First-Principles Molecular Structure Search with a Genetic Algorithm

Adriana Supady,[*,†] Volker Blum,[†,‡] and Carsten Baldauf[*,†]

Fritz-Haber-Institut der MPG, Berlin, Germany, and Department of Mechanical Engineering & Materials Science, Duke University, Durham, U.S.A.

E-mail: supady@fhi-berlin.mpg.de; baldauf@fhi-berlin.mpg.de

Abstract

The identification of low-energy conformers for a given molecule is a fundamental problem in computational chemistry and cheminformatics. We assess here a conformer search that employs a genetic algorithm for sampling the low-energy segment of the conformation space of molecules. The algorithm is designed to work with first-principles methods, facilitated by the incorporation of local optimization and blacklisting conformers to prevent repeated evaluations of very similar solutions. The aim of the search is not only to find the global minimum, but to predict all conformers within an energy window above the global minimum. The performance of the search strategy is: (i) evaluated for a reference data set extracted from a database with amino acid dipeptide conformers obtained by an extensive combined force field and first-principles search and (ii) compared to the performance of a systematic search and a random conformer generator for the example of a drug-like ligand with 43 atoms, 8 rotatable bonds and 1 *cis/trans* bond.

[*]To whom correspondence should be addressed
[†]FHI Berlin
[‡]Duke University

Introduction

One of the fundamental problems in cheminformatics and computational chemistry is the identification of three-dimensional (3D) conformers that are energetically favourable and likely to be encountered in experiment at given external conditions.[1] Conventionally, these conformers are often characterized by specific, fixed sets of nuclear coordinates or ensembles thereof, and their potential energy is given by the electronic degrees of freedom in a Born-Oppenheimer picture of the chemical bond. A variety of conformations can be adopted by flexible organic molecules as the multi-dimensional potential-energy surface (PES) usually contains multiple local minima, with a global minimum among them. Only when the relevant conformers are known, one can predict and evaluate chemical and physical properties of the molecules (e.g. reactivity, catalytic activity, or optical properties). In many practical applications, the PES minima are taken as starting points to explore the free-energy surface (FES). Generating conformers is an integral part of methods such as protein-ligand docking[2–5] or 3D pharmacophore modeling.[6] The propensity to adopt a certain conformation strongly depends on the environment and possible interactions with other compounds. It has been shown that the bioactive conformation of drug-like molecules can be higher in energy than the respective global minimum[7] and that different 3D conformations may be induced by specific interactions with other molecules.[8] Thus, it is crucial to focus not just on a single, global minimum of the PES, but instead to provide a good coverage of the accessible conformational space of a molecule yielding diverse low-energy conformers.

The exploration of a high-dimensional PES is challenging. A selection of popular sampling approaches utilized in conformer generation is summarized in Table 1. We focus specifically on genetic algorithms (GAs)[29–31] that belong to the family of evolutionary algorithms (EAs) that are frequently used for global structure optimization of chemical compounds.[3,4,32–56] GAs for chemical structure searches implement a 'survival of the fittest' concept and adopt evolutionary principles starting from a population of, most commonly, random solutions. GAs use the accumulated information to explore the most promising regions of the conformational space. With this, the number of unhelpful evaluations of physically implausible high-energy solutions can be reduced.

Table 1: Popular sampling approaches. Names of freely available programs are highlighted in boldface.

Method	Description	Implemented, e.g., in
grid-based	based on grids of selected Cartesian or internal coordinates (e.g., grids of different torsional angle values of a molecule)	CAESAR,[9] **Open Babel**,[10] **Confab**,[11] MacroModel,[12] MOE[13]
rule/knowledge - based	use known (e.g., from experiments) structural preferences of compounds	**ALFA**,[14] **CONFECT**,[15] CORINA and ROTATE,[16,17] **COSMOS**,[18,19] OMEGA[20]
population-based metaheuristic	improve candidate solutions in a guided search	**Balloon**,[21] **Cyndi**[22]
distance geometry	based on a matrix with permitted distances between pairs of atoms	**RDKit**[23]
basin-hopping[24] / minima hopping[25]	based on moves across the PES combined with local relaxation	**ASE**,[26] **GMIN**,[27] **TINKER SCAN**[28]

Examples of GA-based structure prediction applications include: (i) conformational searches for molecules like of unbranched alkanes[36] or polypeptide folding;[37] (ii) molecular design;[38,39] (iii) protein-ligand docking;[3,4] cluster optimization;[40–49] (v) predictions of crystal structures;[50–53] (vi) structure and phase diagram predictions.[54] Further, Neiss and Schoos[55] proposed a GA including experimental information into the global search process by combining the energy with the experimental data in the objective function. Since GAs typically rely on internal, algorithmic parameters that control the efficiency of a search, a meta-GA for optimization of a GA search for conformer searches was proposed by Brain and Addicoat.[56]

Aside from the search algorithm itself, the choice of the mathematical model for the PES is critical to ensure results that reliably reflect the experimental reality. Among the available atomistic simulation approaches, "molecular mechanics" models, i.e., so-called force fields are especially fast from a computational point of view and therefore often employed. However, the resulting predictions depend on the initial parametrization of a particular force field and can lead to considerable rearrangements of the true PES for molecules that were not included in the parameterization procedure.[57–59] On the other end of the spectrum of approaches, the PES can be faithfully represented based on the "first principles" of quantum mechanics. Indeed, benchmark quality approaches such as coupled cluster theory at the level of singles, doubles and perturbative triples

(CCSD(T)) are almost completely trustworthy for closed-shell molecules, but still prohibitively expensive towards larger systems and/or large-scale screening of energies of many conformers. Density-functional theory (DFT) approximations are an attractive alternative to balance accuracy and computational cost. The choice of the approximation is critical when using first principles methods like DFT. It has been shown that it is necessary to incorporate dispersion effects for (bio)organic molecules and their complexes.[57,60–62] The challenge of including long-range interactions has been met for example by the dispersion correction schemes described by Grimme[63,64] or by Scheffler and Tkatchenko,[65–67] but validating the DFT approximation employed is critical. In fact, subtle energy balances of competing conformers can require relatively high-level DFT approximations for reliable predictions.[58,68]

The aim of our work is to develop and test an approach to sample the PES of small to medium sized (bio)organic molecules without relying on empirical force fields, utilizing instead electronic-structure methods for the entire search. With the molecular structure problem in mind, we define following requirements for the search strategy and implementation:

- Global search based on user-curated torsional degrees of freedom (bond rotations).
- Local optimization based on full relaxation of Cartesian coordinates and avoidance of re-computing too similar structures to ensure both efficient sampling and economic use of a computationally demanding energy function.
- Design of the program in a way to use an external and easily exchangeable electronic structure code (in our case FHI-aims[69,70]).
- Simple input of molecules (composition and configuration) by means of SMILES codes.[71]
- A robust and simple metaheuristic that ideally identifies the complete ensemble of low-energy conformers.
- Freely available and with a flexible open-source license model.
- Support for parallel architectures.

Based on these requirements, we present in this work a conformational search strategy based on a genetic algorithm. We provide a detailed description of our approach and a software implementation Fafoom (Flexible algorithm for optimization of molecules) that is available under an open-source license (GNU Lesser General Public License[72]) for use by others. For simplicity, we abbreviate 'potential energy' with 'energy' and 'minima of the potential-energy surface' with 'energy minima'.

Methods

In the following, we first motivate and explain assumptions that we met for handling 3D structures of molecules. Further, we outline the algorithm's implementation and describe its technical details. Finally we introduce a data set that we use as a reference for evaluating the performance of our implementation. Our work focuses on both the ability to reliably predict the global minimum and to provide a good conformational coverage with a computationally feasible approach. To achieve that, we formulate some specific algorithmic choices at the outset: (i) only sterically feasible conformations are accepted for local optimization; (ii) a geometry optimization to the next local minimum is performed for every generated conformation; (iii) an already evaluated conformation will not be evaluated again.

Choice of coordinates

In computational chemistry, at least two ways of representing a molecule's 3D structure are commonly used, either Cartesian or internal coordinates. The simplest internal coordinates are based on the 'Z-matrix coordinates', which include bond lengths, bond angles and dihedral angles (torsions) and can also be referred to as 'primitive internal coordinates'. These coordinates reflect the actual connectivity of the atoms and are well suited for representing curvilinear motions such as rotations around single bonds.[73] Bond lengths and bond angles posses usually only one rigid minimum, i.e. the energy increases rapidly if these parameters deviate from equilibrium. In contrast, torsions can

change in value by an appreciable amount without a dramatic change in energy. Similar to the work of Damsbo *et al.*,[37] we use Cartesian coordinates for the local geometry optimizations while internal coordinates, in this work only torsional degrees of freedom (TDOFs), i.e. freely rotatable bonds and, if present, *cis/trans* bonds, are used for the global search. We consider only single, non-ring bonds between non-terminal atoms to be rotatable bonds after excluding bonds that are attached to methyl groups that carry three identical substituents. Further we allow for treating selected bonds in a *cis/trans* mode, i.e. allowing only for two different relative positions of the substituents. In cases in which the substituents are oriented in the same direction we refer to it as to *cis*, whereas, when the substituents are oriented in opposite directions, we refer to it as to *trans*.

Handling of molecular structures

Figure 1 shows different chemical representations of a molecule, here for the example of 3,4-dimethylhex-3-ene. Figure 1A and B depict the standard 3D and 2D representation of the compound together with marked *cis/trans* and rotatable bonds. A SMILES (simplified molecular-input line-entry system) string is shown in Figure 1C. A SMILES representation[71] of a chemical compound encodes the composition, connectivity, the bond order (single, double, triple), as well as stereochemical information in a one-line notation. Finally, a vector representation (Figure 1D) can be created if the locations of *cis/trans* and rotatable bonds are known. The vector will store the corresponding torsion angle values. Our implementation takes as input a SMILES representation of a molecule, while vectors of angles are used to internally encode different structures in the genetic algorithm below.

Frequently used terms

Several terms need to be defined prior to describing the structure of the algorithm. In the following, the parameters of the search are highlighted in boldface. These parameters are input parameters to the algorithm and need to be defined in the input file.

A *sensible geometry* meets two constraints. First, the atoms are kept apart, i.e. none of the

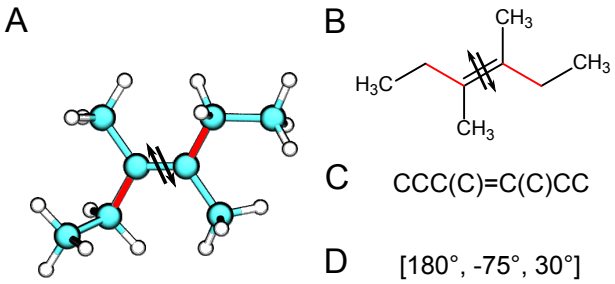

Figure 1: Different chemical representations of 3,4-dimethylhex-3-ene: (A) 3D structure with rotatable bonds marked in red and the *cis/trans* bond marked with double arrows. (B) 2D structure. (C) SMILES string. (D) vector representation of the molecule. The first value encodes the torsion angle value for the *cis/trans* bond and the two remaining position store the torsion angle values of the rotatable bonds.

distances between non-bonded atoms can be shorter than a defined threshold (**distance_cutoff_1**, default=1.3 Å). Secondly, it is fully connected, i.e. none of the distances between bonded atoms can be longer than a defined threshold (**distance_cutoff_2**, default=2.15 Å). The attribute *sensible* can be used further to describe any operation that outputs sensible geometries.

The *blacklist* stores all structures that: (i) were starting structures for the local optimization and (ii) resulted from local optimization, as they may have changed significantly during the optimization. In case of achiral molecules (**chiral**, default=False) also the corresponding mirror images are created and stored.

A structure is *unique* if none of the root-mean-square deviation (RMSD) values calculated for the structure paired with any of the structures stored in the blacklist is lower than a defined threshold (**rmsd_cutoff_uniq**, default=0.2 Å). We consider only non-hydrogen atoms for the calculation of the RMSD.

Basic outline of the search algorithm

We implemented the genetic algorithm (GA) using the Python language (version 2.7) and employ the RDKit library.[23] An overview is presented in Algorithm 1.

```
# initialization
while i < popsize:
    x = random_sensible_geometry
    blacklist.append(x)
    x = DFT_relaxation(x)
    blacklist.append(x)
    population.append(x)
    i+=1
# iteration
while j < iterations:
    population.sort(index=energy)
    (parent1, parent2) = population.select_candidates(2)
    (child1, child2) = sensible_crossover(parent1, parent2)
    (child1, child2) = mutation(child1, child2)
    repeat
        (child1, child2) = mutation(child1, child2)
    until child1 and child2 are sensible and are not in the blacklist

    blacklist.append(child1, child2)
    (child1, child2) = DFT_relaxation(child1, child2)
    blacklist.append(child1, child2)
    population.append(child1, child2)
    population.sort(index=energy)
    population.delete_high_energy_candidates(2)
    if convergence criteria met:
        break
    else:
        j+=1
```

Algorithm 1: Genetic algorithm for sampling the conformational space of molecules.

Initialization of the population

First, a random 3D structure is generated with RDKit directly from the SMILES code. This structure serves as a template for the upcoming geometries. Next, two lists of random values are generated: one for the rotatable bonds and one for the *cis/trans* values. If the resulting 3D geometry is sensible, the structure is then subjected to local optimization. To generate an initial population of size N (**popsize**), N sensible geometries with randomly assigned values for torsion angles need to

be built and locally optimized. The optimized geometries constitute the initial population. Due to the fact that the geometries are created one after another, all randomly built structures can but do not have to be made unique in order to increase the diversity of the initial population.

An iteration of the GA

Our GA follows the established generation-based approach, i.e. the population evolves over subsequent generations. After completion of the initialization, the first iteration can be performed. For this purpose, the population is sorted and ranked based on the total energy values E_i of the different conformers $i = 1, ..., N$. For each individual the fitness F_i is being calculated according to:

$$F_i = \frac{E_{max} - E_i}{E_{max} - E_{min}} \tag{1}$$

E_{max} is the highest energy and E_{min} is the lowest energy among the energies of the conformers belonging to the current population. As a result, $F = 1$ for the 'best' conformer and $F = 0$ for the 'worst' conformer. In the case of a population with low variance in energy values ($E_{max} - E_{min} <$ **energy_var**, default=0.001 eV), all individuals are assigned a fitness of 1.0.

Selection. Two individuals need to be selected prior to the genetic operations. We implemented three mechanisms for the selection.

(i) In the (energy-based) *roulette wheel*,[31] the probability p_i for selection of a conformer i is given by:

$$p_i = \frac{F_i}{\sum_{n=1}^{N} F_i} \tag{2}$$

With this, the probabilities of the conformers are mapped to segments of a line of length one. Next, two random numbers between zero and one are generated and the conformers whose segments contain these random numbers are selected. In the case when the sum of the fitness values is lower than a defined threshold near one (**fitness_sum_limit**, default=1.2) the best and a random individual are selected.

(ii) The *reverse roulette wheel* proceeds similarly to the *roulette wheel* mechanism with the difference that the fitness values are swapped, i.e. new fitness F_i^* is assigned to each conformer:

$$F_i^* = F_{N+1-i} \tag{3}$$

Analogously, the probability p_i for selection of a conformer i is given by:

$$p_i = \frac{F_i^*}{\sum_{n=1}^{N} F_i^*} \tag{4}$$

(iii) In the *random selection* mechanism all individuals have the same chance to be selected.

In all selection mechanisms the selected individuals must be different from each other so that the crossing-over has a chance to produce unique conformers.

Crossing-over. Crossing-over is considered to be the main feature distinguishing evolutionary algorithms from Monte Carlo techniques where only a single solution can evolve. Crossing-over allows the algorithm to take big steps in exploration of the search space.[37] In our algorithm, a crossing-over step is performed if a generated random number (between zero and one) is lower than a defined threshold (**prob_for_crossing**, default=0.95). Between the selected individuals, parts of the representing vectors are then exchanged. To that end, the vectors characterizing the structure of both individuals are "cut" at the same single position (determined at random). The first part of the first individual is then combined with the second part of the second, and vice versa (a scheme explaining the crossing-over procedure is provided in Figure S1). Crossing-over is successful only when the resulting vectors can be used for generating sensible geometries. Otherwise the crossing-over is repeated until sensible geometries are generated or a maximum number of attempts (**cross_trial**, default=20) is exceeded. In the latter case, exact copies of the selected conformers are used for the following step.

Mutations are performed independently for the values of *cis/trans* bonds and of the rotatable bonds and if randomly generated numbers exceed corresponding thresholds (**prob_for_mut_cistrans**, default=0.5; **prob_for_mut_rot**, default=0.5). For each, the number of mutation events is deter-

mined by a randomly picked integer number not higher than the user-defined maximal number of allowed mutations (**max_mutations_cistrans** and **max_mutations_torsions**). For each mutation, a random position of the vector is determined and the mutation is chosen to affect the value of that variable. In case of *cis/trans* bonds, the selected value is changed to 0° if it was above 90° or below -90°, else it is changed to 180°. A selected rotatable bond is changed to a random integer between -179° and 180°. A mutation step is only successful if the geometry built after the mutation of the vector is sensible and unique. Otherwise the entire set of mutations in a mutation step is repeated until a sensible and unique structure is generated or a maximum number of attempts (**mut_trial**, default=100) is exceeded. In this case, the algorithm terminates. The mutation is performed for both vectors generated via crossing-over.

Local optimization and update. As the computational cost of the local optimization is significantly higher than all of the other operations,[54,74] only unique and sensible structures are subject to local optimization. The structures are transferred to an external program for local geometry optimization (here: FHI-aims[69,70], see section DFT calculations). The application of local optimization was shown to facilitate the search for minima by reducing the space the GA has to search.[24,43] Thus, the implemented GA is closer to Lamarckian than to Darwinian evolution, as the individuals evolve and pass on acquired and not inherited characteristics. Afterwards, the population is extended by the newly optimized structures and, after ranking, the two individuals with highest energy are removed in order to keep the population size constant.

Termination of the algorithm is reached if one of the convergence criteria is met: (i) the lowest energy has not changed more than a defined threshold (**energy_diff_conv**, default=0.001 eV) during a defined number of iterations (**iter_limit_conv**, default=10), or (ii) the lowest energy has reached a defined value (**energy_wanted**), or (iii) the maximal number of iterations (**max_iter**, default=10) has been reached. The convergence criteria are checked only after a defined number of iterations (**iter_limit_conv**, default=10).

We are interested not only in finding the global minimum but also in finding low-energy local minima as many of them may be relevant. Thus, all of the generated conformers are saved and

are available for final analysis even if only a subset of them constitutes the final population. Table 2 summarizes practical GA parameters that were employed for one of the reference systems (isoleucine dipeptide).

The parameters listed in Table 2 can be taken as indicative of settings that will work for many small to mid-size molecules. A few exceptions apply. Specifically, the **max_iter** and the **popsize** parameters are set to low values in Table 2, covering only a small set of structures within an individual GA run. This choice would be appropriate for an ensemble of many short independent GA run to generate a broad structural ensemble with a bias towards the low-energy solution space. For larger and more complex molecules, and/or for runs designed to identify the global minimum in a single shot, **max_ite**r could be increased significantly, and **popsize** could be increased somewhat (to 10-20 individuals) as well. Likewise, the mutation probabilities **prob_for_mut_cistrans** and **prob_for_mut_rot** are here set to relatively high values of 0.5, instilling a significantly amount of randomness into the search process. For a more "deterministic" search process, somewhat smaller values (e.g., 0.2) might be chosen. Finally, the **distance_cutoff** criteria are chosen to be appropriate for light elements (first and second row); adjustments may be appropriate if heavier covalently bonded atoms are included in the search.

DFT calculations

For the tests presented below, all DFT calculations are performed with the FHI-aims code.[69,70] We employed the PBE functional[75] with a correction for van der Waals interactions (pairwise[65] for the amino acid dipeptides calculations and MBD[67] for the drug-like ligands) and with *light* computational settings and *tier* 1 basis set.[69] For the local optimization, we use a trust radius enhanced version of the Broyden–Fletcher–Goldfarb–Shanno (BFGS) algorithm[76] initialized by the model Hessian matrix of Lindh.[77] This is the default choice in the FHI-aims code and was implemented by Jürgen Wieferink. The local optimization is set to terminate when the maximum residual force component per atom is equal to $5 \cdot 10^{-3}$ eV/ Å. Density functional, basis set, and numerical settings (e.g. integration grids) are user choices of the underlying density-functional

Table 2: GA parameters for isoleucine dipeptide

	Parameter	Value
Molecule	SMILES	CC(=O)N[C@H](C(=O)NC)[C@H](CC)C
	distance_cutoff_1	1.2 Å
	distance_cutoff_2	2.0 Å
	rmsd_cutoff_uniq	0.2 Å
	chiral	True
Run settings	max_iter	10
	iter_limit_conv	10
	energy_diff_conv	0.001 eV
GA settings	popsize	5
	energy_var	0.001 eV
	selection	roulette wheel
	fitness_sum_limit	1.2
	prob_for_crossing	0.95
	cross_trial	20
	prob_for_mut_cistrans	0.5
	prob_for_mut_rot	0.5
	max_mutations_cistrans	1
	max_mutations_torsions	2
	mut_trial	100

theory code and must be set appropriately outside of Fafoom. The settings for numerical convergence (including basis set) must be chosen converged enough but not introduce artifacts in the landscape of minima found. The choice of the density-functional approximation (DFA) to the exact Born-Oppenheimer potential-energy surface needs to reproduce the local energy minima of the PES faithfully, as discussed in the introduction. We here only note that costs for different electronic structure approximation can vary by orders of magnitude. In practice, and strictly speaking, the scope of our algorithm is to find the PES minima for a given DFT functional, while the physical choice of the "right" DFA is not the focus of this paper. We do show, however, that we can use our algorithm in practice with one specific functional, PBE functional with a correction for van der Waals interactions, that has yielded very reliable results in the past.

Parallelization

Parallel computational resources can be utilized in two ways in order to speed up the computation. First, multiple GA runs can be started in parallel and the blacklist can be shared between different

and subsequent runs. Sharing the blacklist increases diversity of solutions with already a few GA runs. Second, the time needed for the individual energy evaluations can be decreased if the molecular simulation package allows calculations across distributed nodes and is efficiently parallelized (e.g. in FHI-aims[78]). Our code supports both modes of computation.

Availability of the code

The code is distributed as a python package named Fafoom (Flexible algorithm for optimization of molecules) under the GNU Lesser General Public License.[72] It is available from following websites:

- https://aimsclub.fhi-berlin.mpg.de/aims_tools.php
- https://github.com/adrianasupady/fafoom

Although designed for usage with a first-principles method (e.g. FHI-aims, NWChem[79]), Fafoom can also be used with a force field (MMFF94,[80] accessed within RDKit[23,81]) for testing purposes. It is in principle possible to use any molecular simulation package which outputs optimized geometries together with their energies. Nevertheless, this requires adjusting a part of the program to the specific needs of the used software. Details are provided with the program's documentation.

Reference data

In order to evaluate the several aspects of the performance of the implemented algorithm we use two sets of reference data. The first reference data set (**Amino acid dipeptides**) was extracted from a database of computational data for the amino acid dipeptides. The second reference data set (**Mycophenolic acid**) contains conformers of a drug-like ligand that were obtained with three different search techniques.

Amino acid dipeptides

The first reference data set contains conformers of seven amino acid dipeptides[82] (Figure 2) and was extracted from a large database for amino acid dipeptide structures generated in a combined basin-hopping/multi-tempering based search. In that search (published in detail in[83]), the framework of the reference search can be divided into a global search step and a refinement step. In the global search, the basin hopping search technique together with an empirical force field OPLS-AA was employed to perform the initial scan of the PES. The identified energy minima were relaxed at the PBE+vdW level with *light* computational settings in FHI-aims. In the refinement step, *ab initio* replica-exchange molecular dynamics runs were performed to locally explore the conformational space and to alleviate a potential bias of the initial search of a force field PES. The resulting minima were again optimized at the PBE+vdW level with *tight* computational settings and with the *tier* 2 basis set.[69] In order to compare to our data, they were re-optimized with the same functional with *light* computational settings, and the *tier* 1 basis set.[69] After this procedure, duplicates were removed from the set used for the comparison with the GA results. For benchmarking the performance of our search strategy for conformers predictions, we consider all structures with a relative energy up to 0.4 eV. These conformers define the reference energy hierarchy for each of the selected dipeptides. We summarize some characteristics and the number of conformers that were considered in Table 3.

Table 3: Reference data set: seven amino acid dipeptides.

Amino acid dipeptide	Abbr.	No. of atoms	No. of rotatable bonds + No. of *cis/trans* bonds	No. of conformers (below 0.4 eV $\approx$ 38.6 kJ/mol)
Glycine	Gly	19	2+2	15 (15)
Alanine	Ala	22	2+2	28 (17)
Phenylalanine	Phe	32	4+2	64 (37)
Valine	Val	28	3+2	60 (40)
Tryptophan	Trp	36	4+2	141 (77)
Leucine	Leu	31	4+2	183 (103)
Isoleucine	Ile	31	4+2	176 (107)

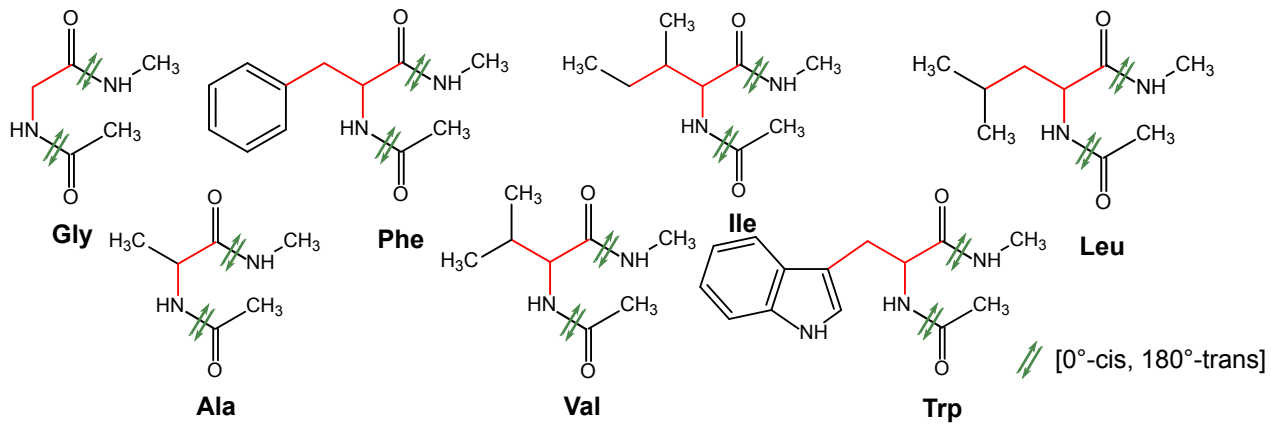

Figure 2: Chemical structures of the amino acid dipeptides. Rotatable bond are single, non-ring bonds between non-terminal atoms that are not attached to methyl groups that carry three identical substituents and are marked in red. Double arrows mark the *cis/trans* bond.

Mycophenolic acid

From the Astex Diverse Set,[84] a collection of X-ray crystal structures of complexes containing ligands from the Protein Data Bank (PDB), one example for a drug-like ligand was selected. This molecule, mycophenolic acid (target protein: 1MEH) has 43 atoms, 8 rotatable bonds and 1 *cis/trans* bond (Figure 3).

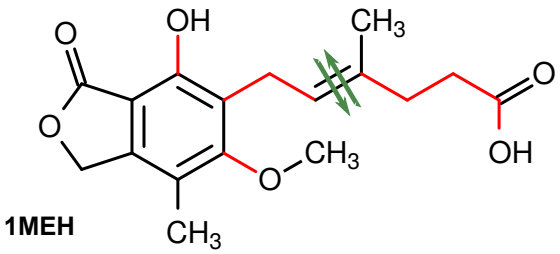

Figure 3: The chemical structure of the selected ligand together with the PDB-ID of the respective X-ray structure of the target protein. Rotatable bonds are marked in red and the *cis/trans* bond is marked with double arrows.

Mycophenolic acid is a very flexible molecule. Even a coarse systematic search with a grid of only 60° for the freely rotatable torsions and 2 values (cis/trans) for the double bond and the X-X-O-H torsions yields already $6^6 * 2 * 2 * 2 = 373248$ conformations to test. This makes this molecule a challenging example to test the performance of three search techniques (A-C below) in combination with first-principles methods.

A) *Genetic algorithm*. 50 independent GA runs with following settings: **max_iter**=30, **iter_limit_conv**=20 and **popsize**=10, were performed with Fafoom. A total of 3208 structures were generated.

B) *Random search*. 3200 random and clash-free structures were generated with Fafoom and further relaxed with DFT.

C) *Systematic search with Confab*[11]. First, 293 conformers were generated with Confab (assessed via Open Babel, used settings: **RMSD cutoff** = 0.65 and **Energy cutoff** = 15 kcal/mol). In order to account for two different values for the *cis/trans* bond and the X-X-O-H torsions (0° and 180°), eight starting structures per each of the conformers generated with Confab were considered. This procedure yields overall 2344 structures. After removing geometries with clashes, **2094** structures were subjected to DFT optimization.

Finally, all DFT optimized structures were merged to a common pool and the duplicates were removed. For this, a two-step criterion was used. First, the compared structures need to have a torsional RMSD (tRMSD) lower than 0.1π rad.[85] Second, the energy difference between the compared structures cannot exceed 10 meV. If both criteria are met, the structure that is higher in energy is labelled as 'duplicate' and is removed from the pool. In total, 1436 unique structures were found. Table S1 shows the number of the obtained unique structures depending on the applied energy cutoff.

Results and discussion

The performance of the GA search is evaluated by the ability to reproduce the reference energy hierarchies and to find the global minimum. We performed multiple GA runs for the test systems to test the impact of varying search settings.

Amino acid dipeptides

For each of the amino acid dipeptides we performed 50 independent GA runs with 10 iterations (**max_iter**) each and a population size of 5 (**popsize**). One GA run with such settings requires

popsize $+ 2 \cdot$ **iterations** $= 25$ geometry optimizations at the PBE+vdW level and yields 25 conformers.

Finding the global minimum

First we assess the probability to find the global minimum (known from the reference energy hierarchy) among them. We check how many of the GA runs succeed in finding the global minimum and subsequently calculate the probability for finding the global minimum in one GA run and present the results in Table 4.

Table 4: Average (from 50 GA runs) probability for finding the energy global minimum in a given run with 25 locally optimized conformers.

Molecule	Gly	Ala	Phe	Val	Trp	Leu	Ile
TDOFs	4	4	6	5	6	6	6
Probability for global minimum (/1 run)	0.82	0.79	0.53	0.60	0.22	0.20	0.10

Table 4 illustrates how the magnitude of the sampling problem does not only depend on the dimensionality, i.e. here the number of TDOFs, but also on the chemical structure. Phenylalanine and isoleucine are two interesting cases, both have the same number of TDOFs and are of similar size, but the probability of finding the global minimum with a single run drops dramatically. The drop in probability is, of course, correlated with the overall number of conformers listed in Table 3.

Conformational coverage

A key point in our approach is to reproduce the known energy hierarchies of the reference systems. For each of the investigated compounds, we randomly choose 5, 10, 15, 20, and 25 runs (from the pool of 50 runs), merge the results, and check how many structures have been found. We repeat this procedure 10,000 times and present the results in Figure 4.

It is evident that for dipeptides with a small number of reference minima (alanine and glycine) we obtain a very good results, i.e. a very good coverage of conformational space, already with 5 repeats of the GA runs. For dipeptides with a slightly higher number of minima (phenylalanine and

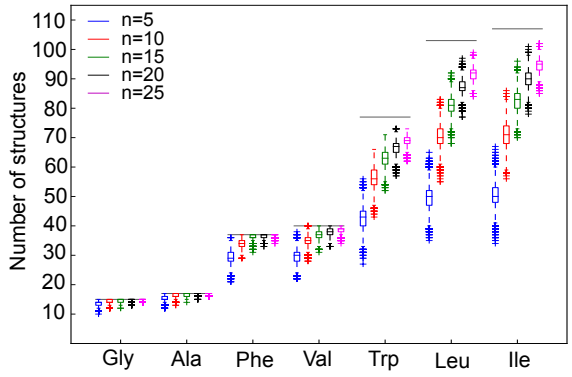

Figure 4: Number of minima found by the GA for seven amino acid dipeptides. The horizontal lines depict the total number of minima for the given compound as predicted by Ropo *et al.*[83] From a total of 50 GA runs, 5, 10, 15, 20, 25 GA runs were randomly selected and the found structures were counted. This procedure was repeated 10,000 times and the resulting distributions are summarized in box plots. The line inside the box is the median, the bottom and the top of the box are given by the lower ($Q_{0.25}$) and upper ($Q_{0.75}$) quartile. The length of the whisker is given by $1.5 \cdot (Q_{0.75} - Q_{0.25})$. Outliers (any data not included between the whiskers) are plotted as a cross.

valine) at least 10 runs of the GA are needed to obtain a good result. For the remaining dipeptides, the GA is not able to find all of the reference minima, even with 25 GA runs. However, the coverage of the reference hierarchy with 20 GA runs is always higher than 80%. We next inspect in more detail which of the amino acid dipeptides' reference minima were missed. To this end we investigate the actual difference between the reference hierarchy and the hierarchy obtained from the 50 GA runs, see Figure 5. Although our search strategy misses a few of the reference structures even when 50 repeats of the GA search are performed, the first missed structure has a relative energy higher than 0.2 eV. This in turn means that no low-energy structures are being missed. Furthermore, there are multiple newly predicted structures that were not present in the reference data set (Figure 5). It should be noted that, considering the fact that the investigated GA runs are rather short, the random component of the search (randomly initialized populations) contributes to the good results of the search.

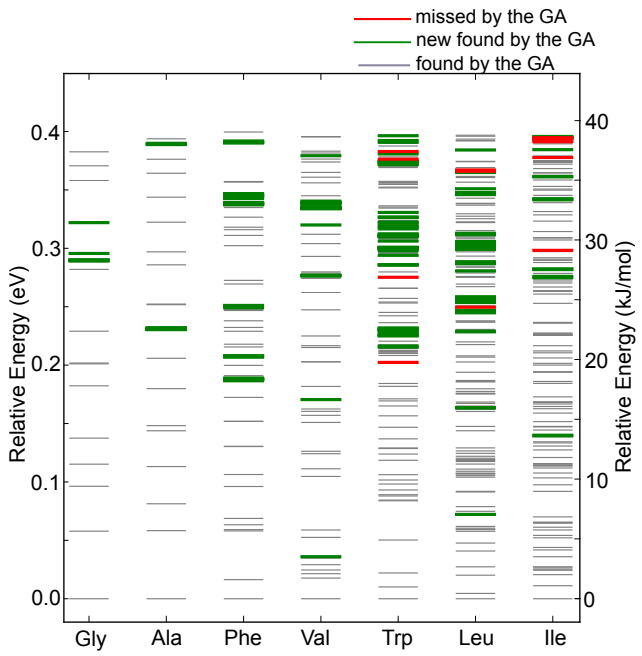

Figure 5: Difference hierarchies for the amino acid dipeptides. Red lines depict structures from the reference data set that have not been found by the GA. Green lines depict structures found by the GA that were absent in the reference data set. Gray lines depict structures from the reference data set that were found by the GA. The results from all 50 GA runs for each dipeptide were taken into account.

Parameter sensitivity

In order to check the robustness of the default run parameters, several alternative settings were tested for the isoleucine dipeptide. The tested parameters include: (i) the impact of the selection mechanism (roulette wheel, reverse roulette wheel, random), (ii) the effect of decreasing the cut-off for blacklisting from the default value of 0.2 Å to 0.05 Å, and (iii) the increase of the maximal number of iterations from the default 10 to 15, 20 and 25. For cases (i) and (ii) 100 GA runs were performed for each of the settings. In order to assess the effect of the number of iterations, 100 runs with a maximal number of iterations equal to 25 have been performed and subsequently only considered up to a maximum of 15, 20, 25 maximum iterations. Additionally, 50 GA runs with a maximal number of iterations equal to 100 were performed. In all mentioned cases convergence criteria were evaluated after each iteration, starting from the **iter_limit_conv**=10th iteration.

We find that none of the three selection mechanisms has a distinct impact on the probability

for finding the global minimum or quality of the conformational coverage. Similarly, no substantial change was observed upon the decrease of the blacklisting cut-off. The probability value for finding the global minimum as well as the number of found reference minima increases with increased number of iterations. This is simply due to the increased number of trials for sampling the conformational space. Table 5 summarizes the probability to find the global minimum in one run with different settings. Detailed data about the conformational coverage is given in Figure S2.

Table 5: Probability of finding the global minimum of isoleucine in one run for different setups. The default settings include roulette wheel selection mechanism, 0.2 Å cut-off for the blacklisting and maximal number of iteration equal to 10. The numbers in brackets denote the mean number of iterations needed for convergence.

Setup		Probability of finding the global minimum (per run)
default		0.17
Selection mechanism	roulette wheel reverse	0.18
	random	0.13
Max. number of iterations	15 (13)	0.20
	20 (15)	0.25
	25 (16)	0.25
	100 (22)	0.46
Cut-off for blacklisting	0.05 Å	0.14

Evaluation of the computational performance

The accuracy of a search/sampling strategy is its most crucial feature. Nevertheless, its computational cost plays a significant role in practical applications. To this end, we quantify the total cost of the GA runs in terms of force evaluations required in the local geometry optimizations. The number of force evaluations, i.e. most expensive steps in the algorithm, is a suitable measure for the computational cost. One force evaluation requires approximately between 1 (glycine) to 3 (tryptophan) CPUminutes. We quantify the number of force evaluations required by the GA for reproducing 85% of the reference hierarchy and present the results in Table 6. The table also includes the number of force evaluations required only in the replica-exchange MD refinement step of the reference search (the number of force evaluations required for the geometry optimizations is not even included).

Table 6: Comparison of the computational cost: amino acid dipeptides. The cost is given in the total number of force evaluations [$*10^3$].

	Total number of force evaluations[$*10^3$]						
	Gly	Ala	Phe	Val	Trp	Leu	Ile
GA (at least 85% reproduction of the reference hierarchy)	11	12	29	24	60	68	61
Reference	380	400	480	440	500	460	460

Mycophenolic acid

In the following we utilize as reference a set of structures that is a result of merging all structures found by three techniques: 3208 structures from 50 GA runs, 3200 random structures and 2094 structures generated with Confab. We define the following subsets: (i) 'GA' is a random selection of 25 GA runs (approx. 1600 structures); (ii) 'SYS (CONFAB)' is a set of all 2094 structures generated in the systematic search; and (iii) 'RANDOM' is random selection of 1600 structures generated in the random search. For the performance evaluation we count how many of the reference structures can be found by the respective search technique. This procedure was repeated 1000 times for each of the energy cutoffs. The results are shown in Figure 6. More details can be found in Table S1.

All of the search techniques found the same global minimum several times. In case if no energy cutoff is applied, none of the searches is able to find all local minima in the conformational space (i.e. more calculation would be needed). With a decreasing energy cutoff, an improved coverage of the conformational space can be observed. The fact that the GA is a global optimization techniques is clearly visible as it performs better in the low-energy (< 0.2 eV) region, whereas the random and systematic search perform uniformly but not perfectly independent of the energy cutoff used for the evaluation.

In order to show the wide and routine applicability of our first-principles structure search approach, we have performed short exploratory structure searches (only 3 GA runs each) to eight drug-like ligands from the Astex Diverse Set, which is widely utilized to assess the performance of, for example, conformer generators. The molecules vary in the size (8 to 32 atoms) and number

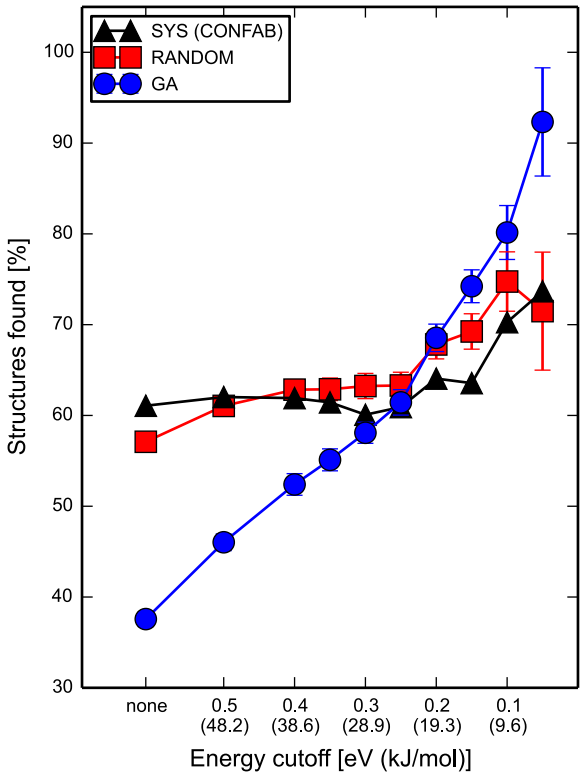

Figure 6: Share of the reference number of structures found by three search techniques: GA (blue circles), random search (red squares) and systematic search with Confab (black triangles) as a function of the applied energy cutoff. Energy values are given in eV and in parentheses also in kJ/mol.

of rotatable bonds (6 to 13). A detailed analysis of this study is shown in the Supporting Information. In brief we find that in all eight instances a diverse pool of conformers can be generated. In each case, a conformer is found that is similar to the protein-bound ligand from the X-ray structure with an RMSD of 1.5 Å. In six of the eight instances, they are similar with RMSD values of less than 0.9 Å. Exploratory first-principles structure searches have a potential application in *in silico* protein-ligand studies:[59] the comparison of the structural space of the isolated ligand and the structure realized by the protein-bound ligand might reveal details about the binding process, for example whether the binding mechanism follows more the conformational-selection or induced-fit type. In contrast to many of the quicker (but simpler) established conformer generators, the first principles energetics that we obtain here are not dependent on initial parametrizations and thus the method is in principle applicable throughout chemical space. It is important to note that, in this test, our goal was not to provide a converged GA search for each molecule but rather to explore the GA's potential to provide approximate conformational coverage with a fixed computational budget. Our investigation of mycophenolic acid indicates that searches for each of these molecule could be reliably converged albeit at significantly higher computational expense.

Literature context

In order to put the algorithm's parameters into perspective, we compare it to four selected applications of EA or GA to the conformational search of molecules in the following. In all considered algorithms, the initial populations are generated randomly and the conformational space of the respective molecules is represented and sampled (by mutation and crossing-over) by means of torsion angles, i.e. rotations around bonds. Table 7 summarizes a few parameters that illustrate the range over which the parameters that are characteristic to these kinds of evolutionary or genetic algorithms can vary. The approaches differ in the energy functions that are employed: Damsbo *et al.*[37] employ the CHARMM force field, Vainio and Johnson[21] use the torsional and the vdW term of the MMFF94 force field separately in a multi-objective genetic algorithm (MO-GA) while Nair and Goodman[36] use the MM2* force field. The study on optimizing the GA parameters

for molecular search with a meta-GA, presented by Brain and Addicoat,[56] uses, similar to our work, a first-principles energy functions. Two choices in the algorithm highlight the difference between theirs and our aim: in order "to reliably find the already known *a priori* correct answer with minimum computational resources", the selection criterion 'rank' focuses on the generation's best solution. Furthermore, crossing-over is considered as not helpful. In contrast, the aim of our work is to provide a GA implementation that ensures broad conformational coverage, i.e. the prediction of an energy hierarchy and not only the reproduction of a global optimum. For that we found it useful to employ random or roulette-wheel selection that also accepts less-optimal structures for genetic operations and a high probability for crossing-over. Both choices (accompanied by blacklisting) can be interpreted as means to increase diversity during the search.

Table 7: Comparison of parameters and schemes that are used in search approaches proposed in four selected publications with the approach presented in this work.

Parameter	Damsbo'04[37]	Vainio'07[21]	Nair'98[36]	Brain'11[56]	this work
Algorithm type	EA	MO-GA	GA	GA	GA
Population size	30	20	2-20	10-15	5
Selection	-	tournament	roulette	rank	roulette
Crossing-over probability	-	0.9	1.0	0.0-1.0	0.95
Mutation probability	-	0.05	0.4	0.3-0.5	0.5

Conclusions

We aimed at designing a user-friendly framework with an implementation of the genetic algorithm for searches in molecular conformational space that is particularly suitable for flexible organic compounds. A SMILES code for the selected molecule is the only required input for the algorithm. Furthermore, a wide selection of parameters (e.g. torsion definition, blacklist cut-off) allows for customizing the search. With minor changes, the code can be interfaced to external packages for molecular simulations that output optimized geometries together with corresponding energies. Besides its adaptability and ease of use, a further advantage of the implementation is the fact that it allows for using first-principles methods. With this, a potential bias resulting from the

parametrization of a particular force-field can be avoided and makes the search applicable to a broad selection of problems. We examined the performance of the implementation in terms of efficiency and accuracy of the sampling. The algorithm is capable of reproducing the reference data with a high accuracy. For a set of amino acid dipeptides, we show that this conformational coverage is achieved much more efficiently than in an earlier, *ab initio* replica-exchange MD based search in our group. For a larger molecule (mycophenolic acid), we show that the low-energy conformational space coverage of the GA surpasses the coverage of two competing methods significantly at similar effort.

Acknowledgement

Matthias Scheffler (FHI Berlin) is kindly acknowledged for support of this work and scientific discussions.

Supporting Information

Crossover procedure

The following scheme (Figure 7) explains the crossover procedure described in the **Methods** section of the article.

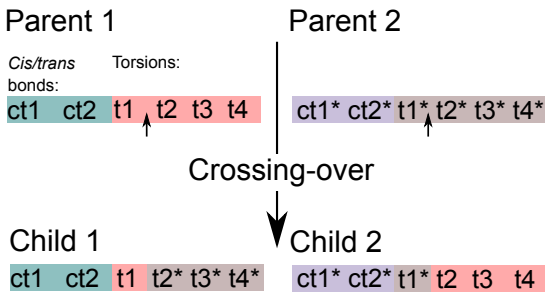

Figure 7: Crossing-over procedure. The lists of values for *cis/trans* bonds and torsions are first combined together. As next, a single cut is performed and the corresponding parts are exchanged.

Conformational coverage: Isoleucine

Figure 8 shows the coverage of the conformational space of isoleucine dipeptide resulting from GA searches with different settings.

Conformational coverage: Mycophenolic acid

Table 8 contains the total number of mycophenolic acid conformers found by three conformer generation techniques depending on the energy cutoff. Further, it contains detailed data used to create Figure 6 in the article. As there are three search techniques there are eight possible cases for finding a structure or not: (i) a structures can be found by all searches ('all') or (ii) a structure can be found by two of three searches ('*GA+RANDOM*','*GA+SYS*', '*SYS+RANDOM*') or (iii) a structure can be found by only one search ('*only GA*', '*only RANDOM*', '*only SYS*') or (iv) a structure can be missed by all of the searches (not included in the table).

Table 8: Share of found structures by the GA, systematic search and random conformer generation depending on the used energy cutoff.

Energy cutoff (eV) / (kJ/mol)	Nr. of structures	Unique structures	[%] of structures found by						
			only GA	only RANDOM	only SYS	GA + RANDOM	GA + SYS	RANDOM + SYS	all
no	8502	1436	6.28	14.00	19.23	6.45	5.20	17.00	19.65
0.5 (48.2)	7164	1006	7.18	12.04	15.26	8.13	5.90	16.05	24.82
0.4 (38.6)	6288	764	8.02	10.74	12.95	9.30	6.17	13.89	28.90
0.35 (33.8)	5452	625	8.43	10.30	12.69	10.22	6.39	12.29	30.07
0.3 (28.9)	4961	531	9.10	9.97	11.53	11.36	6.62	10.92	31.00
0.25 (24.1)	4535	458	9.53	8.39	11.23	12.38	7.16	10.16	32.37
0.2 (19.3)	4079	345	8.53	6.79	9.52	14.27	7.78	8.79	37.96
0.15 (14.5)	3153	225	9.51	5.64	8.17	15.71	7.49	6.37	41.53
0.1 (9.6)	1314	74	7.04	4.33	6.56	14.07	7.36	4.67	51.68
0.05 (4.8)	298	19	6.87	2.81	1.04	15.10	19.05	2.28	51.31

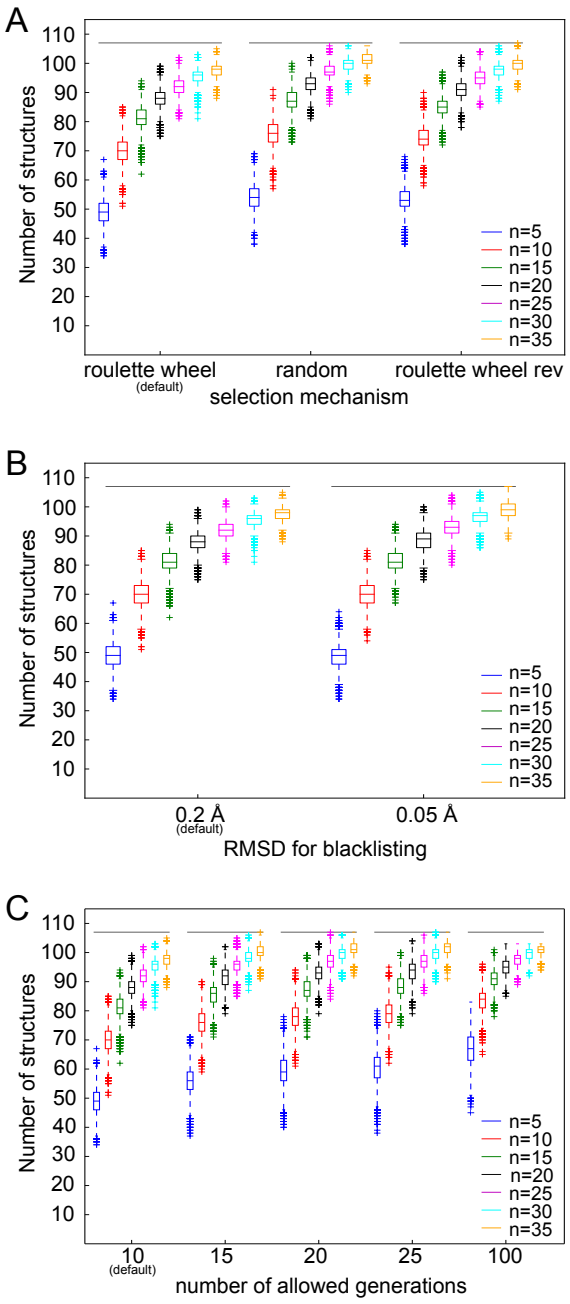

Figure 8: Conformational coverage of the GA search with different settings for Ile. The horizontal lines depict the number of reference minima[83] (107). From a total of 100 GA runs, 5, 10, 15, 20, 25, 30, 35 GA runs were randomly selected and the found structures were counted. This procedure was repeated 10,000 times and the resulting distributions are summarized in box plots. The line inside the box is the median, the bottom and the top of the box are given by the lower ($Q_{0.25}$) and upper ($Q_{0.75}$) quartile. The length of the whisker is given by $1.5 \cdot (Q_{0.75} - Q_{0.25})$. Any data not included between the whiskers is plotted as an outlier with a cross. Conformational coverage hardly changes by using different selection mechanisms (A) or changing the blacklisting cut-off (B). (C) Increasing the number of GA iterations improves conformational coverage.

A small set of flexible organic molecules

We utilize the Astex Diverse Set,[84] a collection of structures obtained from X-ray crystal structures from the Protein Data Bank (PDB), to construct the a small set of flexible organic molecules. The goal here is not to provide a converged GA search for each molecule but rather to explore the GA's potential to provide approximate conformational coverage with a fixed computational budget. Our investigation of mycophenolic acid indicates that searches for each of these molecule could be reliably converged albeit at significantly higher computational expense.

We selected 8 ligands (Figure 9) that differ in composition, the number of heavy atoms (15-32) and the number of rotatable bonds (6-13).[86]

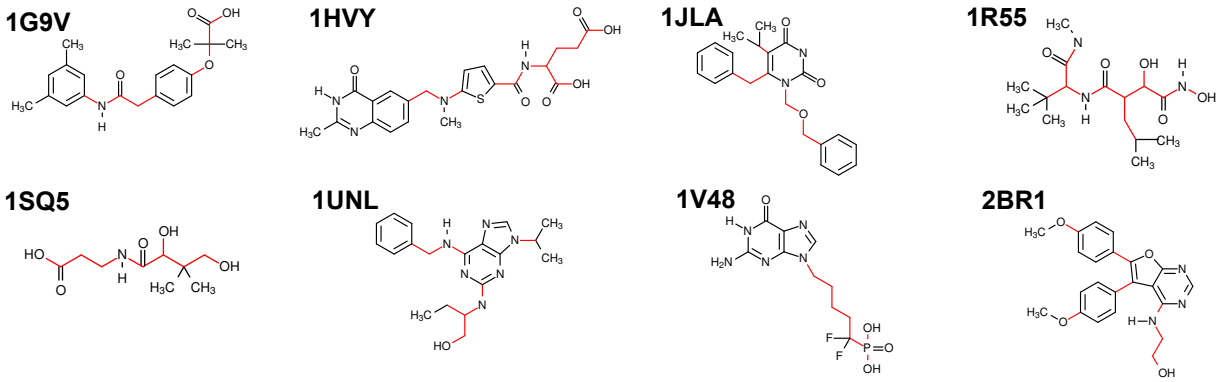

Figure 9: Chemical structures of the selected ligands together with the PDB-IDs of the respective X-ray structures of the target proteins. Rotatable bonds are marked in red.

Genetic algorithm searches

SMILES codes for the respective entries were taken from PubChem to ensure an unbiased starting point for the search. We utilize the parameter values as listed in Table 2 in the article with following exceptions: **max_iter**=30, **iter_limit_conv**=20, **popsize**=10, **prob_for_mut_rot**=0.8, **prob_for_mut_cistrans**=1, **cross_trial**=100 and **max_mutations_torsion**=3. For each of the molecules, three GA runs have been performed, i.e. given the settings, the number of obtained conformers cannot be higher than 210.

Results

We present the summary of the results in Table 9. The number of found conformers is obtained after removing duplicates among all obtained conformers. For each of the molecules, we calculate the RMSD between the non-hydrogen atoms of each of the obtained structures and the reference ligand (Figure 10A) (hydrogens in the Astex Diverse Set set are the result of modeling and not part of the experimental result). With this, we identify the *best match*, i.e. the conformer which is most similar to the reference ligand. Furthermore, the reference ligand structures were optimized with DFT and are added to the respective plots for completeness. Figure 10B shows the overlay between the reference ligand (before the DFT optimization) and the *best match* for all molecules.

Table 9: Selected ligands from the Astex Diverse Set. The number of found conformers is obtained after removing duplicates among all obtained conformers. The *best match* is the conformer which is most similar to the reference ligand.

Target protein	No. of heavy atoms	No. of rotatable bonds	No. of found conformers	RMSD (Å) between the ligand and the		ΔE (eV) between the GA minimum and the	
				best match	GA minimum	*best match*	optimized ligand
1G9V	25	8	70	1.43	1.66	0.536	0.035
1HVY	32	10	176	1.2	2.73	0.707	0.505
1JLA	27	7	41	0.56	1.44	0.339	0.268
1R55	23	13	116	0.88	1.59	0.315	0.326
1SQ5	15	10	152	0.77	2.14	0.661	0.555
1UNL	26	9	166	0.65	2.23	0.076	0.026
1V48	22	6	118	0.72	2.23	0.696	0.722
2BR1	29	8	73	0.28	0.63	0.005	0.002

For all investigated systems, we obtain a large number of conformers that are spread over a wide energy window. This satisfies our primary goal of obtaining a diverse set of conformers with a reliable energy hierarchy in a straightforward fashion. Moreover, in most of the cases, the RMSD between the *best match* and reference ligand is satisfactory (i.e. below 1.0 Å). Here we would like to note, that our energy evaluations are performed in the gas phase while the reference ligand is obtained from a X-ray crystal structure. The energy of the *best match* is significantly higher than the energy of the GA minimum in most of the cases. This finding supports the need for providing a

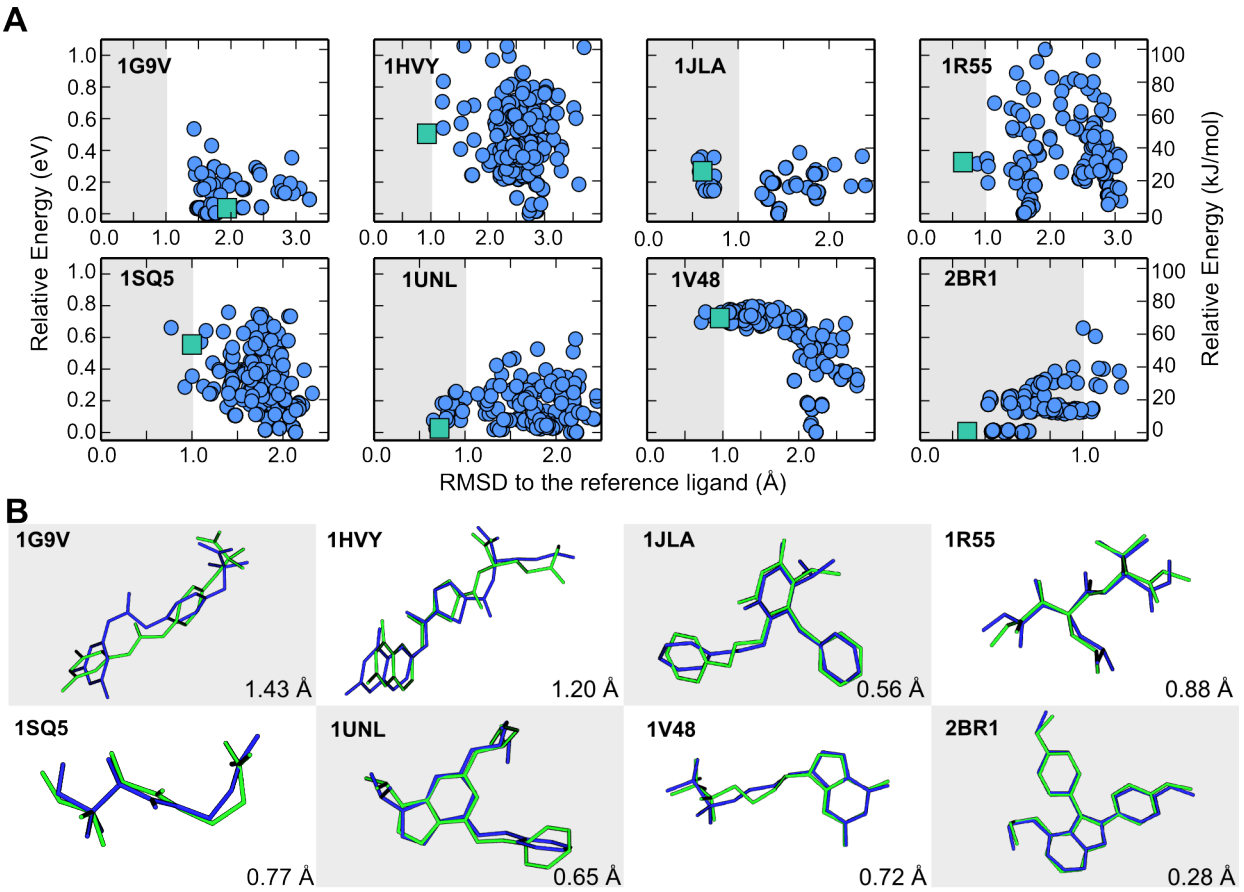

Figure 10: Evaluation of the results for the subset of the Astex Diverse Set. (A) Relative energy of all found conformers as a function of the RMSD to the reference ligand (blue circles). The green squares depict the reference ligand structures after DFT optimization. (B) An overlay between the reference ligand (green) and the *best match* (blue) is presented together with the corresponding RMSD value.

broad range of conformers instead of only focusing on the global minimum of the particular energy function.

A few cases require further analysis. For two ligands, with the targets 1G9V and 1HVY, the RMSD values between the *best match* and the reference ligand structure exceed the threshold value (1.0 Å). One possible reason is the fact, that the reference ligand is not a minimum on the PES sampled by the GA. Another trivial cause might be the insufficient number of the performed GA runs.

Further we note that the optimization of the orientation of the hydroxy groups is required for obtaining a meaningful conformational ensemble.

Apart from performing short GA runs, one long GA run has been performed for each of the selected molecules for comparison. In order to obtain comparable results (by means of the number of performed DFT optimization), following parameters have been adjusted: **max_iter**=80, **iter_limit_conv**=70. We compare the results in terms of: (i) energy of the most stable structure, (ii) matching the reference ligand and (iii) number of found conformers. Detailed data about the results of the single long runs are given in Table 10. In terms of finding the *best match*, the long GA run performs better than the three short GA runs together for some of the molecules. On the other hand, the number of found structures is significantly higher if three short GA runs are performed instead of a single long run for most of the molecules.

The results of the exploratory structure searches for the 8 drug-like ligands suggest that performing one long GA run instead of 3 short GA runs may increase the chance for finding the global minimum and simultaneously decrease the number of identified unique conformers.

Table 10: Comparison of the results obtained from one long GA run (max. 80 iterations) and three short GA runs (each max. 30 iterations). ΔE is the difference between the most stable structures found in the compared setups.

Molecule	Number of found conformers		RMSD (Å) between the ligand and the *best match*		ΔE (eV)
	1 x max. 80 iterations	3 x max. 30 iterations	1 x max. 80 iterations	3 x max. 30 iterations	
1G9V	45	70	0.57	1.43	0.0
1HVY	146	176	1.07	1.2	0.211
1JLA	40	41	0.62	0.56	-0.005
1R55	85	116	0.69	0.88	0.081
1SQ5	99	152	0.64	0.77	0.031
1UNL	93	166	0.8	0.65	0.003
1V48	115	118	0.81	0.72	0.054
2BR1	47	73	0.28	0.28	-0.005

References

(1) Schwab, C. H. Conformations and 3D pharmacophore searching. *Drug Discov. Today. Technol.* **2010**, *7*, e245–e253.

(2) Kuntz, I. D.; Blaney, J. M.; Oatley, S. J.; Langridge, R.; Ferrin, T. E. A geometric approach to macromolecule-ligand interactions. *J. Mol. Biol.* **1982**, *161*, 269–288.

(3) Jones, G.; Willett, P.; Glen, R. C.; Leach, A. R.; Taylor, R. Development and validation of a genetic algorithm for flexible docking. *J. Mol. Biol.* **1997**, *267*, 727–748.

(4) Morris, G. M.; Goodsell, D. S.; Halliday, R. S.; Huey, R.; Hart, W. E.; Belew, R. K.; Olson, A. J. Automated docking using a Lamarckian genetic algorithm and an empirical binding free energy function. *J. Comput. Chem.* **1998**, *19*, 1639–1662.

(5) Meier, R.; Pippel, M.; Brandt, F.; Sippl, W.; Baldauf, C. ParaDockS: a framework for molecular docking with population-based metaheuristics. *J. Chem. Inf. Model.* **2010**, *50*, 879–889.

(6) Kristam, R.; Gillet, V. J.; Lewis, R. A.; Thorner, D. Comparison of conformational analysis techniques to generate pharmacophore hypotheses using catalyst. *J. Chem. Inf. Model.* **2005**, *45*, 461–476.

(7) Kirchmair, J.; Laggner, C.; Wolber, G.; Langer, T. Comparative analysis of protein-bound ligand conformations with respect to catalyst's conformational space subsampling algorithms. *J. Chem. Inf. Model.* **2005**, *45*, 422–430.

(8) Agrafiotis, D. K.; Gibbs, A. C.; Zhu, F.; Izrailev, S.; Martin, E. Conformational sampling of bioactive molecules: a comparative study. *J. Chem. Inf. Model.* **2007**, *47*, 1067–1086.

(9) Li, J.; Ehlers, T.; Sutter, J.; Varma-O'Brien, S.; Kirchmair, J. CAESAR: a new conformer generation algorithm based on recursive buildup and local rotational symmetry consideration. *J. Chem. Inf. Model.* **2007**, *47*, 1923–1932.

(10) O'Boyle, N. M.; Banck, M.; James, C. A.; Morley, C.; Vandermeersch, T.; Hutchison, G. R. Open Babel: An open chemical toolbox. *J. Cheminform.* **2011**, *3*, 33.

(11) O'Boyle, N. M.; Vandermeersch, T.; Flynn, C. J.; Maguire, A. R.; Hutchison, G. R. Confab - Systematic generation of diverse low-energy conformers. *J. Cheminform.* **2011**, *3*, 8.

(12) Mohamadi, F.; Richards, N. G. J.; Guida, W. C.; Liskamp, R.; Lipton, M.; Caufield, C.; Chang, G.; Hendrickson, T.; Still, W. C. Macromodel - An integrated software system for modeling organic and bioorganic molecules using molecular mechanics. *J. Comput. Chem.* **1990**, *11*, 440–467.

(13) *MOE (Molecular Operating Environment)*; Chemical Computing Group, Inc.: Montreal, Canada, 2008.

(14) Klett, J.; Cortés-Cabrera, A.; Gil-Redondo, R.; Gago, F.; Morreale, A. ALFA: Automatic Ligand Flexibility Assignment. *J. Chem. Inf. Model.* **2014**, *54*, 314–323.

(15) Schärfer, C.; Schulz-Gasch, T.; Hert, J.; Heinzerling, L.; Schulz, B.; Inhester, T.; Stahl, M.; Rarey, M. CONFECT: Conformations from an Expert Collection of Torsion Patterns. *ChemMedChem* **2013**, 1690–1700.

(16) Sadowski, J.; Gasteiger, J.; Klebe, G. Comparison of Automatic Three-Dimensional Model Builders Using 639 X-ray Structures. *J. Chem. Inf. Comput. Sci.* **1994**, *34*, 1000–1008.

(17) Renner, S.; Schwab, C. H.; Gasteiger, J.; Schneider, G. Impact of conformational flexibility on three-dimensional similarity searching using correlation vectors. *J. Chem. Inf. Model.* **2006**, *46*, 2324–2332.

(18) Andronico, A.; Randall, A.; Benz, R. W.; Baldi, P. Data-driven high-throughput prediction of the 3-D structure of small molecules: review and progress. *J. Chem. Inf. Model.* **2011**, *51*, 760–776.

(19) Sadowski, P.; Baldi, P. Small-molecule 3D structure prediction using open crystallography data. *J. Chem. Inf. Model.* **2013**, *53*, 3127–3130.

(20) Hawkins, P. C. D.; Skillman, A. G.; Warren, G. L.; Ellingson, B. A.; Stahl, M. T. Conformer generation with OMEGA: algorithm and validation using high quality structures from the Protein Databank and Cambridge Structural Database. *J. Chem. Inf. Model.* **2010**, *50*, 572–584.

(21) Vainio, M. J.; Johnson, M. S. Generating conformer ensembles using a multiobjective genetic algorithm. *J. Chem. Inf. Model.* **2007**, *47*, 2462–2474.

(22) Liu, X.; Bai, F.; Ouyang, S.; Wang, X.; Li, H.; Jiang, H. Cyndi: a multi-objective evolution algorithm based method for bioactive molecular conformational generation. *BMC Bioinformatics* **2009**, *10*, 101.

(23) RDKit: Cheminformatics and Machine Learning Software. `http://www.rdkit.org/`.

(24) Wales, D. J.; Doye, J. P. K. Global Optimization by Basin-Hopping and the Lowest Energy Structures of Lennard-Jones Clusters Containing up to 110 Atoms. *J. Phys. Chem. A* **1997**, *101*, 5111–5116.

(25) Goedecker, S. Minima hopping: an efficient search method for the global minimum of the potential energy surface of complex molecular systems. *J. Chem. Phys.* **2004**, *120*, 9911–9917.

(26) Bahn, S.; Jacobsen, K. An object-oriented scripting interface to a legacy electronic structure code. *Comput. Sci. Eng.* **2002**, *4*, 56–66.

(27) Wales, D. J. GMIN: A program for finding global minima and calculating thermodynamic properties from basin-sampling. `http://www-wales.ch.cam.ac.uk/GMIN/`.

(28) Ponder, J. W. Tinker - Software Tools for Molecular Design. `http://dasher.wustl.edu/tinker/`.

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

(30) Fogel, D. B., Ed. *Evolutionary Computation: The Fossil Record*; IEEE Press: Piscataway, NJ, 1998.

(31) Goldberg, D. E. *Genetic algorithms in search, optimization, and machine learning*; Addison-Wesley: Reading, MA, 1989.

(32) Clark, D. E.; Westhead, D. R. Evolutionary algorithms in computer-aided molecular design. *J. Comput. Aided. Mol. Des.* **1996**, *10*, 337–358.

(33) Wales, D. J. *Energy Landscapes: Applications to Clusters, Biomolecules and Glasses*; Cambridge University Press: Cambridge, 2003.

(34) Wales, D. J.; Scheraga, H. A. Global optimization of clusters, crystals, and biomolecules. *Science* **1999**, *285*, 1368–1372.

(35) Johnston, R. L., Ed. *Applications of Evolutionary Computation in Chemistry*; Structure and Bonding; Springer Berlin Heidelberg: Berlin, Heidelberg, 2004.

(36) Nair, N.; Goodman, J. M. Genetic Algorithms in Conformational Analysis. *J. Chem. Inf. Comput. Sci.* **1998**, *38*, 317–320.

(37) Damsbo, M.; Kinnear, B. S.; Hartings, M. R.; Ruhoff, P. T.; Jarrold, M. F.; Ratner, M. A. Application of evolutionary algorithm methods to polypeptide folding: comparison with experimental results for unsolvated Ac-(Ala-Gly-Gly)$_5$-LysH$^+$. *Proc. Natl. Acad. Sci. U. S. A.* **2004**, *101*, 7215–7222.

(38) Carstensen, N. O.; Dieterich, J. M.; Hartke, B. Design of optimally switchable molecules by genetic algorithms. *Phys. Chem. Chem. Phys.* **2011**, *13*, 2903–2910.

(39) Carlotto, S.; Orian, L.; Polimeno, A. Heuristic approaches to the optimization of acceptor systems in bulk heterojunction cells: a computational study. *Theor. Chem. Acc.* **2012**, *131*, 1191.

(40) Hartke, B. Global geometry optimization of clusters using genetic algorithms. *J. Phys. Chem.* **1993**, *97*, 9973–9976.

(41) Deaven, D. M.; Ho, K. M. Molecular geometry optimization with a genetic algorithm. *Phys. Rev. Lett.* **1995**, *75*, 288–291.

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

(43) Johnston, R. L. Evolving better nanoparticles: Genetic algorithms for optimising cluster geometries. *Dalt. Trans.* **2003**, 4193–4207.

(44) Blum, V.; Hart, G. L. W.; Walorski, M. J.; Zunger, A. Using genetic algorithms to map first-principles results to model Hamiltonians: Application to the generalized Ising model for alloys. *Phys. Rev. B* **2005**, *72*, 165113.

(45) Schönborn, S. E.; Goedecker, S.; Roy, S.; Oganov, A. R. The performance of minima hopping and evolutionary algorithms for cluster structure prediction. *J. Chem. Phys.* **2009**, *130*, 144108.

(46) Sierka, M. Synergy between theory and experiment in structure resolution of low-dimensional oxides. *Prog. Surf. Sci.* **2010**, *85*, 398–434.

(47) Bhattacharya, S.; Levchenko, S. V.; Ghiringhelli, L. M.; Scheffler, M. Stability and Metastability of Clusters in a Reactive Atmosphere: Theoretical Evidence for Unexpected Stoichiometries of Mg_MO_x. *Phys. Rev. Lett.* **2013**, *111*, 135501.

(48) Hartke, B. Global optimization. *Wiley Interdiscip. Rev.: Comput. Mol. Sci.* **2011**, *1*, 879–887.

(49) Heiles, S.; Johnston, R. L. Global optimization of clusters using electronic structure methods. *Int. J. Quantum Chem.* **2013**, *113*, 2091–2109.

(50) Hart, G. L. W.; Blum, V.; Walorski, M. J.; Zunger, A. Evolutionary approach for determining first-principles hamiltonians. *Nature Mater.* **2005**, *4*, 391–394.

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

(52) Oganov, A. R.; Glass, C. W. Crystal structure prediction using ab initio evolutionary techniques: principles and applications. *J. Chem. Phys.* **2006**, *124*, 244704.

(53) Woodley, S. M.; Catlow, R. Crystal structure prediction from first principles. *Nature Mater.* **2008**, *7*, 937–946.

(54) Tipton, W. W.; Hennig, R. G. A grand canonical genetic algorithm for the prediction of multi-component phase diagrams and testing of empirical potentials. *J. Phys. Condens. Matter* **2013**, *25*, 495401.

(55) Neiss, C.; Schooss, D. Accelerated cluster structure search using electron diffraction data in a genetic algorithm. *Chem. Phys. Lett.* **2012**, *532*, 119–123.

(56) Brain, Z. E.; Addicoat, M. A. Optimization of a genetic algorithm for searching molecular conformer space. *J. Chem. Phys.* **2011**, *135*, 174106.

(57) Baldauf, C.; Pagel, K.; Warnke, S.; von Helden, G.; Koksch, B.; Blum, V.; Scheffler, M. How cations change peptide structure. *Chem. Eur. J.* **2013**, *19*, 11224–11234.

(58) Rossi, M.; Chutia, S.; Scheffler, M.; Blum, V. Validation Challenge of Density-Functional Theory for Peptides-Example of Ac-Phe-Ala_5-$LysH^+$. *J. Phys. Chem. A* **2014**, *118*, 7349–7359.

(59) Avgy-David, H. H.; Senderowitz, H. *J. Chem. Inf. Model.*, 2015, DOI: 10.1021/acs.jcim.5b00259.

(60) Wu, Q.; Yang, W. Empirical correction to density functional theory for van der Waals interactions. *J. Chem. Phys.* **2002**, *116*, 515.

(61) Sedlak, R.; Janowski, T.; Pitoňák, M.; Řezáč, J.; Pulay, P.; Hobza, P. The accuracy of quantum chemical methods for large noncovalent complexes. *J. Chem. Theory Comput.* **2013**, *9*, 3364–3374.

(62) Tkatchenko, A.; Rossi, M.; Blum, V.; Ireta, J.; Scheffler, M. Unraveling the Stability of Polypeptide Helices: Critical Role of van der Waals Interactions. *Phys. Rev. Lett.* **2011**, *106*, 118102.

(63) Grimme, S.; Antony, J.; Schwabe, T.; Mück-Lichtenfeld, C. Density functional theory with dispersion corrections for supramolecular structures, aggregates, and complexes of (bio)organic molecules. *Org. Biomol. Chem.* **2007**, *5*, 741–758.

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

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

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

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

(68) Schubert, F.; Rossi, M.; Baldauf, C.; Pagel, K.; Warnke, S.; von Helden, G.; Filsinger, F.; Kupser, P.; Meijer, G.; Salwiczek, M.; Koksch, B.; Scheffler, M.; Blum, V. Exploring the conformational preferences of 20-residue peptides in isolation: Ac-Ala_{19}-Lys+H^+vs. Ac-Lys-Ala_{19}+H^+ and the current reach of DFT. *Phys. Chem. Chem. Phys.* **2015**,

(69) Blum, V.; Gehrke, R.; Hanke, F.; Havu, P.; Havu, V.; Ren, X.; Reuter, K.; Scheffler, M. Ab initio molecular simulations with numeric atom-centered orbitals. *Comput. Phys. Commun.* **2009**, *180*, 2175–2196.

(70) Havu, V.; Blum, V.; Havu, P.; Scheffler, M. Efficient integration for all-electron electronic structure calculation using numeric basis functions. *J. Comput. Phys.* **2009**, *228*, 8367–8379.

(71) Weininger, D. SMILES, a chemical language and information system. 1. Introduction to methodology and encoding rules. *J. Chem. Inf. Comput. Sci.* **1988**, *28*, 31–36.

(72) GNU Lesser General Public License. `https://www.gnu.org/licenses/lgpl.html`.

(73) Schlegel, H. B. Geometry optimization. *Wiley Interdiscip. Rev. Comput. Mol. Sci.* **2011**, *1*, 790–809.

(74) Cheng, J.; Fournier, R. Structural optimization of atomic clusters by tabu search in descriptor space. *Theor. Chem. Acc.* **2004**, *112*, 7–15.

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

(76) Nocedal, J.; Wright, S. J. *Numerical optimization*; Springer New York, 2006.

(77) Lindh, R.; Bernhardsson, A.; Karlström, G.; Malmqvist, P.-Å. On the use of a Hessian model function in molecular geometry optimizations. *Chem. Phys. Lett.* **1995**, *241*, 423–428.

(78) Marek, A.; Blum, V.; Johanni, R.; Havu, V.; Lang, B.; Auckenthaler, T.; Heinecke, A.; Bungartz, H.-J.; Lederer, H. The ELPA library: scalable parallel eigenvalue solutions for electronic structure theory and computational science. *J. Phys.: Condens. Matter.* **2014**, *26*, 213201.

(79) Valiev, M.; Bylaska, E. J.; Govind, N.; Kowalski, K.; Straatsma, T. P.; Van Dam, H. J. J.; Wang, D.; Nieplocha, J.; Apra, E.; Windus, T. L.; de Jong, W. A. NWChem: A comprehensive and scalable open-source solution for large scale molecular simulations. *Comput. Phys. Commun.* **2010**, *181*, 1477–1489.

(80) Halgren, T. A. Merck molecular force field. I. Basis, form, scope, parameterization, and performance of MMFF94. *J. Comput. Chem.* **1996**, *17*, 490–519.

(81) Tosco, P.; Stiefl, N.; Landrum, G. Bringing the MMFF force field to the RDKit: implementation and validation. *J. Cheminform.* **2014**, *6*, 37.

(82) We use the term *dipeptide* for amino acids with an acetylated N terminus and an aminomethylated C terminus.

(83) Ropo, M.; Baldauf, C.; Blum, V. Energy/structure database of all proteinogenic amino acids and dipeptides without and with divalent cations. *arXiv* **2015**, *q-bio.BM*, 1504.03708.

(84) Hartshorn, M. J.; Verdonk, M. L.; Chessari, G.; Brewerton, S. C.; Mooij, W. T. M.; Mortenson, P. N.; Murray, C. W. Diverse, high-quality test set for the validation of protein-ligand docking performance. *J. Med. Chem.* **2007**, *50*, 726–41.

(85) The value of 0.1π tRMSD corresponds to a 55° change of a single dihedral angle or to a change of 18° per each of 9 the dihedral angles.

(86) We treat all X-X-O-H torsion angles as rotatable bonds.