

Unsupervised search of low-lying conformers with spectroscopic accuracy: A two-step algorithm rooted into the island model evolutionary algorithm

Cite as: J. Chem. Phys. **153**, 124110 (2020); <https://doi.org/10.1063/5.0018314>

Submitted: 16 June 2020 • Accepted: 02 September 2020 • Published Online: 24 September 2020

 Giordano Mancini,  Marco Fusè,  Federico Lazzari, et al.

COLLECTIONS

Paper published as part of the special topic on [Machine Learning Meets Chemical Physics](#)



View Online



Export Citation



CrossMark

ARTICLES YOU MAY BE INTERESTED IN

[OrbNet: Deep learning for quantum chemistry using symmetry-adapted atomic-orbital features](#)

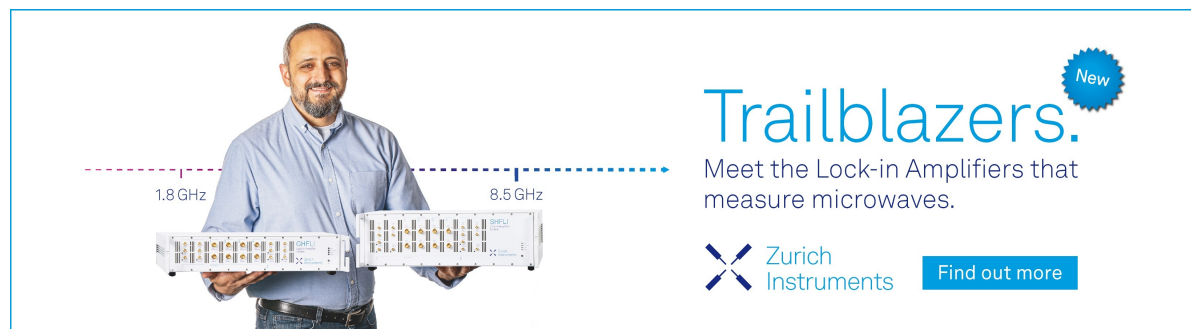
The Journal of Chemical Physics **153**, 124111 (2020); <https://doi.org/10.1063/5.0021955>


[The ORCA quantum chemistry program package](#)

The Journal of Chemical Physics **152**, 224108 (2020); <https://doi.org/10.1063/5.0004608>


[Molecular force fields with gradient-domain machine learning \(GDML\): Comparison and synergies with classical force fields](#)

The Journal of Chemical Physics **153**, 124109 (2020); <https://doi.org/10.1063/5.0023005>



Trailblazers. 

Meet the Lock-in Amplifiers that measure microwaves.

 Zurich Instruments [Find out more](#)

Unsupervised search of low-lying conformers with spectroscopic accuracy: A two-step algorithm rooted into the island model evolutionary algorithm

Cite as: J. Chem. Phys. 153, 124110 (2020); doi: 10.1063/5.0018314

Submitted: 16 June 2020 • Accepted: 2 September 2020 •

Published Online: 24 September 2020



View Online



Export Citation



CrossMark

Giordano Mancini,^{1,a)}  Marco Fusè,¹  Federico Lazzari,¹  Balasubramanian Chandramouli,² 
and Vincenzo Barone¹ 

AFFILIATIONS

¹Scuola Normale Superiore, Piazza dei Cavalieri 7, 56125 Pisa, Italy

²Super Computing Applications and Innovation, CINECA, Via Magnanelli, 6/3, Casalecchio di Reno, BO, Italy

Note: This paper is part of the JCP Special Topic on Machine Learning Meets Chemical Physics.

a) Author to whom correspondence should be addressed: giordano.mancini@sns.it

ABSTRACT

The fruitful interplay of high-resolution spectroscopy and quantum chemistry has a long history, especially in the field of small, semi-rigid molecules. However, in recent years, the targets of spectroscopic studies are shifting toward flexible molecules, characterized by a large number of closely spaced energy minima, all contributing to the overall spectrum. Here, artificial intelligence comes into play since it is at the basis of powerful unsupervised techniques for the exploration of soft degrees of freedom. Integration of such algorithms with a two-stage QM/QM' (Quantum Mechanical) exploration/refinement strategy driven by a user-friendly graphical interface is the topic of the present paper. We will address in particular: (i) the performances of different semi-empirical methods for the exploration step and (ii) the comparison between stochastic and meta-heuristic algorithms in achieving a cheap yet complete exploration of the conformational space for medium sized chromophores. As test cases, we choose three amino acids of increasing complexity, whose full conformer enumeration has been reached only very recently. Next, we show that systems in condensed phases can be treated at the same level and with the same efficiency when employing a polarizable continuum description of the solvent. Finally, the challenging issue represented by the vibrational circular dichroism spectra of some rhodium complexes with flexible ligands has been addressed, showing that our fully unsupervised approach leads to remarkable agreement with the experiment.

Published under license by AIP Publishing. <https://doi.org/10.1063/5.0018314>

I. INTRODUCTION

Artificial intelligence methods are increasingly employed with remarkable success in several fields of chemistry like, e.g., quantum chemistry,^{1,2} organic synthesis,^{3,4} spectroscopy,⁵ or analytical chemistry.⁶ Here, we will be concerned with the identification of low-lying energy minima (conformers, rotamers) related to the soft degrees of freedom of flexible molecular systems. This topic has been investigated in a large number of recent studies,^{7–10} pointing out that the search of the global energy minimum is not sufficient to

analyze the experimental outcome in a number of cases like, e.g., high-resolution spectroscopy^{11,12} or kinetics.^{13,14} In the same vein, the crystal structures of protein-ligand complexes have shown that bioactive conformations tend to be more extended than random ones¹⁵ and may lie up to a few tens of kJ/mol above their respective global energy minima.¹⁶ In other words, molecules of nontrivial size cannot be described in terms of a single three-dimensional structure, but, rather, as ensembles of low-lying conformers with time dependent local fluctuations or, in extreme cases, even global deformations. Hence, an incomplete ensemble of conformers can easily

generate an unsatisfactory modeling of physical-chemical properties, essentially equivalent to a wrong equilibrium structure in the case of semi-rigid molecules.

On these grounds, several semi-automated packages such as CONFAB,¹⁷ CONFLEX,¹⁸ and CREST¹⁹ have been developed to explore Potential Energy Surfaces (PES) using different approaches. General criteria for the classification of search methods can be based either on the computational model used to accept or reject new conformations or on the nature (deterministic or stochastic) of the algorithm employed to generate new tentative structures. The former criterion may refer to different flavors of energy and, possibly, gradient evaluations [by Quantum Chemistry (QC) or Molecular Mechanics (MM) methods], or to simple geometrical criteria.²⁰ Systematic search methods (based on chemoinformatics) are computationally cheap, but of limited applicability, so that stochastic search methods are the most used. In the limit of very long simulations, stochastic methods such as Monte Carlo (MC) and deterministic ones, like molecular dynamics (MD), should be able to sample effectively the conformational space of a (macro)molecule in condensed phase, but the efficiency of the latter approaches is limited for isolated molecules or low-pressure gas-phases.²¹ The most widely employed methods for conformational searches include, besides MC, several types of metaheuristic algorithms (e.g., genetic algorithms,^{22–25} artificial bee colony,^{26,27} differential evolution,²⁸ particle swarm optimization,²⁹ and ant colony³⁰). Metaheuristic, nature-inspired algorithms are able to produce solutions beyond those that are normally generated in a quest for local optimization,³¹ learning from past moves to improve candidate solutions by means of suitable trade-offs between randomization and local search (exploitation vs exploration). Each method has its own strengths and limits, but, since none of them is the best for all situations, selection of the most suitable approach for a specific problem requires careful testing and validation.³²

In addition to a search strategy, a general conformational exploration approach requires an effective yet reliable method to evaluate energies (or other physical chemical properties). For some applications (e.g., virtual screening and pharmacophore modeling) where well-tested specialized force fields (FF) are available (e.g., AMBER³³

or Optimized Potentials for Liquid Simulations³⁴) and only energetic properties are of interest, MM is often the method of choice. However, robust tools aiming at the prediction of spectroscopic properties^{35,36} require either robust and general FFs under active development³⁷ or fast quantum chemical methods. In a previous study,³⁸ we have followed the second alternative, showing that last generation semi-empirical (SE) methods (in particular DFTBA,³⁹ PM7,⁴⁰ and "HF-mini"⁴¹) yield sufficiently accurate geometries to be employed in a two-step procedure, in which final energies are evaluated at a higher level of theory.¹⁹ The present contribution builds on these premises with two main purposes: (i) to further assess the performance of SE approaches, adding the recently proposed GFN-xTB model⁴² to the set of tested methods and (ii) to improve the effectiveness of the exploration step by switching from the Monte Carlo sampling to an evolutionary algorithm.⁴³ As showcase systems, we selected organic and coordination compounds both in the gas phase⁴⁴ and in solution. For the method assessment, we selected some neutral amino acids in the gas phase because (i) they possess several low lying energy minima, (ii) they are important systems in investigations broadly related to the origin of the life, (iii) a large number of state-of-the-art quantum chemical results are available, providing ideal validation and benchmarking data sets, and (iv) they are challenging systems for high resolution spectroscopic techniques. In particular, we have studied threonine,⁴⁵ serine,⁴⁶ and cysteine⁴⁷ (see Fig. 1). Next, we switched to systems in condensed phases by analyzing bulk solvent effects on the conformational landscapes of the anionic, zwitterionic, and cationic forms of threonine in the framework of an implicit solvent approach. Finally, we have tackled a challenging spectroscopic problem, namely the vibrational circular dichroism (VCD) spectra of two related chiral rhodium complexes (see Fig. 2) involving quite flexible ligands⁴⁸ in a non-innocent solvent like acetonitrile, which can be involved in some specific solute-solvent interaction.

The paper is organized as follows: we start by illustrating the selection and validation of suitable EA methods, including the essential details of the management of chemical topology. Then, we introduce a new graphical user interface (GUI) for driving the PES exploration and provide the computational details of quantum

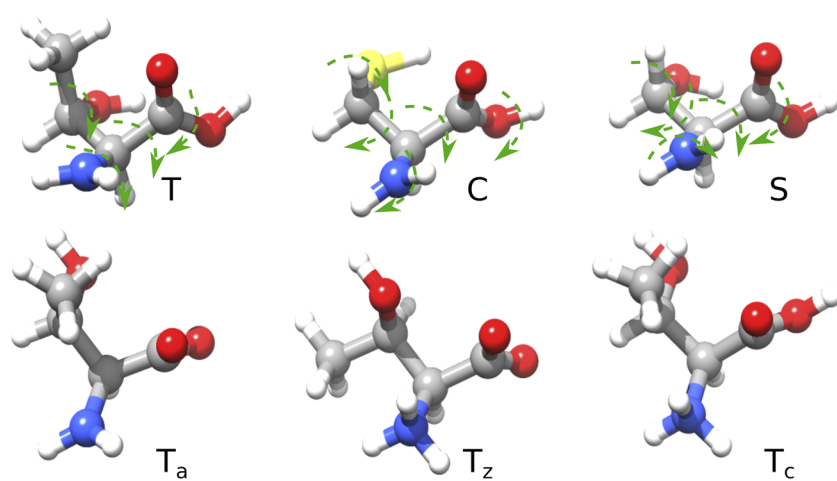


FIG. 1. Neutral and charged amino acids used as test systems in the present study. Upper row: neutral forms of threonine (T), cysteine (C), and serine (S) with rotatable bonds highlighted. Bottom row: anionic (T_a), zwitterionic (T_z), and cationic (T_c) forms of threonine.

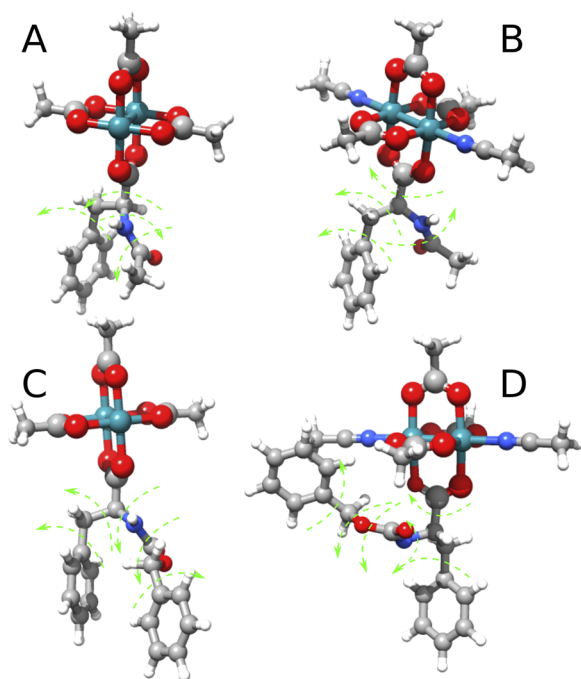


FIG. 2. In the top panels, the $[\text{Rh}_2(\text{O-Phe-Ac})(\text{O-Ac})_3]$ (Rh_2Ac) and the coordinatively saturated complex ($\text{Rh}_2\text{Ac-MeCN}$) are sketched, whereas in the bottom ones the O-He-Cbz complexes are represented (Rh_2Z and $\text{Rh}_2\text{Z-MeCN}$, respectively). Larger spheres indicate atoms blocked during the quantum chemical geometry optimization.

chemical calculations. In the results section, after pointing out the importance of topology and chirality checks in the pipeline, we first assess the performance of GFN-XTB with respect to DFTBA and PM7 employing a Monte Carlo sampling, and the performance of direct Density Functional Theory (DFT) energy evaluations on SE geometries with respect to the most reliable (but more costly) geometry re-optimization at the DFT level in retrieving the complete set of 56 minima for neutral threonine.⁴⁵ Then, the most promising SE method has been employed to compare the performance of Monte Carlo sampling with respect to different evolutionary algorithms (see Sec. III C). The best combination of the search algorithm and QC method is next tested for serine and cysteine in the gas-phase, different protomeric forms of threonine in aqueous solution, and the VCD spectra in acetonitrile solution of two flexible rhodium complexes with and without axial ligands. Additional technical details are given in the [supplementary material](#).

II. METHODS

A. Search methods

1. The $(\lambda + \mu)$ algorithm

Evolutionary algorithms have been previously applied to several problem involving soft degrees of freedom, such as drug candidates²⁴ and metal⁴⁹ or metallorganic complexes.⁵⁰ After their first

introduction in 1970,⁵¹ several variants and modifications of genetic algorithms have been proposed, most of which have been critically reviewed by Whitley.⁵² A short list of the basic GA terminology used in the following is given in the [supplementary material](#).

At variance with other search methods,⁵³ evolutionary algorithms do not rely on a simple physical model, thus typically requiring several parameters, like the specific strategy and the probability threshold for applying a specific operator. A comprehensive study of the application of canonical genetic algorithms to searches in the space of dihedral angles²⁴ concluded that, when applied to small or medium sized molecules, these methods behave like *hill climbers* rather than *hyperspace samplers*⁵² and that the mutation rate is the most critical parameter with a sort of sweet spot for small population sizes around values of 0.5. In the canonical genetic algorithms, the entire population is replaced by new chromosomes at each generation; however, this choice can remove potentially important, albeit sub-optimal, genomes, especially when they are considerably different from the best ones. When aiming, as in our case, to a nearly complete PES exploration, a slightly different algorithm, the $(\lambda + \mu)$ Evolution Strategy is more suitable.^{52,54} In the $(\lambda + \mu)$ algorithm, at each generation, μ parents generate λ offspring, then survival occurs and the population size is reduced back to μ . In our implementation, we always employ a unitary λ/μ ratio and $\lambda = s \cdot P$, where s is the selection rate and P is the population size. For analogous reasons, the selection operator was determined by the so called rank selection or elitism method for the last 10% of the planned iterations, whereas tournament was used at the start. [Figure 3](#) shows a flow chart of the complete procedure. To facilitate the submission of searches (for a future release of the software), we have create a Tk based GUI, whose main windows are shown in [Fig. 4](#). Full details about the evolutionary algorithm and GUI implementation are given in the [supplementary material](#).

2. Structure manipulation and chirality detection

The searches for all amino acids were performed in the space of the soft dihedral angles (see [Fig. 1](#)) discretizing the $[-\pi, \pi]$ domain with a resolution of 30° . Each time a new dihedral value was generated, the selected bin was smoothed by a gaussian with a half-height width of 5° . Crossover operations on dihedral angles were carried out directly by interpolating values and then checking for periodicity. Searches for rhodium complexes were carried out either using the same procedure for a subset of dihedral angles (see [Fig. 2](#)), or employing cartesian coordinates. In the latter case, crossover operations were carried out using the Simulated Binary Crossover, as proposed by Llanio-Truillo *et al.*⁵⁰ Mutations were carried out by “rattling” the atomic positions of non-blocked atoms (see [Fig. 2](#) and Sec. II C), i.e., by changing the atomic positions by superimposing three dimensional gaussians with widths of 0.15 \AA and then picking new values. In all cases, each time a new conformation was generated, we checked the new topology as recently proposed by Ferro-Costas and Fernández-Ramos,⁵⁵ but using the Proxima library.⁵⁶ In particular, we checked that covalent bonds and R/S chiral atoms did not change either because of coordinate manipulation by the evolutionary algorithm or after the geometry optimization by a quantum chemical method. As for covalent bonds, the adjacency matrix of the bond graph⁵⁶ was compared with the starting one. Conversion between cartesian and internal coordinates was done

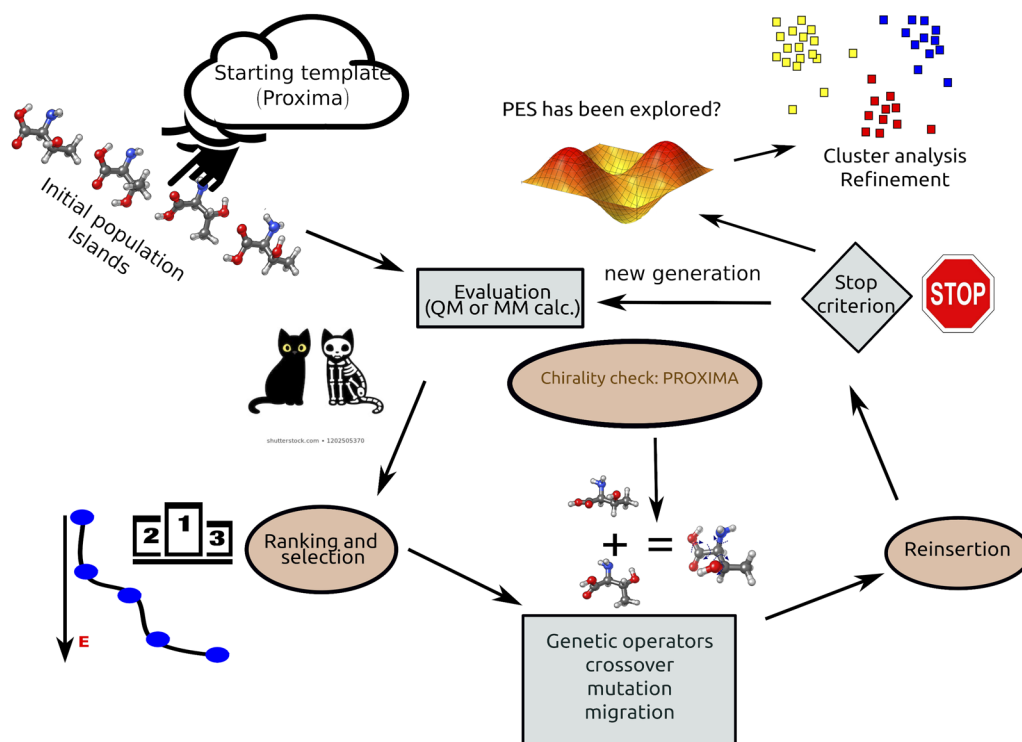


FIG. 3. Flowchart of the search procedure.

using the algorithm devised by Parsons *et al.*⁵⁷ Chiral sp^3 R/S centers were checked with a novel procedure implemented in the Proxima library and based on standard priority rules. Accidental clashes were always avoided by checking that the minimum distance between non-bonded atoms was always larger than 0.7 Å.

3. The island model (IM)

As a further alternative, we extended the single population ($\lambda + \mu$) method to a multi-population version using the so called island model.⁵⁸ In this method, after the generation of the initial population, chromosome sub-populations (called *islands*) are defined by dividing the obtained fitness values in a histogram and then assigning structures from each bin to different islands enforcing comparable average fitness. During the search, islands evolve independently, i.e., genetic operators of selection and crossover work only within islands, which are arranged in a ring topology and are allowed to exchange chromosomes (*migration*) at fixed generation intervals (*migration frequency*). Following Withely *et al.*,⁵⁸ we used a round robin mechanism: at the first migration, an island sends a fixed amount (here 5%) of its best chromosomes to its right neighbor and swaps an equal amount of its worst chromosomes with the best ones sent by the left-lying island. During the next migration cycle, the operation is repeated using second nearest neighbors and so on. The main advantages of the island model in the present context are: (i) it should maintain a large degree of diversity, being thus able to explore wider regions of the PES before convergence; (ii) on large systems with separate weakly interacting moieties, islands can

be faster in finding sub-optimal genomes; and (iii) it allows an effective implementation of multi-node parallel searches for electronic structure codes offering only shared memory parallelism.

A last point deserves attention. All metaheuristic methods are able to find only sub-optimal solutions and, despite all efforts, they can remain trapped in local minima. In the present context, this is a serious drawback as the search will either fail to find all the relevant structures or waste a large number of costly quantum chemical calculations. Aiming to a wide coverage of PESs, we have implemented a linear-search restart²⁴ to deal with premature convergence when a search failed to retrieve all the structures in a given reference dataset or if the best fitness is not improved when elitism was applied.

4. Analysis of obtained structures

Comparison against a reference dataset composed of m structures was based on the full (weighted) Root Mean Square Distance (RMSD) matrix between different structures (excluding non-polar H atoms) and checking that all structures in the reference set had at least one neighbor within a given threshold. For rhodium complexes, we combined a clustering procedure and energy considerations to analyze the results of the search in analogy to what was done in our previous paper and as recommended in the CREST method.¹⁹ In particular, we discarded all the structures outside a given energy threshold above the global energy minimum (see the Results section) and then tried to extract a subset of the sampled structures, retaining the most relevant features by means of a clustering

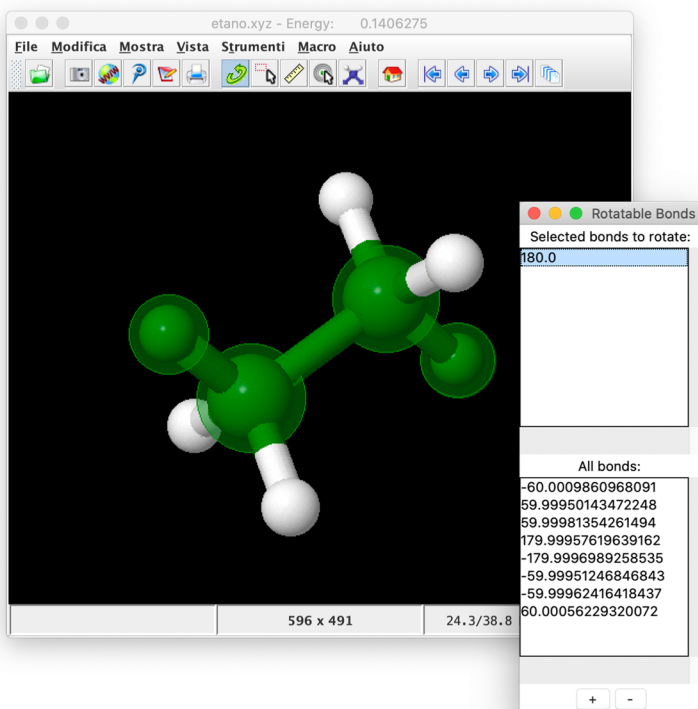
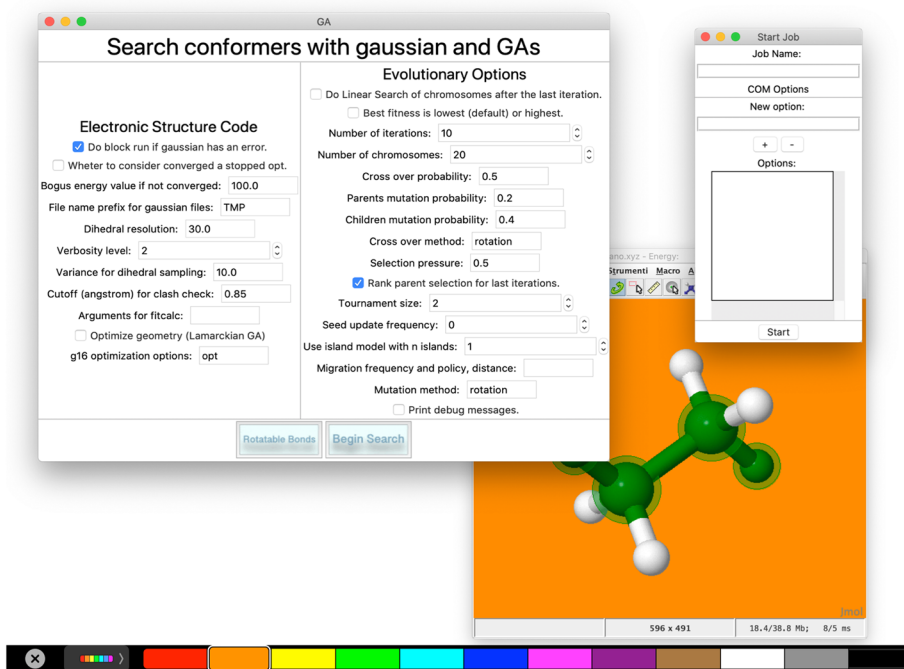


FIG. 4. The main panels of the GUI. (a) The main configuration panel on the left with the COM generator panel on the right and the Jmol window in the background. The options are divided in electronic structure code options and evolutionary algorithm options. The background color of the Jmol window can be selected from the TouchBar portion (at the bottom of the image). (b) In this panel, the user can select the rotatable bonds to be used in the computation. Each time a torsion is selected, the atoms involved are highlighted in the Jmol interface.



performed with the Partition Around Medoids (PAM) algorithm⁵⁹ and using cluster centroids as representative structures. The PAM algorithm optimizes the partition of objects in the dataset into k (an input parameter) distinct groups; the determination of the best value

of k is usually performed by running the algorithm over a range of values and then applying some external validation criteria able to measure cluster compactness and separation.^{60,61} We used three criteria: the change in the slope of the Within Sum of Squares error

(WSS) curve (the so-called *elbow*), together with the Silhouette (SI) and Davies Bouldin (DBI) indexes. Noted is that the best partition is obtained by maximizing SI and minimizing DBI. The choice of PAM was made for different reasons: (i) it is straightforward to validate a unique parameter and (ii) in the present case, we are seeking a relatively high number of clusters, hence it is less problematic to assume a local spherical domain. We have already used PAM⁶² with success when clustering structures for spectroscopic or biological applications.^{63,64} Clustering methods are applied over a *feature space* upon which a distance definition is employed in order to evaluate the similarity/dissimilarity of data points. We built the feature space as follows: for all the structures selected after the application of the energy threshold, we calculated the RMSD matrix of the non-frozen atoms (scaled in the [0, 1] interval) and the inertia tensors (scaled in the [0, 1] interval and then divided by three), and then used RMSD and the differences of inertia tensor elements as features to calculate the L_1 dissimilarity.

B. Case studies

The first set of showcase systems includes threonine, serine, and cysteine in their neutral forms. As mentioned above, in these cases, conformational searches were carried out entirely in internal coordinates using the rotatable bonds highlighted in Fig. 1. The low-lying energy minima of these three systems in the gas-phase have been previously studied at several QC levels. In particular, Szidarovszky and co-workers systematically explored the conformational space of threonine by performing 7776 energy evaluations at the B3LYP/6-311++G** level;⁴⁵ next, seven of these conformers were identified using microwave spectroscopy⁶⁵ and refined at the MP2/6-311++G(d,p) level. It is also noteworthy that the full space of dihedral angles with a 30° resolution would include 2×10^6 points in this case (threonine having six rotatable bonds). A comparable effort was made by He and Allen,⁴⁶ who carried out an exploration of the conformational space of serine by performing 15 552 HF/6-31G* geometry optimizations that resulted in 89 unique minima, reduced to 85 by refinement at the MP2/cc-pVTZ level. Wilke and co-workers⁴⁷ performed an analogous exploration for cysteine by running 11 664 calculations at the HF/3-21G level, which yielded 71 structures after refinement at the MP2/cc-pVTZ level. Finally, we performed a systematic exploration of the conformational space of two rhodium complexes originally investigated in Ref. 48 as a prerequisite for computing their VCD spectra.

C. QC calculations

Monte Carlo calculations were carried out performing geometry optimizations with DFTBA,³⁹ PM7,⁴⁰ or GFN-xTB⁴² semi-empirical methods. The searches for amino acids by evolutionary algorithms were carried out at either the DFTBA or PM7 level, followed by B3LYP/6-31G+(d) single point energy calculations⁶⁶ with the addition of empirical dispersion⁶⁷ using the Gaussian suite of programs.⁶⁸

Soft degrees of freedom used in the preliminary searches were kept frozen in the final QC geometry optimizations. Selected structures of charged threonine underwent a second stage optimization using the double hybrid B2PLYP⁶⁹ functional including empirical dispersion (D3BJ)⁷⁰ in conjunction with the jun-cc-pVTZ^{71,72} basis

set (hereafter B2), which has proven to provide reliable results in the characterization of conformers in solution⁷³ at a not too-high computational cost. Starting structures were built with the help of a molecular editor and then relaxed with the UFF force field.⁷⁴ DFT calculations on rhodium bimetallic complexes were carried out at the B3LYP level including empirical dispersion (D3BJ)⁷⁰ in conjunction with the Jul-cc-pVDZ^{71,72} basis set on light atoms and the Stuttgart (SDD) valence basis set⁷⁵ with the corresponding pseudopotential⁷⁶ on Rh atoms (hereafter B3). The initial structures were built from the crystallographic structure of rhodium(II) acetate⁷⁷ and then optimized at the B3 level of theory. In this case, the lack of parameters ruled out DFTBA and the computational cost forced us to eliminate the DFT single point energy evaluation. Note that the presence of two heavy metal atoms results in a huge computational cost, and thus, optimizations were performed freezing the inner core of the complex (see larger atoms in Fig. 2). Further optimizations and spectroscopic calculations were performed at the B3 level of theory on the structures obtained from the clustering procedure.

III. RESULTS AND DISCUSSION

A. Chirality detector

We performed a number of *dry runs* (in which structures were generated using random numbers in place of energy evaluations in order to assign fitness) in order to set the minimum distance threshold and test the ability of the software to preserve topology and chirality. Figure 5 shows two structures obtained in a dry run without R/S check, in which the chiral carbon atom in the first dirhodium complex has changed its chirality; it is also worth noting that, with the distance checks in place, the distance between the hydrogen atoms approaches but never goes below the cutoff of 0.7 Å. In addition, we used dry runs to compare the ratio of collisions generated by internal or cartesian moves, by performing four runs (two in cartesian and two in internal coordinates) for each complex. Since 49.5 % of attempted moves was refused, we abandoned the use of cartesian moves for all the production runs.

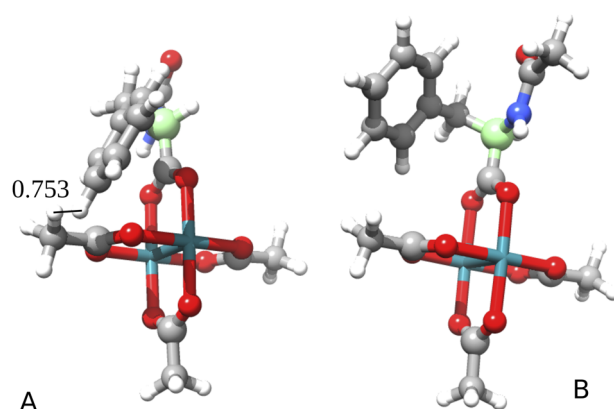


FIG. 5. Structures with different chirality in dirhodium complexes obtained in dry runs; the chiral carbon atom is shown in green; the distance between two close hydrogen atoms is also shown.

TABLE I. Summary statistics for stochastic samplings. The first (Q1) and third (Q3) quartiles of the relative energies along with the maximum (ΔE), the number of retrieved structures, and the RMSE are reported. Energetic values in kJ/mol.

Method	Acc. ratio	Q1	Q3	ΔE_{max}	No. of found	RMSE
DFTBA	9.22	13.12	20.44	41.11	39	10.8
PM7	5.53	9.05	17.69	44.60	26	10.6
XTB	7.16	9.86	17.90	79.90	27	9.1

B. Performance of semiempirical methods

As a first step, the performances of the GFN-xTB⁴² semiempirical method were compared with those of PM7⁴⁰ and DFTBA³⁹ by adopting the same protocol used in our previous study,³⁸ i.e., a Monte Carlo search with geometry optimization and restart and one of the same case studies, namely threonine. More in detail, we carried out conformational explorations of threonine, with PM7, DFTBA, and GFN-xTB geometries and energies and compared the effectiveness of each method in retrieving as many structures as possible from the reference dataset⁴⁵ using as a benchmark the two-level approach of our previous study. For each method, three independent conformer searches were performed using the stochastic method setting the number of iterations, target temperature, and grid resolution to 5000 K, 400 K, and 30°, respectively. During the search, trial geometries were generated changing three out of five dihedral angles at random and assigning to the current geometry the highest energy located so far when the search remained stuck for more than 2% of the planned steps. After the search, all generated structures were post filtered using energetic (structures below the third energy quartile) and geometric (heavy atom RMSD larger than 0.5 Å) descriptors; these reduced datasets underwent further DFT optimizations. Table I summarizes the results of the searches

carried out using the Monte Carlo method. It can be observed that DFTBA yields an higher acceptance ratio and a wider energy spread as compared to the other methods. Panel (a) in Fig. 6 shows that comparable energy distributions of different structures are obtained by the three semi-empirical methods (RMSD distribution is also shown in Fig. S1 of the supplementary material). Unique energy minima, thus identified, are shown in a 2D plane defined by their relative energy vs dipole moment [panel (b)]. The plot confirms that all the employed models successfully identified the conformers lying less than 12 kJ/mol above the global energy minimum, including the experimentally detected conformers. Beyond this threshold, the identified minima differ among the various models, none of which was able to retrieve all the 56 conformers of the reference dataset. However, this may be related to the non-exhaustiveness of the searches; also the accuracy in distinguishing equivalent structures is likely to be more important in more demanding applications. To address this point, we compared the root mean square error (RMSE) of single point energies obtained with each of the low-level models on a set of 150 GFN-xTB geometries against the DFT-D3 ones and indeed the GFN-xTB energies showed the smallest error with respect to the DFT-D3 values. However, from a practical point of view, it is worth observing that all the methods are able to identify the seven structures of gas phase neutral threonine that have been observed experimentally.^{38,65}

C. Threonine

1. Gas phase: ($\lambda + \mu$) vs Monte Carlo

Initial searches were performed using a population of 28 chromosomes either with a single population or using the island model (see Table II, low end values). These runs produced 1394 and 532 structures, respectively, before stalling or reaching of the programmed maximum number of generations. Comparison of the sampled structures against the reference dataset⁴⁵ using RMSD (cutoff of

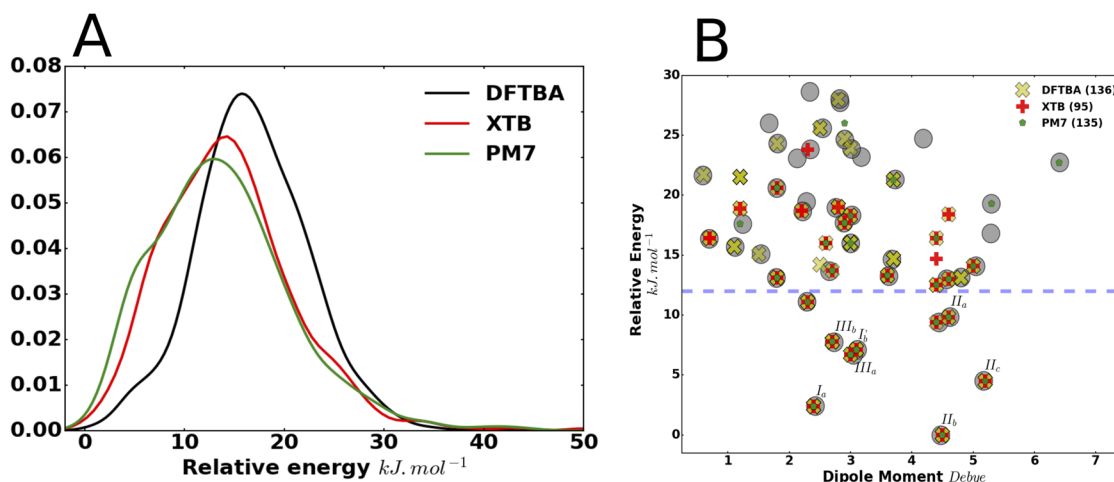
**FIG. 6.** Results of Monte Carlo searches for threonine using DFTBA, PM7, and XTB. Panel (a) distribution of relative energies with respect to the GEM for the three SE methods tested. Panel (b) relative energy and dipole moment specific structures present in the reference dataset and retrieved by the three methods; the dashed line is the 12 kJ/mol threshold used to select structures to be refined.

TABLE II. Run parameters and values for L-threonine searches.

Parameter	Single population	Island model
Population size	28–100	28–100
Number of generations (max)	50	50
Selection rate	0.5	0.5
Selection method	Tournament	Tournament
Tournament size	2	2
Elitism (last 10% generations)	T	T
Crossover method	SBX	SBX
Crossover probability	0.5	0.5
Mutation rate (parents)	0.3	0.3
Mutation rate (children)	0.5	0.5
Number of islands	1	4
Migration frequency	NA	4
Migration size (IM)	NA	0.05

0.2 Å) shows that both evolutionary algorithms missed 8 structures out of 56, but with significantly different convergence rates: while the single population runs had a similar behavior, slowly improving until the last few generations, the run performed with the island model was able to converge in just 25 generations or about 550 QC calculations. Anyway, even the worst outcome represents a significant improvement with respect to stochastic methods like Monte Carlo (less than 1400 calculations vs 3000). Panel (a) in Fig. 7 shows a comparison of the completeness achieved in the best runs by different search algorithms in conjunction with DFT geometry/energy evaluation.

Since in all these searches some structures below a threshold of 15 kJ/mol were still missing, we tried to force the evolutionary

algorithm to explore less accessible regions of the PES by enlarging the population size to 100, while leaving other parameters unchanged. On the basis of the Monte Carlo results, we carried out three pairs of runs employing 27 or 100 chromosomes, respectively: one pair with PM7 using one population and two pairs with the island model using either DFTBA or PM7. For the single population run, the enlargement of the size slightly increased the number of retrieved structures (6 and 4) at a considerably higher cost (2250 and 2600 calculations, respectively). The island model led to a further improvement: DFTBA retrieved 53 out of 56 structures in both replica, stalling at 1600 and 2250 calculations, whereas PM7 missed just one structure (number 21 at 19.45 kJ/mol above the global energy minimum) stalling at 2800 and 3200 calculations. The overall performance can be appreciated looking at panel (b) of Fig. 7 where the number of missing structures (average of replicas) vs the number of calculations carried out is shown (roughly 60 points for each generation); it is apparent that both the runs performed with the island model converge faster than the single population one and that PM7 offers an additional (small) advantage. Taken together, the results show that less than 10 generations should be sufficient to sample a system with these numbers of degrees of freedom with an accuracy largely sufficient to obtain converged estimates of most observables.

2. Threonine in aqueous solution

After the extensive exploration of threonine in the gas phase, we proceeded to characterize its charged forms in solution, employing the conductor-like polarizable continuum model (CPCM)⁷⁸ to take into account bulk solvent effects. In this case, run-time topology checks are critical since proton transfers may take place during the search. The computations were performed using the settings shown in Table II, carrying out a single search using PM7 in view

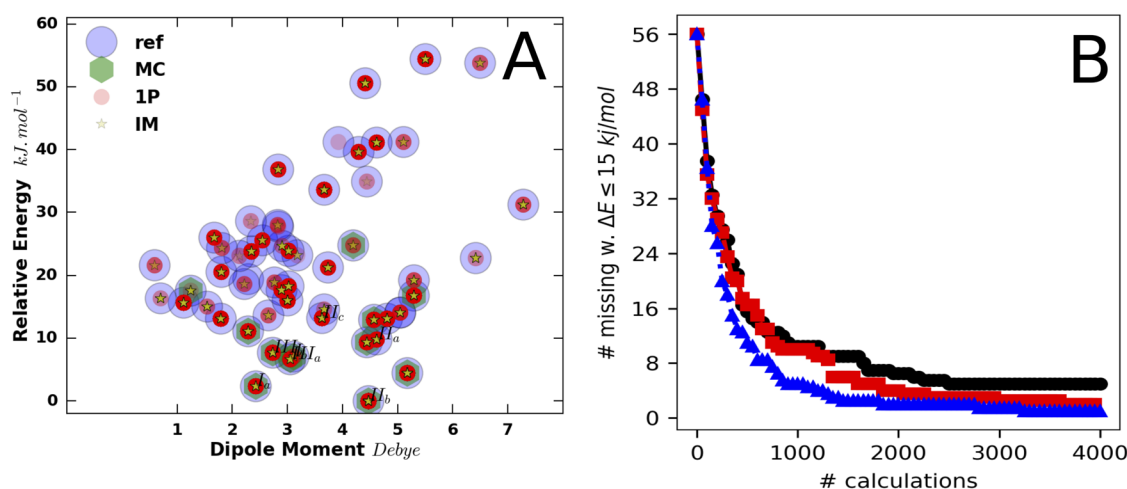


FIG. 7. Panel (a) structures retrieved in searches for L-threonine using a population of 28 chromosomes with a RMSD threshold of 0.2 Å. Reference structures⁴⁵ are shown as cyan circles, neighbors found by MC as green hexagons, neighbors found by single population EA with a smaller red circle, and neighbors found with the IM with a yellow star. Panel (b) the number of missing structures as a function of calculations done using 100 chromosomes for a single population (PM7, black circles), IM/DFTBA (red squares), and IM/PM7 (blue triangles).

TABLE III. ΔG at B2 level of theory of the unique structure obtained from the optimization of the selected structure in the MC searches.

Anion		Zwitterion		Cation	
Conf.	ΔG	Conf.	ΔG	Conf.	ΔG
T-A-1	0.0	T-Z-1	0.0	T-C-1	0.0
T-A-2	0.4	T-Z-2	1.6	T-C-2	7.2
T-A-3	1.5	T-Z-3	4.2	T-C-3	19.0
T-A-4	11.9	T-Z-4	20.0		
		T-Z-5	20.1		

of the reduced conformational freedom of charged species in solution. Structures within 25 kJ/mol from the global energy minimum were then re-optimized at higher level of theory in order to obtain a better estimate of relative stabilities (see Sec. II C); no clustering was carried out since the dataset was already small enough. Relative stabilities of the low lying conformers are reported in Table III.

As well known, in aqueous solutions at neutral pH, the zwitterionic form of amino acids is more stable than its neutral counterpart. PH changes then lead to either protonation of the carboxylate group or deprotonation of the NH_3 moiety. Thus, an extensive exploration of the charged forms of threonine is pivotal to analyze the relationships among the low-lying conformers of the different charged species and to identify the preferred paths for protonation or deprotonation. In Fig. 8, the geometries and relative free energies of the low-energy conformers (within 12 kJ/mol above the global energy minimum of each form) for anionic, zwitterionic, and cationic forms of threonine are reported, with orange lines connecting closely related structures. The most stable conformer of the cationic form (the blue energy level in Fig. 8) is characterized by hydrogen bonds of the positively charged NH_3 group with both the

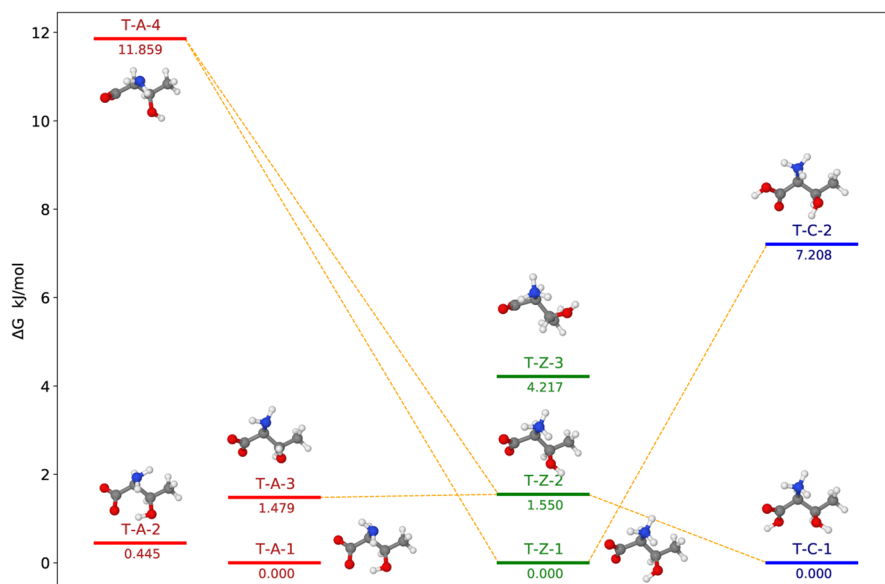
carboxylic and hydroxylic oxygens. This structure is closely related to the second low-energy conformer of the zwitterionic form, which is only 1.5 kJ/mol above the global energy minimum of this form. Interconversion between the two conformers is ruled by rotation of the hydroxyl hydrogen atom. Only slight structural rearrangements occur during the deprotonation of the carboxylic group. In the case of the anionic form, after deprotonation of the ammonium group, the strongest hydrogen bond is formed between one carboxylic oxygen and the hydrogen of the hydroxyl group (rather than with an aminic hydrogen). The interaction between the NH_2 and OH groups is retained in the less stable T-A-3 and T-A-4 conformers, which represent the possible connections between the zwitterionic and the anionic forms upon NH_3 deprotonation.

Though beyond the scope of the present work, we recall that a complete knowledge of the most stable conformers of each species and of the possible paths of proton dissociation gives access to the estimation of the acid dissociation constants. Considering the same experimental conditions, the pKa values can be obtained evaluating the $\Delta\Delta G$ between the conformers of the amino acid under study and the conformers of a related amino acid whose experimental values are known.

3. Serine and cysteine

Having evaluated the performance of the $(\lambda + \mu)$ IM strategy, we proceeded to test it further with the next two systems, i.e., serine and cysteine.

For serine (Fig. 9), we run two replicated runs using PM7 (with the same settings employed for cysteine) and another pair using DFTBA, with the single population model and 100 chromosomes. In all cases, the populations stopped to improve between 14 and 35 generations (i.e., between 1150 and 2000 SE/DFT calculations), being unable to find a single structure among the 85 present in the dataset, even after a linear search restart. This missing conformer was actually the highest-energy structure, lying 214.7 kJ/mol above

**FIG. 8.** Energy levels below 12 kJ/mol of the anionic (red), zwitterionic (green), and cationic (blue) forms of L-threonine are reported. Orange lines connect structures upon acid dissociation. The insets show 3D representations of the conformers.

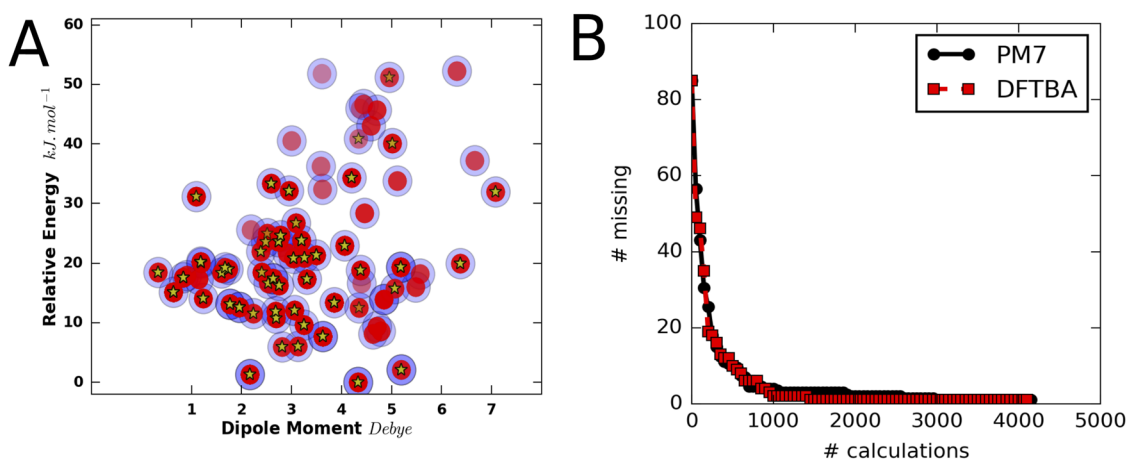


FIG. 9. Panel (a): structures retrieved in searches for serine with RMSD thresholds of 0.2 and 0.125 Å, respectively. Reference structures⁴⁷ are shown as cyan circles, neighbors within 0.2 Å with a smaller red circle, and neighbors within 0.125 Å with a yellow star. Panel (b): the number of missing structures as a function of calculations for PM7 (black circles) and DFTBA (red squares).

the global energy minimum. Superimposition between the latter structure and its nearest neighbor (see Fig. S2 in the [supplementary material](#)) yielded a RMSD of 0.0128 Å. Lowering the RMSD threshold made the number of missing structures jump to 25 in both PM7 runs with an average energy difference of these missing conformers with respect to the global energy minimum of 56.3 kJ/mol and 52.0 kJ/mol, respectively. Note that, at variance with cysteine, even with a low threshold, all but two structures within 15 kJ/mol were still detected. The number of missing structures at the PM7 and DFTBA levels [panel (b) of Fig. 9] and the energies of the missing structures

(see Fig. S3 in the [supplementary material](#)) show the quite good convergence of both methods at about 8–9 generations.

For cysteine, we run two replicated PES explorations with the settings used for the second set of threonine runs (see Table II, third column) using only PM7 (owing to the lack of sulfur parameters in DFTBA) and single point DFT evaluations. The results are shown in Fig. 10: looking at panel A, it can be observed that, using the RMSD between heavy atoms and polar hydrogens and a threshold of 0.2 Å, the single population model was able to retrieve all the structures in the reference dataset (within 36 generations, i.e., optimizations)

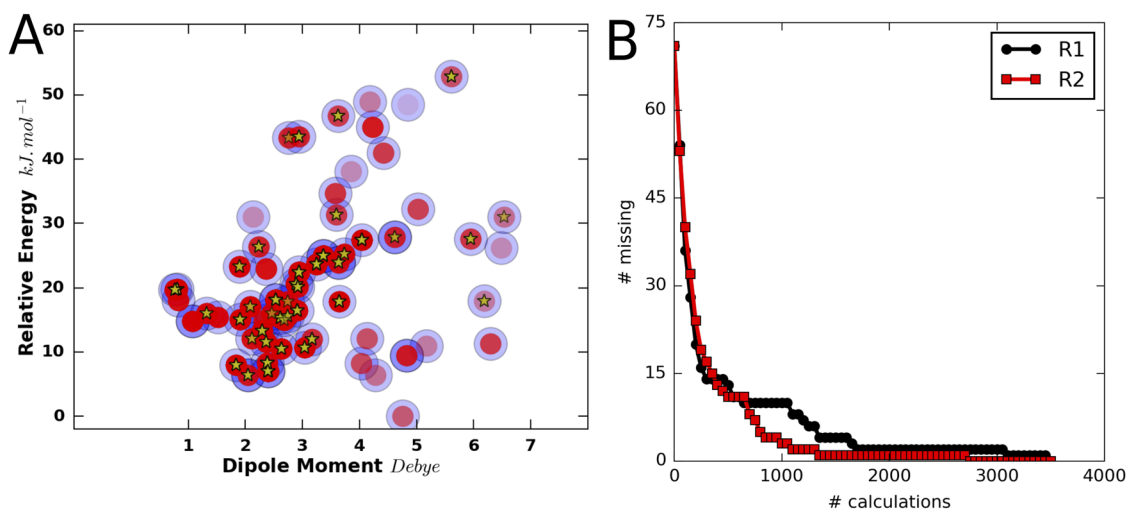


FIG. 10. Panel (a): structures retrieved in searches for cysteine with a RMSD thresholds of 0.2 Å and 0.125 Å, respectively. Reference structures⁴⁷ are shown as cyan circles, neighbors within 0.2 Å with a smaller red circle, and neighbors within 0.125 Å with a yellow star. Panel (b): the number of missing structures as a function of calculations for the two replicas.

in one replica and all but one in the other replica (all at least 44.3 kJ/mol above the global energy minimum [using geometries from the reference dataset optimized at the B3LYP-D3/6-31+G(d) level]). Lowering the RMSD threshold to 0.125 Å increased the number of non-retrieved structures to 22 and 23, respectively; with a single exception (see below), all these structures were above the 15 kJ/mol threshold (on average 71 kJ/mol). Comparison of SE/DFT and pure DFT searches did not highlight any significant difference. Superimposition between the global energy minimum in the reference dataset and its nearest neighbor (Fig. S4 in the [supplementary material](#)) yielded a RMSD of 0.155 Å. It is worth observing that use of the $(\lambda + \mu)$ evolutionary algorithm allows us to keep high fitness chromosomes during the whole exploration, thus seeming better suited for our needs: even if this feature may slow down the convergence, finding the lowest lying structures was never an issue in our tests. Finally, the speed of exploration is in line with those already observed for the other two amino acids.

D. Rhodium complexes

The last set of case studies includes the four rhodium coordination complexes described in Ref. 48. The remarkable robustness of the exploration of amino acid PES's convinced us to retain the same parameter set for these systems, see [Table II](#). However, due to the huge computational cost of evaluating even single point DFT energies at each fitness calculation, the search was carried out using PM7 geometries and energies; in addition, we limited the number of generations to 30 and the number of chromosomes to 70. For each system, we carried out a replicated search using either cartesian or internal coordinates. The role of two explicit solvent molecules (CD₃CN) filling the axial positions of the octahedral coordination shell of rhodium atoms was also investigated. Among the structures found during the searches, only those lying less than 25 kJ/mol above

the GEM were retained and grouped by means of clustering, thus finally obtaining a small set of highly descriptive structures.

[Figure 11](#) shows the validation scores obtained for the four dirhodium complexes used to select the best value of k , i.e., the number of structures used in the second step (see [Sec. II](#)). Taking into account that neither of the used validation scores is rigorous (from a general point of view, the same concept of correct clustering is ill defined), we are actually seeking either a consensus among the scores or a strong signal. It is also worth remembering that the internal variance in clusters is inversely proportional to k , so that the *smallest best* k is actually sought. For Rhac (blue lines), WSS and SI were not very informative (there are neither big changes in slope in WSS nor peaks in SI), so that the value of k obtained by DBI (18) was selected (the structures collapsed to 15 after full DFT optimization); for Rhac-MeCN (orange lines), SI clearly indicates 14 clusters and DBI 10; however, the value of DBI at $k = 14$ is small enough to convince us to retain this value. In analogy with Rhac, we selected $k = 16$ for Rhz (green lines) and $k = 24$ for Rhz-MeCN (red lines) since only DBI gave sufficient information.

In the second stage of the general two-step procedure, the resulting candidates underwent a full DFT geometry optimization (see [Sec. II C](#)) followed by computation of spectroscopic parameters. The clustering procedure ended with 14–24 representative structures for each complex, some of which collapsed to the same minima after B3 geometry optimization, thus further reducing the final set of structures tuning the overall spectra. The reduction was particularly significant for RhAc, where only 5 of the initial 15 structures survived after geometry re-optimization.

In [Fig. 12](#), the experimental VCD spectra of the four complexes are compared to those issuing from Boltzmann averaging of the computed harmonic spectra of all the final structures for each complex. As always, the computed harmonic frequencies in the 1300 cm⁻¹–1800 cm⁻¹ region have been scaled by 0.99 to take

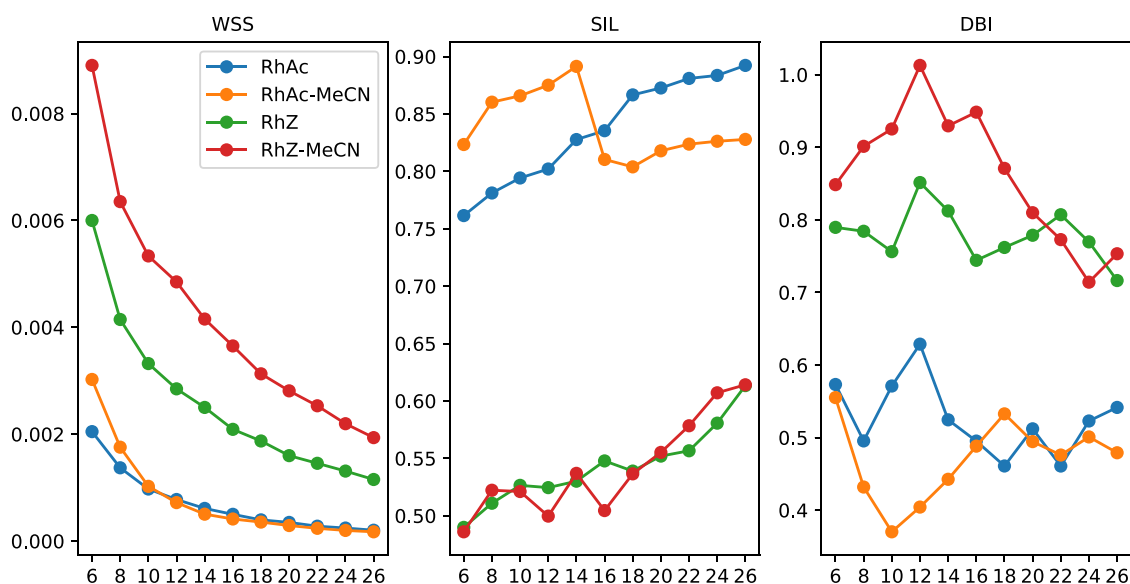


FIG. 11. WSS, SI, and DBI scores for the four dirhodium complex as a function of the number of clusters.

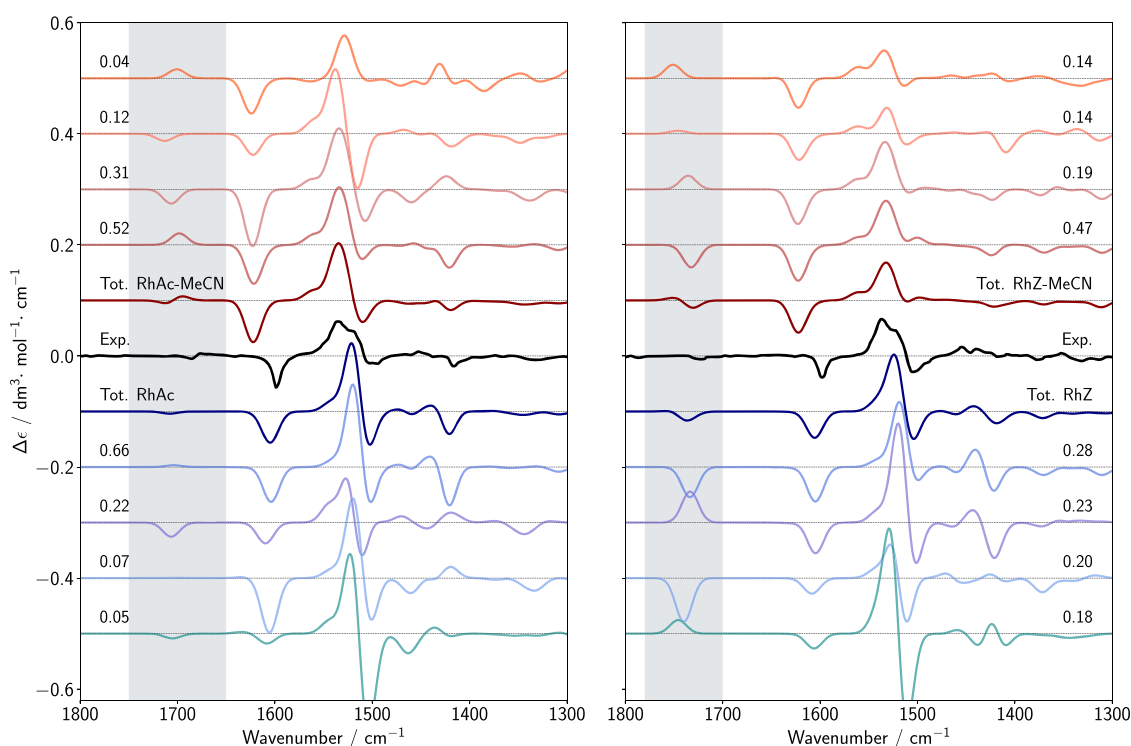


FIG. 12. Experimental (black lines) and theoretical harmonic vibrational circular dichroism spectra of the four lowest energy conformers and Boltzmann averaged spectra at 298 K (bold lines) of Rh_2Ac (left panel) and Rh_2Z (right panel). The spectra of the conformers with explicit acetonitrile molecules are reported in shade of red, whereas the coordinatively unsaturated conformers are reported in shades of blue. The percentage probability of occurrence estimated at 298 K is reported. The theoretical line-shapes have been convoluted by means of gaussian distribution functions with half-widths at half-maximum of 10 cm^{-1} .

into account anharmonicity. The computed spectra show significant differences among the various conformers, thus confirming the paramount role of a comprehensive exploration of the conformational space. For instance, the isolated signal at about 1700 cm^{-1} (the gray areas in Fig. 12) is characteristic of the amide group and its position and intensity depend on both the orientation and the local environment of this moiety within the molecular system. Indeed, small variations of the conformation can even lead to sign changes and none of the conformers, taken alone, reproduces the experimental spectrum. Proper description of the extremely sensitive amide transition requires also the inclusion of explicit acetonitrile molecules in axial positions, since their presence led to small, but non-negligible changes of the conformer populations. In particular, only Boltzmann averaging of the low-energy conformers of the coordinatively saturated complex (RhAc-MeCN) reproduces the small bisignate signal of the experimental high-resolution spectrum of RhAc (the red bold line in the left panel in Fig. 12). Since the only source of chirality is the stereogenic center of the acetate ligands, some general features were expected to be shared among the different complexes. Indeed, the $-, +, -$ sequence of signals of the spectra in the 1450 cm^{-1} – 1600 cm^{-1} region is found in most of the conformers; nevertheless, significant differences in the peak intensity are present, which, only after tuning by the populations of the different conformers, approach the experimental shapes.

IV. CONCLUSIONS AND PERSPECTIVES

The identification and characterization of the most statistically significant (low-energy) minima of flexible medium-size molecular systems is a demanding task due to the high dimensionality of the problem and the ruggedness of the potential energy surfaces. Two general tools are needed to accomplish this task, namely, an effective way to cross barriers in the PES and an equally efficient way of evaluating properties at a given geometry.

In this contribution, we set up a general strategy, which tries to fulfill both goals, combining, in the best possible way, speed, feasibility for large systems, and accuracy in the exploration of multi-dimensional rugged surfaces with many shallow minima. We validated our procedure using as test cases some amino acids (which have their own specific interest in high resolution spectroscopy and astrochemistry) by retrieving almost complete datasets of several minima at a fraction of the computational cost of more conventional approaches. The goal was achieved by tuning and combining some well-known techniques for our needs: in particular, use of the $(\lambda + \mu)$ variant in place of a conventional genetic algorithm allows us to retain in the selected population good performing structures even if their geometrical parameters are not closely related to the best ones. Furthermore, use of large populations for fewer generations, combined with the island model, helps in keeping a degree of

diversity in the population. Finally, the island model is particularly effective for parallelization and distribution among nodes. Concerning next the evaluation of energies, we extended the benchmarks reported in a previous study, confirming the effectiveness of semi-empirical quantum chemical methods, which are able to combine general applicability and good accuracy to reasonable computational cost.¹⁹

One strength of our framework is the integration of the exploration layer of the software with the Proxima library, which allows for checking and manipulating topological properties in an easy and transparent way; in the future, when more features will be added to Proxima, this may allow us to explore different chemical phenomena by accepting or rejecting newly generated geometries based on topological properties (e.g., the presence of a hydrogen bond, just to mention one of the simplest cases). This is actually linked to one of the most important improvements we plan to introduce in future releases, i.e., the optimization of genomes based on general internal coordinates, generated automatically by the software and/or selected by the user, rather than on just torsions or cartesian coordinates. As a matter of fact, torsions (or other internal coordinates, e.g., ring puckering) are much more efficient than cartesian coordinates in exploring large amplitude motions, whereas the latter are simpler and can be used in any problem, but their tuning is intrinsically not easy. In fact, this aspect has been taken into account, at least partially, since the geometry optimization by an electronic structure code is done precisely in that way freezing the genome degrees of freedom. However, we are currently restricted to very basic genomes. At the same time, the sub-optimal efficiency of cartesian based searches can be outflanked in different ways: for instance, the use of metadynamics implemented in CREST seems to be a quite effective approach. More general coordinate manipulation would also allow combining systematic and metaheuristic approaches⁵⁵ by carrying out an initial pruning of candidate structures, possibly on a less dense grid, using only topological properties.

Another aspect that deserves further analysis is the type of local relaxation carried out on candidate structures, which, in our case, is a geometry optimization, while other authors used molecular dynamics. From the perspective of an evolutionary algorithm, these choices are interchangeable, since they can be viewed as an additional mutation operator, which carries away a chromosome from its initial state. In this respect, it would be interesting to combine evolutionary algorithms with local Monte Carlo simulations in condensed phase (in place of molecular dynamics), whose implementation in generalized curvilinear coordinates is quite straightforward. Finally, although our software can use different electronic structure codes, its very effective interface with the Gaussian software allows for combining cartesian and internal coordinates and performing multilayer [QM/QM, (Quantum Mechanical) QM/MM, etc.] computations, thus further enlarging the range of applications of conformer searches.

Three further aspects are, in our opinion, particularly significant for future developments: first, while we have used a general recipe in all the presented case studies, a wider generalization (and perhaps a metaEA optimization) will be perhaps needed when a general coordinate manipulation stack will be implemented; second, the implementation of multi-objective evolutionary algorithms, able to optimize simultaneously energy and other properties, would be very useful in spectroscopic applications; and third, automatic

procedures should not completely overcome domain competences and chemical intuition and in this respect the use of perception tools and immersive virtual reality may be of significant help, as hinted in recent studies.^{63,79}

In conclusion, we think that, even pending those future enhancements, we already dispose *hic et nunc* of a robust and general platform aiding the spectroscopic analysis of large and flexible systems in a user-friendly and effective way.

SUPPLEMENTARY MATERIAL

The [supplementary material](#) files include further information on metaheuristic methods, details on the implementation of the GUI and its functioning, and additional results. In particular, the latter include distribution of energies in MC searches and the superimposition between global energy minima of amino acids found in EA searches and present in reference datasets.

ACKNOWLEDGMENTS

The authors thank the Avogadro Staff at SNS for managing the HPC systems and Dr. Sara del Galdo and Andrea Salvadori for useful discussions.

DATA AVAILABILITY

The data that support the findings of this study are available from the corresponding author upon reasonable request.

REFERENCES

- 1 K. Choo, A. Mezzacapo, and G. Carleo, "Fermionic neural-network states for *ab-initio* electronic structure," *Nat. Commun.* **11**(1), 2368 (2020).
- 2 S. A. Tawfik, O. Isayev, C. Stampfl, J. Shapter, D. A. Winkler, and M. J. Ford, "Efficient prediction of structural and electronic properties of hybrid 2d materials using complementary DFT and machine learning approaches," *Adv. Theory Simul.* **2**(1), 1800128 (2019).
- 3 F. Peiretti and J. M. Brunel, "Artificial intelligence: The future for organic chemistry?," *ACS Omega* **3**(10), 13263–13266 (2018).
- 4 J. P. Reid and M. S. Sigman, "Holistic prediction of enantioselectivity in asymmetric catalysis," *Nature* **571**(7765), 343–348 (2019).
- 5 K. Ghosh, A. Stuke, M. Todorović, P. B. Jørgensen, M. N. Schmidt, A. Vehtari, and P. Rinke, "Deep learning spectroscopy: Neural networks for molecular excitation spectra," *Adv. Sci.* **6**(9), 1801367 (2019).
- 6 J. C. Cancellia and J. S. Torrecilla, C. V. Proestos, and J. O. Valderrama, "Editorial: Artificial intelligence in chemistry," *Front. Chem.* **8**, 275 (2020).
- 7 A. R. Leach, "A survey of methods for searching the conformational space of small and medium-sized molecules," in *Reviews in Computational Chemistry*, edited by K. B. Lipkowitz and D. B. Boyd (John Wiley & Sons, Inc., 1991), Vol. 2, pp. 1–55.
- 8 M. Saunders, K. N. Houk, Y. D. Wu, W. C. Still, M. Lipton, G. Chang, and W. C. Guida, "Conformations of cycloheptadecane. A comparison of methods for conformational searching," *J. Am. Chem. Soc.* **112**(4), 1419–1427 (1990).
- 9 J. T. Ngo and M. Karplus, "Pseudosystematic conformational search. Application to cycloheptadecane," *J. Am. Chem. Soc.* **119**(24), 5657–5667 (1997).
- 10 D. K. Agrafiotis, A. C. Gibbs, F. Zhu, S. Izrailev, and E. Martin, "Conformational sampling of bioactive molecules: A comparative study," *J. Chem. Inf. Model.* **47**(3), 1067–1086 (2007).
- 11 V. Barone and A. Polimeno, "Integrated computational strategies for UV/vis spectra of large molecules in solution," *Chem. Soc. Rev.* **36**(11), 1724 (2007).

- ¹²V. Barone, M. Biczysko, and G. Brancato, "Extending the range of computational spectroscopy by QM/MM approaches: Time-dependent and time-independent routes," in *Advances in Quantum Chemistry* (Elsevier, 2010), Vol. 59, pp. 17–57.
- ¹³M. P. Haag, A. C. Vaucher, M. Bosson, S. Redon, and M. Reiher, "Interactive chemical reactivity exploration," *ChemPhysChem* **15**(15), 3301–3319 (2014).
- ¹⁴G. N. Simm, A. C. Vaucher, and M. Reiher, "Exploration of reaction pathways and chemical transformation networks," *J. Phys. Chem. A* **123**(2), 385–399 (2019).
- ¹⁵D. J. Diller and K. M. Merz, Jr., "Can we separate active from inactive conformations?," *J. Comput.-Aided Mol. Des.* **16**(2), 105–112 (2002).
- ¹⁶J. Kirchmair, C. Laggner, G. Wolber, and T. Langer, "Comparative analysis of protein-bound ligand conformations with respect to catalyst's conformational space subsampling algorithms," *J. Chem. Inf. Model.* **45**(2), 422–430 (2005).
- ¹⁷N. M. O'Boyle, T. Vandermeersch, C. J. Flynn, A. R. Maguire, and G. R. Hutchison, "Confab—Systematic generation of diverse low-energy conformers," *J. Cheminform.* **3**(1), 8 (2011).
- ¹⁸H. Goto, U. Nagashima, K. Ohta, T. Takahashi, and Y. Takata, "Conflex: Conformational behaviors of polypeptides as predicted by a conformational space search," Tech Connect Briefs, <https://briefs.techconnect.org/papers/conflex-conformational-behaviors-of-polypeptides-as-predicted-by-a-conformational-space-search/> (2003).
- ¹⁹P. Pracht, F. Bohle, and S. Grimme, "Automated exploration of the low-energy chemical space with fast quantum chemical methods," *Phys. Chem. Chem. Phys.* **22**(14), 7169–7192 (2020).
- ²⁰S. Riniker and G. A. Landrum, "Landrum. Better informed distance geometry: Using what we know to improve conformation generation," *J. Chem. Inf. Model.* **55**(12), 2562–2574 (2015).
- ²¹S. B. Ozkan and H. Meirovitch, "Conformational search of peptides and proteins: Monte Carlo minimization with an adaptive bias method applied to the heptapeptide deltorphin," *J. Comput. Chem.* **25**(4), 565–572 (2004).
- ²²N. Nair and J. M. Goodman, "Genetic algorithms in conformational analysis," *J. Chem. Inf. Comput. Sci.* **38**(2), 317–320 (1998).
- ²³Y. Sakae, T. Hiroyasu, M. Miki, and Y. Okamoto, "Protein structure predictions by parallel simulated annealing molecular dynamics using genetic crossover," *J. Comput. Chem.* **32**(7), 1353–1360 (2011).
- ²⁴Z. E. Brain and M. A. Addicoat, "Optimization of a genetic algorithm for searching molecular conformer space," *J. Chem. Phys.* **135**(17), 174106 (2011).
- ²⁵J. Zhao, R. Shi, L. Sai, X. Huang, and Y. Su, "Comprehensive genetic algorithm for *ab initio* global optimisation of clusters," *Mol. Simul.* **42**(10), 809–819 (2016).
- ²⁶H. A. A. Bahamish, R. Abdullah, and R. Abdul Salam, *Protein Conformational Search Using Bees Algorithm* (IEEE, 2008). pp. 911–916.
- ²⁷H. A. A. Bahamish, R. Abdullah, and R. Abdul Salam, *Protein Tertiary Structure Prediction Using Artificial Bee Colony Algorithm* (IEEE, 2009). pp. 258–263.
- ²⁸G.-J. Zhang, X.-G. Zhou, X.-F. Yu, X.-H. Hao, and L. Yu, "Enhancing protein conformational space sampling using distance profile-guided differential evolution," *IEEE/ACM Trans. Comput. Biol. Bioinf.* **14**, 1288–1301 (2016).
- ²⁹Y. Guo and Y. Wang, "Predicting the non-compact conformation of amino acid sequence by particle swarm optimization," in *2013 7th International Conference on Systems Biology (ISB)* (IEEE, 2013), pp. 119–122.
- ³⁰F. Daeyaert, M. De Jonge, L. Koymans, and M. Vinkers, "An ant algorithm for the conformational analysis of flexible molecules," *J. Comput. Chem.* **28**(5), 890–898 (2007).
- ³¹F. Glover and M. Laguna, *Tabu Search* (Kluwer Academic Publishers, 1997).
- ³²D. H. Wolpert and W. G. Macready, "No free lunch theorems for optimization," *IEEE Trans. Evol. Comput.* **1**(1), 67–82 (1997).
- ³³L.-P. Wang, K. A. McKiernan, J. Gomes, K. A. Beauchamp, T. Head-Gordon, J. E. Rice, W. C. Swope, T. J. Martínez, and V. S. Pande, "Building a more predictive protein force field: A systematic and reproducible route to AMBER-FB15," *J. Phys. Chem. B* **121**(16), 4023–4039 (2017).
- ³⁴E. Harder, W. Damm, J. Maple, C. Wu, M. Rebol, J. Y. Xiang, L. Wang, D. Lupyan, M. K. Dahlgren, J. L. Knight, J. W. Kaus, D. S. Cerutti, G. Krilov, W. L. Jorgensen, R. Abel, and R. A. Friesner, "OPLS3: A force field providing broad coverage of drug-like small molecules and proteins," *J. Chem. Theory Comput.* **12**(1), 281–296 (2016).
- ³⁵B. Chandramouli, S. Del Galdo, G. Mancini, N. Tasinato, and V. Barone, "Tailor-made computational protocols for precise characterization of small biological building blocks using QM and MM approaches," *Biopolymers* **109**, e23109 (2018).
- ³⁶*Computational Strategies for Spectroscopy: From Small Molecules to Nano Systems*, edited by V. Barone (Wiley, 2011).
- ³⁷S. Spicher and S. Grimme, "Robust atomistic modeling of materials, organometallic and biochemical systems," *Angew. Chem., Int. Ed.* **132**(1), 1–12 (2020).
- ³⁸B. Chandramouli, S. Del Galdo, M. Fusè, V. Barone, and G. Mancini, "Two-level stochastic search of low-energy conformers for molecular spectroscopy: Implementation and validation of MM and QM models," *Phys. Chem. Chem. Phys.* **21**, 19921–19934 (2019).
- ³⁹D. Porezag, T. Frauenheim, T. Köhler, G. Seifert, and R. Kaschner, "Construction of tight-binding-like potentials on the basis of density-functional theory: Application to carbon," *Phys. Rev. B* **51**(19), 12947–12957 (1995).
- ⁴⁰J. P. Stewart, "Optimization of parameters for semiempirical methods V: Modification of NDDO approximations and application to 70 elements," *J. Mol. Model.* **13**(12), 1173–1213 (2007).
- ⁴¹V. K. Prasad, A. Otero-de-la-Roza, and G. A. DiLabio, "Atom-centered potentials with dispersion-corrected minimal-basis-set Hartree-Fock: An efficient and accurate computational approach for large molecular systems," *J. Chem. Theory Comput.* **14**(2), 726–738 (2018).
- ⁴²C. Bannwarth, S. Ehlert, and S. Grimme, "GFN2-XTB—An accurate and broadly parametrized self-consistent tight-binding quantum chemical method with multipole electrostatics and density-dependent dispersion contributions," *J. Chem. Theory Comput.* **15**(3), 1652–1671 (2019).
- ⁴³J. Brownlee, *Clever Algorithms: Nature-Inspired Programming Recipes*, LuLu.com, s.l., 2nd revision edition (OCLC, 2012).
- ⁴⁴V. Barone, M. Biczysko, J. Bloino, and C. Puzzarini, "Characterization of the elusive conformers of glycine from state-of-the-art structural, thermodynamic, and spectroscopic computations: Theory complements experiment," *J. Chem. Theory Comput.* **9**(3), 1533–1547 (2013).
- ⁴⁵T. Szidarovszky, G. Czakó, and A. G. Császár, "Conformers of gaseous threonine," *Mol. Phys.* **107**(8), 761–775 (2009).
- ⁴⁶K. He and W. D. Allen, "Conformers of gaseous serine," *J. Chem. Theory Comput.* **12**(8), 3571–3582 (2016).
- ⁴⁷J. J. Wilke, M. C. Lind, H. F. Schaefer, A. G. Császár, and W. D. Allen, "Conformers of gaseous cysteine," *J. Chem. Theory Comput.* **5**(6), 1511–1523 (2009).
- ⁴⁸G. Szilvágyi, Z. Majer, E. Vass, and M. Hollósi, "Conformational studies on chiral rhodium complexes by ECD and VCD spectroscopy," *Chirality* **23**(4), 294–299 (2011).
- ⁴⁹B. Assadollahzadeh, P. R. Bunker, and P. Schwerdtfeger, "The low lying isomers of the copper nonamer cluster, Cu₉," *Chem. Phys. Lett.* **451**(4), 262–269 (2008).
- ⁵⁰J. L. Llanio-Trujillo, J. M. C. Marques, and F. B. Pereira, "An evolutionary algorithm for the global optimization of molecular clusters: Application to water, benzene, and benzene cation," *J. Phys. Chem. A* **115**(11), 2130–2138 (2011).
- ⁵¹J. H. Holland, "Adaptation in natural and artificial systems: An introductory analysis with applications to biology, control, and artificial intelligence," in *Complex Adaptive Systems*, 1st MIT press ed. (MIT Press, 1992).
- ⁵²D. Whitley, "A genetic algorithm tutorial," *Stat. Comput.* **4**(2), 65–85 (1994).
- ⁵³Y. Nourani and B. Andresen, "A comparison of simulated annealing cooling strategies," *J. Phys. A: Math. Gen.* **31**(41), 8373–8385 (1998).
- ⁵⁴T. Bäck and H.-P. Schwefel, "An overview of evolutionary algorithms for parameter optimization," *Evol. Comput.* **1**(1), 1–23 (1993).
- ⁵⁵D. Ferro-Costas and A. Fernández-Ramos, "A combined systematic-stochastic algorithm for the conformational search in flexible acyclic molecules," *Front. Chem.* **8**, 16 (2020).
- ⁵⁶F. Lazzari, A. Salvadori, G. Mancini, and V. Barone, "Molecular perception for visualization and computation: The proxima library," *J. Chem. Inf. Model.* **60**, 2668–2672 (2020).

- ⁵⁷J. Parsons, J. B. Holmes, J. M. Rojas, J. Tsai, and C. E. M. Strauss, "Practical conversion from torsion space to Cartesian space for silico protein synthesis," *J. Comput. Chem.* **26**(10), 1063–1068 (2005).
- ⁵⁸D. Whitley, S. Rana, and R. B. Heckendorn, "The island model genetic algorithm: On separability, population size and convergence," *J. Comput. Info. Technol.* **7**(1), 33–47 (1999).
- ⁵⁹L. Kaufmann and P. J. Rousseeuw, "Partitioning around Medoids (Program PAM)," in *Finding Groups in Data: An Introduction to Cluster Analysis*, edited by L. Kaufman and P. J. Rousseeuw (Wiley, Hoboken, NJ, 1990), pp. 68–125.
- ⁶⁰J. Han and M. Kamber, *Data Mining: Concepts and Techniques*, 3rd ed. (Elsevier, 2011).
- ⁶¹C. C. Aggarwal and C. K. Reddy, *Data Clustering: Algorithms and Applications* (Chapman and Hall/CRC, 2014).
- ⁶²T. Hastie, R. Tibshirani, and J. Friedman, *The Elements of Statistical Learning*. Springer Series in Statistics (Springer New York, 2009).
- ⁶³D. Licari, M. Fusè, A. Salvadori, N. Tassinato, M. Mendolicchio, G. Mancini, and V. Barone, "Towards the SMART workflow system for computational spectroscopy," *Phys. Chem. Chem. Phys.* **20**, 26034–26052 (2018).
- ⁶⁴G. Mancini and C. Zazza, "F429 regulation of tunnels in cytochrome p450 2b4: A top down study of multiple molecular dynamics simulations," *PLoS One* **10**(9), e0137075 (2015).
- ⁶⁵J. L. Alonso, C. Pérez, M. Eugenia Sanz, J. C. López, and S. Blanco, "Seven conformers of L-threonine in the gas phase: A LA-MB-FTMW study," *Phys. Chem. Chem. Phys.* **11**(4), 617–627 (2009).
- ⁶⁶A. D. Becke, "Density-functional exchange-energy approximation with correct asymptotic behavior," *Phys. Rev. A* **38**, 3098–3100 (1988).
- ⁶⁷R. Sure and S. Grimme, "Corrected small basis set Hartree-Fock method for large systems," *J. Comput. Chem.* **34**(19), 1672–1685 (2013).
- ⁶⁸M. J. Frisch, G. W. Trucks, H. B. Schlegel, G. E. Scuseria, M. A. Robb, J. R. Cheeseman, G. Scalmani, V. Barone, G. A. Petersson, H. Nakatsuji, X. Li, M. Caricato, A. V. Marenich, J. Bloino, B. G. Janesko, R. Gomperts, B. Mennucci, H. P. Hratchian, J. V. Ortiz, A. F. Izmaylov, J. L. Sonnenberg, D. Williams-Young, F. Ding, F. Lipparini, F. Egidi, J. Goings, B. Peng, A. Petrone, T. Henderson, D. Ranasinghe, V. G. Zakrzewski, J. Gao, N. Rega, G. Zheng, W. Liang, M. Hada, M. Ehara, K. Toyota, R. Fukuda, J. Hasegawa, M. Ishida, T. Nakajima, Y. Honda, O. Kitao, H. Nakai, T. Vreven, K. Throssell, J. A. Montgomery, Jr., J. E. Peralta, F. Ogliaro, M. J. Bearpark, J. J. Heyd, E. N. Brothers, K. N. Kudin, V. N. Staroverov, T. A. Keith, R. Kobayashi, J. Normand, K. Raghavachari, A. P. Rendell, J. C. Burant, S. S. Iyengar, J. Tomasi, M. Cossi, J. M. Millam, M. Klene, C. Adamo, R. Cammi, J. W. Ochterski, R. L. Martin, K. Morokuma, O. Farkas, J. B. Foresman, and D. J. Fox, Gaussian 16 Revision B.01, Gaussian, Inc., Wallingford, CT, 2016.
- ⁶⁹L. Goerigk and S. Grimme, "Efficient and accurate double-hybrid-meta-GGA density functionals—Evaluation with the extended GMTKN30 database for general main group thermochemistry, kinetics, and noncovalent interactions," *J. Chem. Theory Comput.* **7**(2), 291–309 (2011).
- ⁷⁰S. Grimme, S. Ehrlich, and L. Goerigk, "Effect of the damping function in dispersion corrected density functional theory," *J. Comput. Chem.* **32**(7), 1456–1465 (2011).
- ⁷¹T. H. Dunning, "Gaussian basis sets for use in correlated molecular calculations. I. The atoms boron through neon and hydrogen," *J. Chem. Phys.* **90**(2), 1007–1023 (1989).
- ⁷²E. Papajak, J. Zheng, X. Xu, H. R. Leverentz, and D. G. Truhlar, "Perspectives on basis sets beautiful: Seasonal plantings of diffuse basis functions," *J. Chem. Theory Comput.* **7**(10), 3027–3034 (2011).
- ⁷³M. Fusè, G. Mazzeo, G. Longhi, S. Abbate, M. Masi, A. Evidente, C. Puzzarini, and V. Barone, "Unbiased determination of absolute configurations by vis-à-vis comparison of experimental and simulated spectra: The challenging case of diplopyrone," *J. Phys. Chem. B* **123**(43), 9230–9237 (2019).
- ⁷⁴A. K. Rappe, K. S. Colwell, and C. J. Casewit, "Application of a universal force field to metal complexes," *Inorg. Chem.* **32**(16), 3438–3450 (1993).
- ⁷⁵J. M. L. Martin and A. Sundermann, "Correlation consistent valence basis sets for use with the Stuttgart–Dresden–Bonn relativistic effective core potentials: The atoms Ga–Kr and in–Xe," *J. Chem. Phys.* **114**(8), 3408–3420 (2001).
- ⁷⁶D. Andrae, U. Häußermann, M. Dolg, H. Stoll, and H. Preuß, "Energy-adjusted *ab initio* pseudopotentials for the second and third row transition elements," *Theor. Chim. Acta* **77**(2), 123–141 (1990).
- ⁷⁷F. A. Cotton, B. G. DeBoer, M. D. LaPrade, J. R. Pipal, and D. A. Ucko, "The crystal and molecular structures of dichromium tetraacetate dihydrate and dirhodium tetraacetate dihydrate," *Acta Crystallogr., Sect. B* **27**(8), 1664–1671 (1971).
- ⁷⁸M. Cossi, N. Rega, G. Scalmani, and V. Barone, "Energies, structures and electronic properties of molecules in solution by the C-PCM solvation model," *J. Comput. Chem.* **24**, 669–681 (2003).
- ⁷⁹A. Salvadori, G. Del Frate, M. Pagliai, G. Mancini, and V. Barone, "Immersive virtual reality in computational chemistry: Applications to the analysis of QM and MM data," *Int. J. Quantum Chem.* **116**(22), 1731–1746 (2016).