the preparation of the initial pools, see Ref. 55. Both initial pools were locally optimized using the DFT settings presented in Section 2.2.

2.2 DFT Settings

For energy evaluations and local structural optimizations within the GA, the generalized gradient approximation of Perdew-Burke-Ernzerhof (PBE)[56,57] is used with the pairwise Tkatchenko-Scheffler (TS) dispersion-correction[58] with *lower-level* numerical settings, which correspond to the tier 1 basis sets and light numerical settings of FHI-aims.[59] For all calculations, a $3 \times 3 \times 3$ k-point grid is used. During local optimization, the space group symmetry is allowed to vary. Atomic ZORA scalar relativity[59] settings are used for the heavier halogen elements. For postprocessing, the top structures produced 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 tier 2 basis sets and tight numerical settings of FHI-aims.[59]

2.3 GA Settings

Four different GA runs were performed starting from each initial pool using GAtor.[47] For all runs, the GA was terminated when the final population reached a total of 385 structures including the initial pool. All runs used 50% crossover and 50% mutation and tournament selection with a tournament size of 20 structures. For closeness checks, the minimum distance between two atoms of different molecules was set to 0.7 of the sum of their van der Waals radii. A relative energy cutoff of 0.75 eV per molecule was used for single-point energy (SPE) evaluations, such that struc-

tures generated with energies per molecule greater than the current global minimum plus 0.75 eV were rejected without performing local optimization.

For each initial pool, a control run was conducted using a traditional energy-based fitness function. In this fitness scheme, the total energy E_i of the ith structure in the population is evaluated using dispersion-inclusive DFT as detailed in Section 2.2. The energy-based fitness f_i of each structure is given by,

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

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

where ε_i is the ith structure's relative energy, and $E_{\max}$ and $E_{\min}$ correspond to the structures with the dynamically updated highest and lowest total energies in the population, respectively.[43,44,47,60] Hence, in this fitness scheme structures with lower relative energies have higher fitness values.

For each initial pool, three additional GA runs were conducted using a cluster-based fitness function. In this scheme, the population is dynamically-clustered into niches of structural similarity using different molecular crystal descriptors (See Section 2.4). Then, a fitness sharing scheme is implemented such that the cluster-based fitness f_i^c of each structure is given by

$$f_i^c = \frac{\bar{f}_i}{\bar{f}_{\max}} \qquad 0 \leq f^c \leq 1 \tag{3}$$

$$\bar{f}_i = \frac{f_i}{m_i} \tag{4}$$

where m_i is each structure's niche count given by the number of structures in each individual's cluster and f_i is the ith structure's energy-based fitness. With this fitness scheme, structures in less populated clusters with under-represented structural motifs have higher fitness values, and hence a higher probability of being selected for mating. Penalizing the fitness of over-sampled clusters and steering the GA towards under-sampled clusters provides a more uniform sampling of the configuration space. If all cluster sizes were equal, this fitness scheme would equivalent to energy-based fitness.

2.4 Clustering

When using GAtor's cluster-based fitness function, every time a new structure is added to the common pool of structures, clustering is performed to group the common population into niches. As the GA evolves, the clusters are updated automatically, affecting each structure's cluster-based fitness value. For clustering, the affinity propagation (AP)[61] clustering algorithm is used, as implemented in scikit-learn.[62] AP detects the number of clusters in a data set as opposed to pre-specifying them *a priori*. The input of AP is a pairwise similarity matrix between data points. All data points are initially considered as possible "exemplars", i.e., points that are representative members of each cluster. The algorithm iteratively refines the candidate exemplars until well-defined exemplars and corresponding clusters are identified. For a more detailed description of the AP algorithm, see Ref. 61. AP clustering has been shown to successfully resolve small isolated clusters with distinct structural motifs.[55]

We used three different molecular crystal descriptors for clustering. The first descriptor L is given by[47]

$$\mathrm{L} = \frac{1}{\sqrt[3]{V}}(a,b,c) \tag{5}$$

where V is the unit cell volume and a, b, and c are the structure's lattice parameters after employing Niggli reduction[63–66] and unit cell standardization. All unit cells 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. For the lattice parameter based descriptor, a negative squared euclidean metric is used for construction of the pairwise similarity matrix for AP clustering.

The second descriptor combines several radial distribution function (RDF) vectors of selected interatomic contacts. The radial distribution function for atom types A and B is given by

$$G_{AB}(r) = \frac{\sum_{ij} \exp(-(r - r_{ij})^2)}{N_A} \tag{6}$$

where r_{ij} is the distance between the ith and jth atom of atom types A and B, respectively, and N_A is the number of A atoms. RDF vectors for atom types A and B are constructed by sampling $G_{AB}(r)$ at finite intervals in a user-defined range. Multiple RDF vectors of different atom type pairs are concatenated to form a combined RDF descriptor. For Target XIII, $G_{AB}(r)$ is sampled for interatomic Br$\cdots$Br, H$\cdots$Br, H$\cdots$F, and H$\cdots$Cl contacts in bins of 1 Å for 1 Å $\leq r \leq$ 8 Å. For the RDF descriptor, a negative squared euclidean metric is used for construction of the pairwise similarity matrix for AP clustering.

The third molecular crystal descriptor is the relative coordinate descriptor (RCD).[55] RCD was developed to capture the packing motif of molecular crystals by using the relative center of mass position $\vec{P}$ and orientation $\vec{Q}$ of N (16 by default) neighboring molecules with respect to a reference molecule. The relative orientation $\vec{Q}$ between two molecules is computed by taking the dot product of orthogonal reference axes centered on each molecule, constructed as described in Ref. 55. The RCD descriptor for a given molecular crystal is given by

$$\mathrm{RCD} = [(\vec{P}^1, \vec{Q}^1), ..., (\vec{P}^N, \vec{Q}^N)]. \tag{7}$$

AP clustering allows the user to input custom similarity metrics that determine how similar two input vectors are to one another. When using the RCD descriptor with AP clustering in GAtor, the distance matrix between two RCD vectors is given by

$$D_{i,j} = \left(\frac{|\vec{P}_1^{\,i} - \vec{P}_2^{\,j}|^2}{|\vec{P}_1^{i}||\vec{P}_2^{j}|} \right) + \frac{1}{3}\left(|\vec{Q}_1^{\,i} - \vec{Q}_2^{\,j}|^2 \right). \tag{8}$$

The M (8 by default) smallest entries of D are selected, such that no two entries have the same value for i or j. The sum of these M entries serves as the input pairwise similarity for the the AP clustering algorithm.

3 Results and Discussion

For each initial pool GAtor was run four times. One of the four runs employed a traditional energy-based fitness function, while the other three runs used a cluster-based fitness function with the lattice parameter based descriptor L, the combined RDF descriptor, and the RCD descriptor. Panel (a) of Fig. 2 shows the average energy of the common population as a function of GA iteration for all eight runs, scaled with respect to the total energy of the global minimum structure. Here an iteration corresponds to when a single structure has been added to the common population. The average energy of the initial population corresponds to iteration 0. The common populations produced from the random (R) and diverse (D) energy-based control runs, shown in light and dark red, respectively, display the lowest average energy (approximately 9 kJ/mol per molecule) when the GA was terminated. The run that used the random initial pool and the cluster-based fitness function with the RDF descriptor, shown in light blue, has an average energy slightly above 9 kJ/mol per molecule when the GA was terminated. All other runs that used a cluster-based fitness function show average energies 1-2 kJ/mol per molecule higher than the energy-based fitness runs for the last 200 iterations. This behavior may be attributed to the fact that the use of a cluster-based fitness function promotes the generation of structures with under-represented structural motifs, which may have higher energies. Panel (b) of Fig. 2. shows the minimum energy structure as a function of GA iteration. The experimental structure (corresponding to the PBE+TS global minimum energy) and the second-lowest energy structure are displayed. By the time the GA was terminated, five of the six runs that employed a cluster-based fitness function generated the experimental structure, the exception being the run that used the diverse initial pool and the lattice parameter based descriptor. Neither of the control runs generated the experimental structure, but both generated the second-lowest energy structure within the first 50 GA iterations. The layered packing motif of the second-lowest energy structure is prevalent in the low-energy structures generated by the control runs.[47]

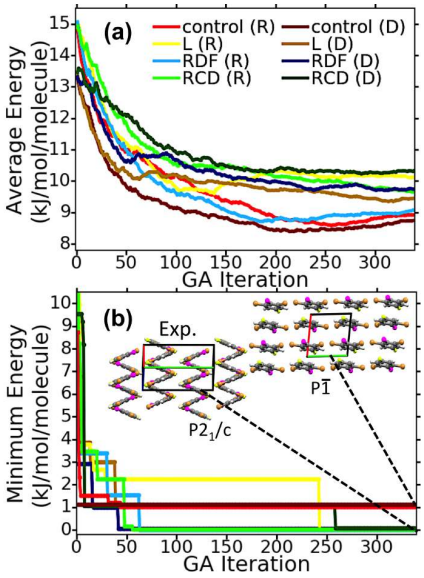

Fig. 2 (a) The average energy of the common population and (b) the global minimum structure produced as a function of GA iteration for different GA runs. For the displayed crystal structures, the $\vec{a}$, $\vec{b}$, and $\vec{c}$ crystallographic lattice vectors are colored in red, green, and blue, respectively.

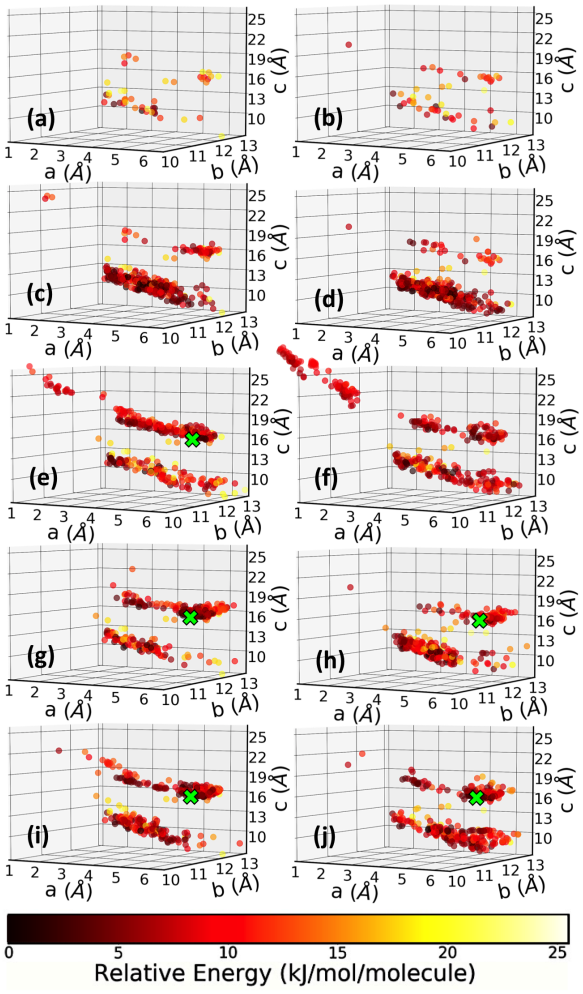

Fig. 3 Structures of Target XIII produced from the evolution of the random (left) and diverse (right) initial pools. Structures are represented on a 3D space consisting of each structure's niggli-reduced lattice parameters a, b, and c and colored according to their energy relative to the global minimum. The distribution of the initial population structures are shown in panels (a-b). The structures produced by the GA runs that used the energy-based fitness function are shown in (c-d). The structures produced by the GA runs that used cluster-based fitness functions with L, RDF, and RCD are shown in panels (e-f), (g-h), and (i-j), respectively. The location of the experimental structure, if generated in a given run, is indicated by a green 'X'.

Fig. 3. demonstrates the effect of evolutionary niching by visualizing the structures generated by each GA run in a 3D space consisting of each structure's unique, Niggli-reduced a, b, and c lattice parameters. All points are colored by their total PBE+TS energy (with *lower-level* numerical settings) relative to the global minimum. For the runs that generated the experimental structure, its location is indicated by a green 'X'. Panels (a) and (b) show the random and diverse initial populations, respectively. The di-

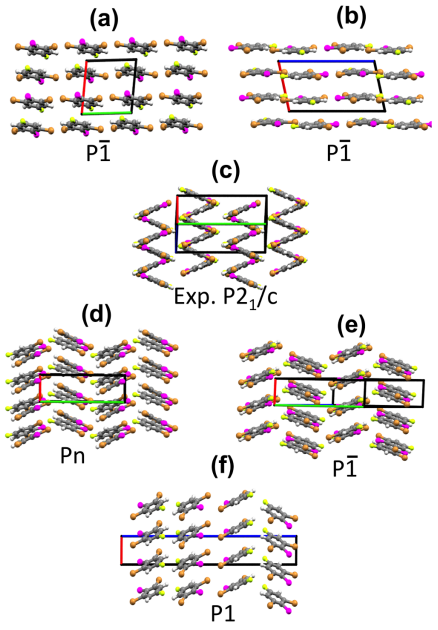

Fig. 4 Structures representative of (a-b) layered packing motifs (c) zig-zag packing motifs (d-e) herringbone packing motifs and (f) large c-parameters. The $\vec{a}$, $\vec{b}$, and $\vec{c}$ crystallographic lattice vectors are colored in red, green, and blue, respectively.

verse initial pool displays structures with a greater variety of lattice parameters than the random pool, as expected. Panels (c) and (d) depict the structures produced by the GA runs that used the energy-based fitness function with the random and diverse initial pools, respectively. Both of these control runs exhaustively explored the low-lying region of structures that had c-parameters with a maximum value of approximately 13 Å, a region that contains many low-energy structures with layered packing motifs but does not contain the experimental structure. Structures representative of this region are shown in Fig. 4, panels (a) and (b). Although the control run that used the diverse initial pool had more low-energy structures with a variety of lattice parameters to start with, over time it thoroughly sampled the same regions of the PES as the control run that used the random initial pool. This is an example of genetic drift, where the search is biased towards exhaustively exploring particular basins of the potential energy surface, which may or may not contain the experimental structure and/or other desirable low-energy crystal structures. Panels (e-f), (g-h), and (i-j) of Fig. 3 show the final populations of structures produced by the GA runs that used cluster-based fitness with the lattice parameter descriptor, the combined RDF descriptor, and the RCD descriptor, respectively. All of the runs that used a cluster-based fitness function explored the region containing the experimental structure more frequently than the control runs, with five out of the six runs successfully generating the experimental structure. The region containing the experimental structure contains numerous structures with zig-zag and herringbone packing motifs and elongated unit cells, examples of which are shown in Fig. 4, panels (c-e). The runs that used the cluster-based fitness function with the lattice parameter descriptor, shown in Fig. 3, panels (e-f), additionally explored a higher-energy region containing structures with large c lattice parameters greater than approximately 20 Å. A representative structure of this region is shown in panel (f) of Fig. 4. The exploration of structures with large c-parameters was particularly the case for the run that used the diverse initial pool with the lattice parameter descriptor, which did not generate the experimental structure. The lattice parameter descriptor, given by equation (5), scales as the inverse of the cube root of the unit cell volume. Therefore, the cluster-based fitness function based on it occasionally promotes the selection of structures with larger unit cell volumes, which may have higher energies.

This demonstrates how evolutionary niching helps overcome initial pool biases and genetic drift by evolving several subpopulations simultaneously. GA runs employing cluster-based fitness more frequently sample regions of the potential energy surface that are not well represented in the initial pool compared to control runs employing energy based fitness. Furthermore, GA runs using cluster-based fitness more frequently sample different regions than those sampled preferentially by GA runs employing energy-based fitness. Some of these underrepesented populations contain important low-energy structures, whereas others may not necessarily produce structures that could be viable polymorphs. This is consistent with the findings reported for Target XXII in Ref. 47. We now proceed to perform a more detailed analysis of the effect of the descriptor on evolutionary niching.

Fig. 5 presents an analysis of the final clusters generated by the GA runs that were started from the random initial pool and used the lattice parameter descriptor (panel a), the RDF descriptor (panel b), and the RCD descriptor (panel c). For each run that utilized evolutionary niching, AP was used to cluster the final population of structures with respect to the same descriptor used in the GA. The resulting distributions are shown in blue. AP was then used to predict the cluster labels of the structures generated in the energy-based control run, shown in red. For all descriptors, a negative squared euclidean metric is used for construction of the pairwise similarity matrix for AP clustering. The distribution of initial pool structures contained in each cluster is shown in grey. The cluster assigned to the experimental structure is indicated by a green arrow. Additionally, the average and standard deviation of the relative energies of the structures in each cluster is plotted in orange. Histograms produced from the GA runs that were started from the diverse initial pool are provided in the Supporting Information†.

For all three descriptors, the use of a cluster-based fitness function suppressed the sampling of clusters that are over-sampled in the control run (e.g. clusters 6-7 in panel (a), clusters 5 and 10 in panel (b), and clusters 7 and 10 in panel (c)). In addition, the GA runs that utilized evolutionary niching generally sampled more structures in clusters that were not represented in the initial pool than the control runs. For the runs that used the lattice parameter descriptor and the RDF descriptor, the respective clusters that contained the experimental structure (clusters 8 and 19, respectively) were more frequently sampled than in the control run. The runs based on L and RDF yield more uniform distributions than the run based on RCD. Additionally, the RCD descriptor produces

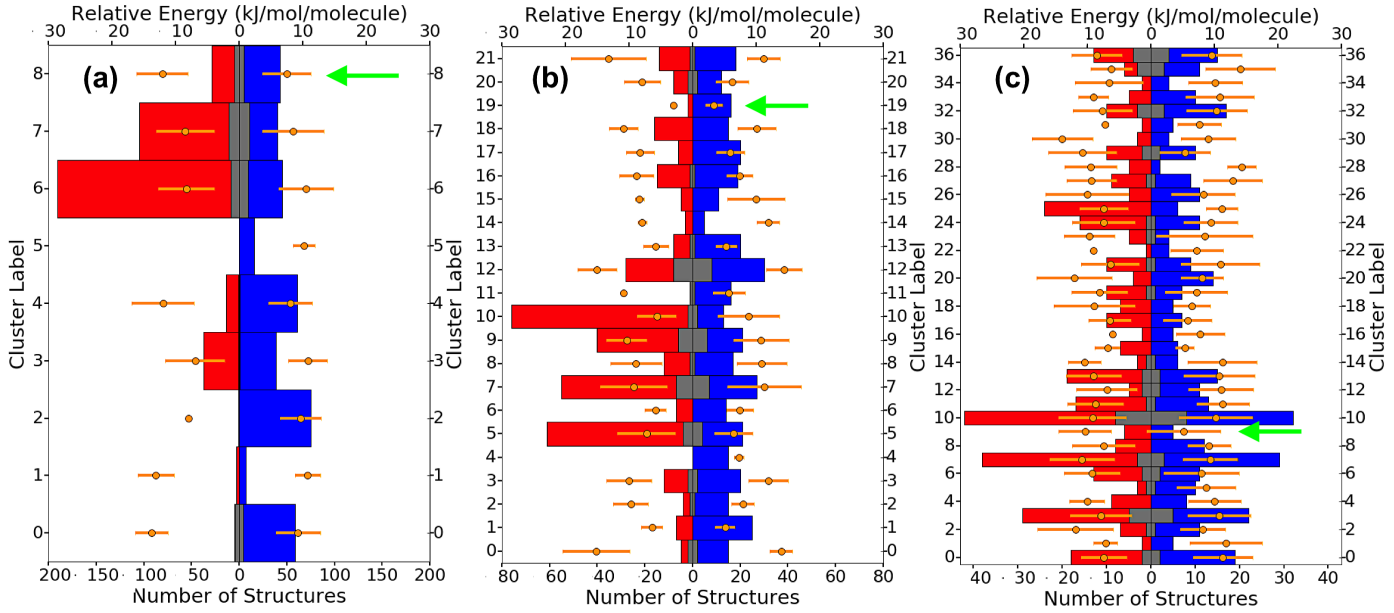

Fig. 5 The distribution of clusters produced by different GA runs initialized from the random initial pool. Histograms corresponding to runs that used evolutionary niching are shown in blue and histograms corresponding to the control run are shown in red. The structures are clustered with respect to (a) L, (b) RDF, and (c) RCD. In all plots the average relative energy and standard deviation for each cluster are shown in orange. The cluster assigned to the experimental structure for each run is indicated by a green arrow.

more overall variance in the average energies of the clusters than L and RDF. The fact that fewer structures with similar energies are grouped together with RCD may be related to the weaker correlation of RCD with unit cell volume, demonstrated in Ref. 55.

The three descriptors differ in the final number of clusters produced. The lattice parameter based run yields 9 final clusters, while the runs that used the RDF and RCD descriptors yield 22 and 37 clusters, respectively. By construction, the RDF and RCD descriptors capture the various packing motifs of Target XIII better than the lattice parameter descriptor. While elongated unit cells may correlate with zig-zag and herringbone motifs, and more box-like unit cells may correlate with layered motifs, there are more subtle differences between packing motifs that are not captured with the lattice parameter descriptor. Additionally, the lattice parameter descriptor promoted sampling of higher-energy structures with large *c*-parameters. For these reasons, the lattice parameter descriptor is probably too crude for the purposes of evolutionary niching.

The recommended best practice when using GAtor is to run the GA several times with different settings and then collect all the resulting structures, remove duplicates, and re-relax all unique structures with *higher-level* numerical settings. The final structures may then be re-ranked using a series of increasingly accurate energy methods.[47] The primary focus of this work is to study the effect of evolutionary niching. Therefore, the best structures produced from the GA runs here are locally optimized using PBE+TS and *higher-level* numerical settings, and no further re-ranking is performed. Thermal contributions to the total energy, which may influence the relative stabilities of polymorphs,[67] are also not included in the present study.

Table 1 presents the top 10 lowest energy structures produced from all the runs combined, as computed using PBE+TS and *higher-level* numerical settings. In the last column, the specific runs that generated a given low-energy structure are indicated. The experimental structure, which has a zig-zag packing motif and elongated unit cell, is ranked as #1 and was only generated in GA runs that used clustering. The structures ranked #6 and #8, shown in panels (d-e) in Fig. 4, have herringbone packing motifs and were also only generated in runs that used clustering. These structures were not produced in our previous study of this molecule, in which evolutionary niching was not used.[47] Seven of the ten top structures (ranked #2-#5, #7, #9-#10) were generated by the control runs using energy-based fitness. The majority of these structures display a layered packing motif. For example, the structures ranked #2 and #4 are shown in panels (a) and (b) of Fig. 4, respectively. Six of the seven top structures found in runs that used energy-based fitness were also generated by at least one GA run that used cluster-based fitness. This shows that evolutionary niching can be a valuable tool for generating novel structures that may be overlooked when using a traditional energy-based fitness function.

4 Concluding Remarks

The effect of using evolutionary niching in the GAtor genetic algorithm was investigated for the crystal structure prediction of 1,3-dibromo-2-chloro-5-fluorobenzene (Target XIII from the fourth CCDC blind test[50]). Evolutionary niching was performed by using the affinity propagation machine learning algorithm to dynamically cluster the population based on structural similarity. Clustering was conducted with respect to three structural descriptors: the first was based on the lattice parameters (L); the second was based on the combined radial distribution functions (RDFs) of Br···Br, H···Br, H···F, and H···Cl interatomic distances; the

third was a relative coordinate descriptor (RCD), based on the relative positions and orientations of neighboring molecules. A cluster-based fitness function was then used to steer the GA towards under-sampled regions of the potential energy surface by penalizing the fitness of structures found in populous clusters, thus reducing their probability of being selected for mating.

The results of GA runs that used evolutionary niching based on L, RDF, and RCD were compared to control runs that employed a traditional energy-based fitness function. To examine the effect of the initial population, two sets of runs were launched from a random pool and from a diverse pool, generated by Genarris. Evolutionary niching promoted sampling in regions of the potential energy landscape that were not well represented in the initial population as well as in regions rarely sampled by the control runs. We have thus demonstrated that evolutionary niching can help overcome initial pool biases and evolutionary drift in a genetic algorithm.

The experimental structure of Target XIII is characterized by a zig-zag packing motif, in contrast to most of the other low-energy structures that are characterized by layered packing arrangements. The region of the experimental structure was therefore rarely sampled by the control runs with the energy-based fitness function. Employing evolutionary niching based on all three descriptors significantly enhanced the likelihood of generating the experimental structure, as well as other low-energy structures with herringbone packing motifs.

Some differences were observed in the GA behavior when evolutionary niching was performed based on different descriptors. Using the L descriptor resulted in increased sampling of a region characterized by structures with very long c-parameters, particularly when the GA was started from the diverse initial pool. This was counterproductive in the case of Target XIII because this region did not contain any important low-energy structures. The L descriptor was additionally found to produce fewer clusters than the RDF and RCD descriptors. This suggests that the sensitivity of the L descriptor may be insufficient to resolve different packing motifs, which do not necessarily correlate with the unit cell shape and volume.

In conclusion, the recommended best practice for molecular crystal structure prediction with GAtor is to run the GA several times with different settings. At least one of the runs should employ a cluster-based fitness function using the RDF or RCD descriptors. The best structures produced from all GA runs should be combined for postprocessing. This increases the likelihood of generating more low-energy structures overall than when running GAtor with the energy-based fitness function alone. Evolutionary niching using cluster-based fitness functions is a promising strategy for generating important structures located in different basins of the potential energy surface for CSP applications.

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. Part of this research was performed while the authors were visiting the Institute for Pure and Applied Mathematics (IPAM), which is supported by the National Science Foundation.

References

1 J. Bernstein, *Polymorphism in Molecular Crystals*, Oxford University Press: Oxford, England, 2002, vol. 14.
2 G. M. Day, W. D. S. Motherwell and W. Jones, *Phys. Chem. Chem. Phys.*, 2007, **9**, 1693–704.
3 A. M. Reilly and A. Tkatchenko, *Phys. Rev. Lett.*, 2014, **113**, 055701.
4 D. P. Elder, J. E. Patterson and R. Holm, *J. Pharm. Pharmacol.*, 2015, **67**, 757–772.
5 C. Reese and Z. Bao, *Mater. Today*, 2007, **10**, 20–27.
6 T. Hasegawa and J. Takeya, *Sci. Technol. Adv. Mater.*, 2009, **10**, 024314.
7 S. Bergantin and M. Moret, *Cryst. Growth Des.*, 2012, **12**, 6035–6041.
8 P. Cudazzo, M. Gatti and A. Rubio, *Phys. Rev. B*, 2012, **86**, 195307.
9 P. Cudazzo, F. Sottile, A. Rubio and M. Gatti, *J. Phys.: Condens. Matter*, 2015, **27**, 113204.
10 N. Panina, F. J. J. Leusen, F. F. B. J. Janssen, P. Verwer, H. Meekes, E. Vlieg and G. Deroover, *J. Appl. Crystallogr.*, 2007, **40**, 105–114.
11 M. Fitzgerald, M. G. Gardiner, D. Armitt, G. W. Dicinoski and C. Wall, *J. Phys. Chem. A*, 2015, **119**, 905–10.
12 M. A. Neumann, J. van de Streek, F. P. A. Fabbiani, P. Hidber and O. Grassmann, *Nat. Commun.*, 2015, **6**, 7793.
13 S. L. Price, D. E. Braun and S. M. Reutzel-Edens, *Chem. Commun.*, 2016, **52**, 7065–7077.
14 A. G. Shtukenberg, Q. Zhu, D. J. Carter, L. Vogt, J. Hoja, E. Schneider, H. Song, B. Pokroy, I. Polishchuk, A. Tkatchenko, A. R. Oganov, A. L. Rohl, M. E. Tuckermanejk and B. Kahral, *Chem. Sci.*, 2017, **8**, 4926–4940.
15 A. G. Shtukenberg, C. Hu, Q. Zhu, M. U. Schmidt, W. Xu, M. Tan and B. Kahr, *Cryst. Growth Des.*, 2017, **17**, 3562–3566.
16 B. Meredig and C. Wolverton, *Nat. Mater.*, 2013, **12**, 123–127.
17 N. Marom, R. A. DiStasio, V. Atalla, S. Levchenko, A. M. Reilly, J. R. Chelikowsky, L. Leiserowitz and A. Tkatchenko, *Angew. Chem. Int. Ed.*, 2013, **52**, 6629–6632.
18 A. J. Cruz-Cabeza, S. M. Reutzel-Edens and J. Bernstein, *Chem. Soc. Rev.*, 2015, **44**, 8619–8635.
19 G. J. Beran, *Angew. Chem. Int. Ed.*, 2015, **54**, 396–398.
20 G. J. Beran, *Chem. Rev.*, 2016, **116**, 5567–5613.
21 R. L. Johnston, *Dalton Trans.*, 2003, 4193–4207.
22 M. Sierka, *Prog. Surf. Sci.*, 2010, **85**, 398–434.
23 S. Heiles and R. L. Johnston, *Int. J. Quantum Chem.*, 2013, **113**, 2091–2109.
24 A. R. Oganov and C. W. Glass, *J. Chem. Phys.*, 2006, **124**, 244704.

Table 1 The top 10 crystal structures of Target XIII as ranked with PBE+TS and *higher-level* numerical settings. In the last column, control indicates runs that used energy-based fitness, while L, RDF, and RCD stand for runs that used cluster-based fitness. The initial pool used in a given run is indicated in parentheses, with R indicating the random pool and D indicating the diverse pool.

Rank	ΔE (kJ/mol/molecule)	Volume (Å^3)	a (Å)	b (Å)	c (Å)	α (°)	β (°)	γ (°)	Z	SG	Runs Generated
1	0.00	783.6	4.0	13.6	14.5	90	87	90	4	$P2_1/c$	L (R), RDF (R,D), RCD (R,D)
2	0.57	403.5	7.0	8.0	8.2	67	74	78	2	$P\bar{1}$	control (R,D), L (D), RDF (D)
3	0.84	784.7	4.0	13.3	14.9	90	96	90	4	$P2_1$	control (D), RCD (R)
4	0.98	797.9	7.4	9.2	13.0	84	77	69	4	$P\bar{1}$	control (R,D)
5	1.11	799.8	8.8	7.0	15.7	90	125	90	4	$P2_1/c$	control (R,D), L (D), RDF (R), RCD (D,R)
6	1.15	788.0	4.0	13.4	15.8	110	90	90	4	Pn	L (D), RDF(R)
7	1.23	794.3	7.6	7.8	14.0	97	96	103	4	$P\bar{1}$	control (R,D), L (D), RDF (R), RCD(R)
8	1.34	791.0	4.0	13.6	16.0	68	86	89	4	$P\bar{1}$	RDF(R), RCD (R)
9	1.43	798.2	7.2	9.3	12.7	83	89	71	4	$P\bar{1}$	control (R,D), L (D)
10	1.46	795.2	7.9	9.3	10.9	94	96	92	4	$P\bar{1}$	control (R,D), RDF (D)

25 C. W. Glass, A. R. Oganov and N. Hansen, *Comput. Phys. Commun.*, 2006, **175**, 713–720.

26 N. L. Abraham and M. I. J. Probert, *Phys. Rev. B*, 2006, **73**, 224104.

27 G. Trimarchi and A. Zunger, *Phys. Rev. B*, 2007, **75**, 104113.

28 S. Wu, K. Umemoto, M. Ji, C. Z. Wang, K. M. Ho and R. M. Wentzcovitch, *Phys. Rev. B*, 2011, **83**, 184102.

29 S. Woodley, P. Battle, J. Gale and C. A. Catlow, *Phys. Chem. Chem. Phys.*, 1999, **1**, 2535–2542.

30 G. Trimarchi and A. Zunger, *J. Phys.: Condens. Matter*, 2008, **20**, 295212.

31 D. C. Lonie and E. Zurek, *Comput. Phys. Commun.*, 2011, **182**, 372–387.

32 G. H. Jóhannesson, T. Bligaard, A. V. Ruban, H. L. Skriver, K. W. Jacobsen and J. K. Nørskov, *Phys. Rev. Lett.*, 2002, **88**, 255506.

33 Q. Zhu, A. R. Oganov, C. W. Glass and H. T. Stokes, *Acta Crystallogr., Sect. B: Struct. Sci.*, 2012, **68**, 215–226.

34 A. M. Lund, G. I. Pagola, A. M. Orendt, M. B. Ferraro and J. C. Facelli, *Chem. Phys. Lett.*, 2015, **626**, 20–24.

35 P. Avery, Z. Falls and E. Zurek, *Comput. Phys. Commun.*, 2017, **217**, 210–211.

36 Z. Falls, D. C. Lonie, P. Avery, A. Shamp and E. Zurek, *Comput. Phys. Commun.*, 2016, **199**, 178–179.

37 J. Morris, D. Deaven, K. Ho, C. Wang, B. Pan, J. Wacker and D. Turner, *IMA Vol. Math. Its Appl.*, 1999, **111**, 167–176.

38 A. N. Alexandrova and A. I. Boldyrev, *J. Chem. Theory Comput.*, 2005, **1**, 566–580.

39 J. M. C. Marques and F. B. Pereira, *Chem. Phys. Lett.*, 2010, **485**, 211–216.

40 B. Hartke, *J. Comput. Chem.*, 1999, **20**, 1752–1759.

41 C. R. A. Catlow, S. T. Bromley, S. Hamad, M. Mora-Fonz, A. A. Sokol and S. M. Woodley, *Phys. Chem. Chem. Phys.*, 2010, **12**, 786–811.

42 V. E. Bazterra, O. Oña, M. C. Caputo, M. B. Ferraro, P. Fuentealba and J. C. Facelli, *Phys. Rev. A*, 2004, **69**, 053202.

43 S. Bhattacharya, S. V. Levchenko, L. M. Ghiringhelli and M. Scheffler, *Phys. Rev. Lett.*, 2013, **111**, 135501.

44 S. Bhattacharya, S. V. Levchenko, L. M. Ghiringhelli and M. Scheffler, *New J. Phys.*, 2014, **16**, 123016.

45 M. S. Jørgensen, M. N. Groves and B. Hammer, *J. Chem. Theory Comput.*, 2017, **13**, 1486–1493.

46 W. W. Tipton and R. G. Hennig, *J. Phys.: Condens. Matter*, 2013, **25**, 495401.

47 F. Curtis, X. Li, T. Rose, A. Vázquez-Mayagoitia, S. Bhattacharya, L. M. Ghiringhelli and N. Marom, *J. Chem. Theory Comput*, 2018, DOI: 10.1021/acs.jctc.7b01152.

48 F. Curtis, X. Wang and N. Marom, *Acta Crystallogr., Sect. B: Struct. Sci., Cryst. Eng. Mater.*, 2016, **72**, 562–570.

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

50 G. M. Day, T. G. Cooper, A. J. Cruz-Cabeza, K. E. Hejczyk, H. L. Ammon, S. X. M. Boerrigter, J. S. Tan, R. G. Della Valle, E. Venuti, J. Jose, S. R. Gadre, G. R. Desiraju, T. S. Thakur, B. P. van Eijck, J. C. Facelli, V. E. Bazterra, M. B. Ferraro, D. W. M. Hofmann, M. A. Neumann, F. J. J. Leusen, J. Kendrick, S. L. Price, A. J. Misquitta, P. G. Karamertzanis, G. W. A. Welch, H. A. Scheraga, Y. A. Arnautova, M. U. Schmidt, J. van de Streek, A. K. Wolf and B. Schweizer, *Acta Crystallogr., Sect. B: Struct. Sci.*, 2009, **65**, 107–125.

51 K. E. Riley and P. Hobza, *J. Chem. Theory Comput.*, 2008, **4**, 232–42.

52 G. Cavallo, P. Metrangolo, R. Milani, T. Pilati, A. Priimagi,

G. Resnati and G. Terraneo, *Chem. Rev*, 2016, **116**, 2478–2601.

53 S. Kozuch and J. M. L. Martin, *J. Chem. Theory Comput.*, 2013, **9**, 1918–31.

54 A. Otero-de-la-Roza, E. R. Johnson and G. A. DiLabio, *J. Chem. Theory Comput.*, 2014, **10**, 5436–5447.

55 X. Li, F. Curtis, T. Rose, C. Schober, A. Vázquez-Mayagoitia and N. Marom, *J. Chem. Phys.*, 2018, DOI: 10.1063/1.5014038.

56 J. P. Perdew, K. Burke and M. Ernzerhof, *Phys. Rev. Lett.*, 1996, **77**, 3865–3868.

57 J. P. Perdew, K. Burke and M. Ernzerhof, *Phys. Rev. Lett.*, 1997, **78**, 1396–1396.

58 A. Tkatchenko and M. Scheffler, *Phys. Rev. Lett.*, 2009, **102**, 073005.

59 V. Blum, R. Gehrke, F. Hanke, P. Havu, V. Havu, X. Ren, K. Reuter and M. Scheffler, *Comput. Phys. Commun.*, 2009, **180**, 2175–2196.

60 S. Bhattacharya, B. H. Sonin, C. J. Jumonville, L. M. Ghiringhelli and N. Marom, *Phys. Rev. B*, 2015, **91**, 241115.

61 B. J. Frey and D. Dueck, *Science*, 2007, **315**, 972–976.

62 F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot and E. Duchesnay, *J. Mach. Learn. Res.*, 2011, **12**, 2825–2830.

63 P. Niggli, *Krystallographische Und Strukturtheoretische Grundbegriffe*, Akademische Verlagsgesellschaft: Leipzig, Germany, 1928.

64 B. Gruber, *Acta Crystallogr., Sect. A: Cryst. Phys., Diffr., Theor. Gen. Crystallogr.*, 1973, **29**, 433–440.

65 I. Křivỳ and B. Gruber, *Acta Crystallogr., Sect. A: Cryst. Phys., Diffr., Theor. Gen. Crystallogr.*, 1976, **32**, 297–298.

66 R. W. Grosse-Kunstleve, N. K. Sauter and P. D. Adams, *Acta Crystallogr., Sect. A: Found. Crystallogr.*, 2004, **60**, 1–6.

67 J. Nyman and G. M. Day, *CrystEngComm*, 2015, **17**, 5154–5165.