



GaudiMM: A Modular Multi-Objective Platform for Molecular Modeling

Jaime Rodríguez-Guerra Pedregal ^{*}, Giuseppe Sciortino, Jordi Guasp, Martí Municoy, and Jean-Didier Maréchal ^{*}

GaudiMM (for Genetic Algorithms with Unrestricted Descriptors for Intuitive Molecular Modeling) is here presented as a modular platform for rapid 3D sketching of molecular systems. It combines a Multi-Objective Genetic Algorithm with diverse molecular descriptors to overcome the difficulty of generating candidate models for systems with scarce structural data. Its grounds consist in transforming any molecular descriptor (i.e. those generally used for analysis of data) as a guiding objective for PES explorations. The platform is written in Python with flexibility in mind: the user can choose which descriptors

to use for each problem and is even encouraged to code custom ones. Illustrative cases of its potential applications are included to demonstrate the flexibility of this approach, including metal coordination of multidentate ligands, peptide folding, and protein-ligand docking. GaudiMM is available free of charge from <https://github.com/insilichem/gaudi>. © 2017 Wiley Periodicals, Inc.

DOI: 10.1002/jcc.24847

Introduction

Molecular modeling has become a major tool in many fields of chemistry and its interfaces. Its final objective is to accurately describe the structural and energetic properties of molecular systems; ultimately from scratch. Most of nowadays computational chemistry exercises need to start with clear structural data on the system of interest (i.e., X-ray or NMR structures). Unfortunately, this first piece of information is frequently missing or highly incomplete and in fact many projects strive to start because of the complexity to find convenient starting points in a reasonably short amount of time. When modelers try to overcome this “blank sheet” syndrome, they generally go through an iterative process of trial and error and manual adjustments, with the only guidance of his or her chemical intuition and/or experimental observations.

Most of the procedures used to rapidly identify physically sound initial models of a molecular system stand on finding the way to ally the exploration of wide conformational spaces and the adequate guiding variables. Among the most frequent tools for this task are Monte Carlo,^[1,2] Random Walk,^[3] Simulated Annealing,^[4–6] and evolutionary algorithms (EA).^[7,8] In those approaches, the way to bias the exploration generally stands on (1) using energetic evaluation of the molecular geometry (i.e., force field) and (2) impose additional simple Euclidian restraints like distances, angles, or dihedrals that could account on structural aspects hypothesis. The construction and assessment of an initial molecular candidate is therefore generally limited to potential energy surface (PES) considerations, eventually accounting for complex force field parametrization, and few guiding elements. However, there is much more structural information that the researchers could account on that are generally neglected at the moment.

Genetic Algorithms (GA)^[9] are a kind of EA that have been increasingly applied for complex molecular problems such as molecular matching,^[10,11] protein-ligand docking,^[12,13] or conformational searches.^[14,15] The most popular implementation in these applications are GAs that use a single objective strategy with a unique fitness function targeted. However, for years now, multi-objective genetic algorithms (MOGA) like NSGA^[16–18] and SPEA^[19–21] families are readily available and could bring novelties in the way we deal with molecular modeling. MOGAs are particularly helpful when different variables of the system fight ones against others to reach a tradeoff, especially if their importance is not known beforehand—a prototypical situation in initial model building. MOGAs tend to be applied when substantially different solutions could exist for the same problem. In principle, this kind of approaches could be useful on complex molecular modeling exercises when only partial information is available at the starting point of the study. The potential of multi-objective optimization in molecular modeling has been demonstrated by some recent developments,^[22,23] but there is plenty of room for advances of MOGA applications in Molecular computational chemistry and more particularly in providing a modular platform able to deal with relevant molecular descriptors.

Here, we present GaudiMM (for Genetic Algorithms with Unrestricted Descriptors for Intuitive Molecular Modeling), a GA-based platform for 3D sketching of molecular systems that expands the idea of PES guiding using molecular objectives.

J. R.-G. Pedregal, G. Sciortino, J. Guasp, Martí Municoy, J.-D. Maréchal
Departament de Química, Universitat Autònoma de Barcelona, Cerdanyola
del Vallès, Barcelona 08193, Spain E-mail: jaime.rodriguezguerra@uab.cat
E-mail: jeandidier.marechal@uab.cat

Contract grant sponsors: Spanish MINECO (project CTQ2014-54071-P), Generalitat de Catalunya (project 2014SGR989; to J.R.G.P.), and UAB (to G.S.)

© 2017 Wiley Periodicals, Inc.

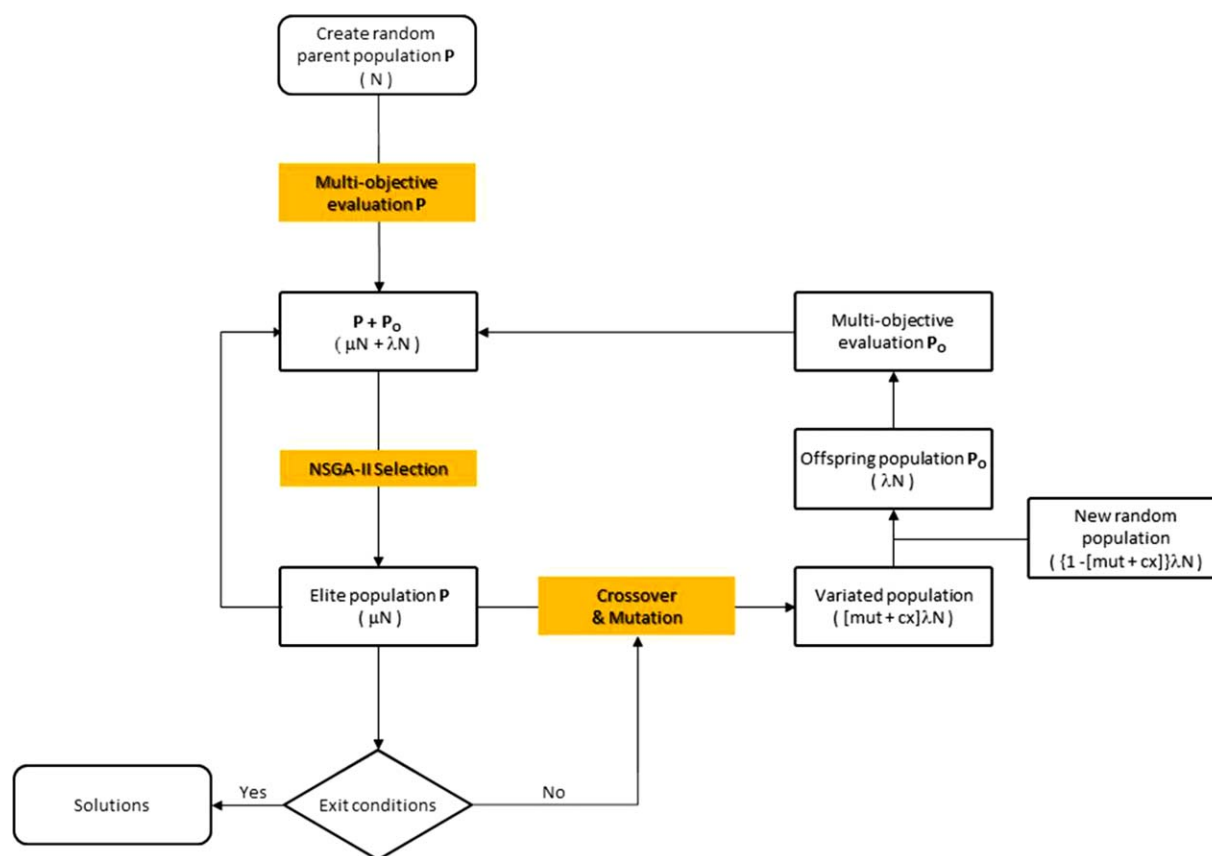


Figure 1. Flowchart of the modular NSGA-II multi-objective genetic algorithm implemented in GaudiMM. N is the number of individuals in the initial population P ; μ and λ , are, respectively, related to the number of individuals selected for the next generation and the number of children produced at each generation (offspring population P_o); mut and cx are the probabilities associated to mutation and crossover.^[24] [Color figure can be viewed at wileyonlinelibrary.com]

GaudiMM aims at generating accurate enough geometric candidates on systems where pre-existing structural data is limited. The architecture of this modular platform is written in Python and mainly combines a MOGA with molecular descriptors to explore physically sound geometries by escaping usual force field and scoring function limitations. GaudiMM is therefore an interesting tool to perform hypothesis driven traversal of the conformational landscape of a molecular system with the support of the chemical intuition and experimental knowledge of the researcher. To illustrate its potential, we present several practical cases mainly focusing on biochemical systems and recognition processes.

Methods

GaudiMM is built on top of a modular NSGA-II multi-objective GA, which has been thoroughly tested and benchmarked in well-characterized multi-objective problems.^[17] Multi-objective means that it can optimize all the needed variables (*objectives*) at once, without compromising any of them over the rest. A MOGA starts with the generation of a random set of potential solutions (*individuals*) which comprise the so-called initial population. This first set of individuals is then evaluated with several objectives and each one is assigned a *fitness* value to find the best ones, which are then allowed to recombine and

mutate. The results of these variations can be better or worse than their preceding individuals (parents), but only the best of both the offspring and the parental generation ($\mu + \lambda$ strategy) will be selected by the algorithm and propagated to the next generation. After a number of iterations, the initial population will evolve and, eventually, answer with a number of reasonable solutions to the problem. A flowchart of the NSGA-II algorithm is shown in Figure 1.

Implementation details

On one hand, the core GA is based on the implementation found in the Deap Python package.^[24] On top of Deap, GaudiMM provides an object-oriented programming interface that emphasizes the conceptual difference between problem exploration and evaluation: that is, between *genes* and *objectives*. Diversity is ensured through a fitness-tie and structural similarity analysis. If two candidate solutions share the same fitness value and their RMSD value is under a threshold, one of them will be discarded. As the user can specify both the RMSD threshold and the decimal precision of the reported fitness values, different levels of diversity and/or crowding can be achieved easily.

On the other hand, UCSF Chimera^[25] provides a robust molecular framework that allows rapid development of such

Table 1. List of genes implemented in GaudiMM.

Name	Description	Depends on
Molecule	Load and build structures	UCSF Chimera
Rotamers	Explore side chain flexibility	UCSF Chimera
Mutamers	Explore mutation of residues	UCSF Chimera
NormalModes	Explore collective motions	ProDy
Search	Translation and rotation of Molecules	UCSF Chimera
Torsion	Dihedral rotation of bonds	UCSF Chimera

genes and objectives, and also doubles as powerful visualization tool. While most genes and objectives are currently wrappers or extended interfaces of parts of UCSF Chimera, additional third-party libraries can be used, like OpenMM for force field energies,^[26] DrugScoreX^[27] and IMP^[28] for docking scoring functions, or ProDy for normal modes analysis.^[29]

Exploration: Individuals and genes. Each possible solution to the problem is represented by an Individual object. These objects contain a sorted list of genes and a series of helper methods. In our implementation, a gene describes a molecular feature, such as the topology of the molecule itself, a flexibility model (rotamers, torsions, normal modes), or the spatial orientation. As each gene is a separate module and class, they can be loaded only when required, as many times as needed. This also allows to offer custom mutation and recombination strategies to suit the nature of the feature, instead of using a single global strategy for all genes.

By creating different Individual objects, each with the same type of genes but with different values or *alleles*, we get to explore the biochemical and conformational search spaces. As a result, depending on the problem at hand, an Individual can be as simple as a rigid molecule that only moves around, or as complex as a two competing, dynamically built peptide with backbone torsions and rotameric side chains (Table 1).

Evaluation: Environment and objectives. After creating a number of individuals, those must be evaluated to tell how good of a solution they are. All the individuals are subject to the same conditions: the environment, represented by an Environment class that wraps a sorted list of objectives each individual must face. Each objective is separate module and class, which defines a single method named evaluate that contains the fitness assessment code. As long as this function returns a

quantity that can be maximized or minimized, it will be a valid objective. The same approach works equally for energy estimations, structure-focused optimizations and trivial restraints (distances, angles, dihedrals, surface areas, volume...). The code itself can be as complex as a H-bond detection engine, or a simple wrapper around a well-known executable that takes Protein Data Bank^[30] files as inputs (Table 2).

As a result, any geometric or energetic parameters that could describe a molecular system can be used as objectives to drive the GA exploration. This allows us to turn the tables on routine protocols based on computing energetic optimizations and then analyzing the results in hopes of finding a suitable model that fits the intended restraints; that is, those same analysis tools can guide the optimization process from the beginning.

Usability

GaudiMM only requires a plain text YAML^[31] input file, which specifies the genes that describe the molecular system, the objectives that will guide the simulated evolution, and the output parameters. (i.e., where to write the results files). A prototypical example of an input file can be found in the Supporting Information. Calculations are then started from the command line with `gaudi run input_file.yaml`, or from an in-house developed GUI, GAUDIInspect, that will also help you configure the input file. The results can be analyzed with that same GUI, or with the help of GaudiView, another GUI built as an UCSF Chimera extension.^[32]

Being user-friendly does not mean developers are not welcome in this project. As GaudiMM is designed with extensibility in mind, the users are able and encouraged to develop their own genes and objectives if the built-in ones do not fulfill their requirements. Ultimately, such extensibility could be taken even further to provide different MOGAs for the users to choose, or allow them to code their own ones.

Further details on how to create input files, analyze results or code custom genes or objectives can be found in the attached documentation in the Supporting Information.

Results on Illustrative Cases

Peptide folding under a given volume

Peptides are receiving an increasing amount of interest at the interface between biology and chemistry. From models of

Table 2. List of objectives implemented in GaudiMM.

Name	Description	Depends on
Angle	Optimize angle of three atoms, or dihedral of four atoms	UCSF Chimera
Contacts	Minimize steric clashes, maximize hydrophobic interactions	UCSF Chimera
Coordination	Optimize coordination geometry of metal center	In-house
Distance	Optimize distance between two or more atoms	UCSF Chimera
DSX	Docking scoring function	DrugScoreX
Energy	Minimize molecular mechanics potential energy	OpenMM
HBonds	Detect hydrogen bonds	UCSF Chimera
Inertia	Align axes of inertia of two or more molecules	In-house
LigScore	Docking scoring function	IMP
Solvation	Measure solvent accessible solvent area	UCSF Chimera
Volume	Measure volume occupied by molecule	UCSF Chimera

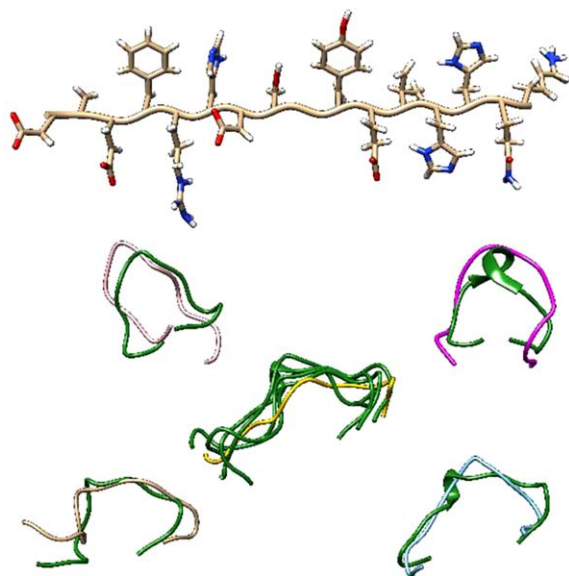


Figure 2. On top, unfolded linear structure of the peptide. On the bottom, alignment of some predicted structures (green) with some of the NMR structures of 1ZE9 (other colors). Only the backbone is displayed in the latter case. [Color figure can be viewed at wileyonlinelibrary.com]

protein behavior to their role as drugs or biosensors, the use of peptides is constantly growing.^[33–39] However, it is well-known that peptides are small polymers with a high degree of flexibility and the exploration of relevant conformations for a given purpose can be particularly complex. After all, only certain conformations are usually biologically active.

To illustrate how molecular objectives could be interesting in the context of a hypothesis-driven simulation, we decided to look at the capabilities of GaudiMM to generate structures of a peptide so that it could fit into a particular volume.

As a case study, calculations were performed on the sequence for Alzheimer β -amyloid structure 1–16 (βA_{1-16}) starting from a purely linear geometry as provided by the peptide builder of UCSF Chimera. The input data consisted of three genes—(1) the linear peptide structure, (2) backbone torsion of the peptide, (3) rotameric exploration of sidechains of the peptide— and four objectives: (1) minimization of clashes to rapidly discard unfeasible conformations, (2) minimization of the potential energy as calculated within the AMBER99SB-ILDN force field,^[40] and (3) a spatial optimization to match 15,854 Å³. This value was selected because it represents the average volume occupied by the Zn-folded NMR structures of the same peptide^[41] and so intrinsically shows the possible pre-organization of the isolated peptide for metal binding (PDB ID: 1ZE9). The complete input is reported in Supporting Information.

As all the objectives competed between them during the optimization in GaudiMM, after 60,300 evaluations, there was obviously not a unique solution for this problem. However, by filtering the proposed solutions of the multi-objective run with reasonable cutoffs (volumetric overlap of Van der Waals spheres under 100 Å³), one could clearly see that the solutions with a good compromise between all the different objectives

present a relatively good folding geometry when compared to the experiment.

From the different solutions, most were able to generate a geometry that satisfy the volumetric objective with a difference between the final objective and the best solution being within a 16% deviation from the target value. As a final analysis, we compared the geometry obtained with the zinc bound system of the beta amyloid peptide. As one would expect, the geometry was not a perfect match as the metal was not explicitly considered in the calculation. Even under those limitations, the mean RMSD fell within the 3.4–3.8 Å range (Fig. 2). In Table 3, we report the best solutions of the simulation. A complete sample can be found in Supporting Information table ESI.1.

Prediction of metal bound form of siderophore

Truth to be said, GaudiMM was born based on our experience in dealing with bioinorganic systems and more particularly in our axes of research toward artificial metalloenzymes and metallocdrugs. On these systems, one of the most interesting aspects consists of rapidly generating structures for metal coordinating organic molecules or biomolecules circumventing the major complexities of parametrizing metal moieties for a given force field at the early stage of the modeling.

As GaudiMM offers a flexible platform based on modules, the user can mix non-trivial geometrical objectives with standard energetic terms. In this problem, the geometry of the non-metallic part could be dealt with a standard force field and the metallic part could be computed with carefully selected geometric terms. As a result, predicting the structure of a multichelating ligand to a metal becomes a feasible exercise: the only requirement is to select well those geometric terms.

The coordination geometry of said complex can be identified by comparing the positions of the potential ligand atoms around the metal center against the vertices of a number of polyhedra and checking if the directionality of ligand-metal interactions is correct.

To test how GaudiMM could answer to the question, we intended to reproduce the experimental structure of the enterobactin siderophore using only the isolated minimized structure of the apo enterobactin (PubChem CID: 34231)^[42] with an

Table 3. Best[†] solutions found for the folding simulation.

Energy ^[a]	Clashes ^[b]	Volume ^[c]	%D ^[d]	RMSD _{best} ^[e]
16685.084	31.052	17520.165	9.5	3.416
15621.915	41.606	17243.505	8.1	3.865
15165.133	36.958	17540.693	9.6	3.108
14909.464	23.340	17781.583	10.8	3.362
13781.994	9.382	18920.091	16.2	3.280
RMSD _{mean}				3.406

[†] Table obtained using the cutoffs: 13,000 < Volume < 19,000 and Clashes < 100 Å³. [a] Energy in Kcal/mol. [b] Clashes measured in Å³. [c] Volume in Å³. [d] Percent deviation from the mean value (15,854 Å³). [e] Best RMSD on the heavy atoms obtained comparing the GaudiMM solution with the NMR solved structure of the peptide (PDB ID: 1ZE9).

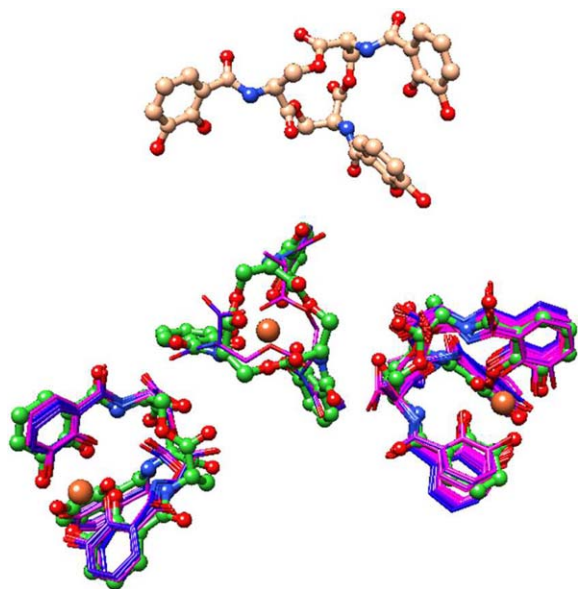


Figure 3. Unfolded enterobactin siderophore (above, tan color, ball, and stick) was successfully folded (below, in blue and pink, thin sticks) and fitted into the three of the four siderophores found in *E. coli* holo FepB (PDB ID: 3TLK, in green, ball and stick). [Color figure can be viewed at wileyonlinelibrary.com]

iron atom as the starting structure. GaudiMM was therefore run using four genes—(1) a Molecule containing the initial structure of the siderophore as provided by the PubChem file, but with one less bond in the central ring to allow bond torsions, (2) said dihedral torsions for rotatable bonds in siderophore, (3) another Molecule with a bare iron ion, (4) free movement of the iron ion within 5 Å from the center of mass of the initial structure— and four objectives—(1) minimization of clashes to discard non-feasible torsions, (2) optimization of the first coordination sphere of the iron so it matches a tetrahedral geometry as observed in the siderophores found on *E. coli* holo FepB (PDB ID: 3TLK),^[43] (3) a helper distance optimization to get the potential ligand atoms closer to the metal at around 2.75 Å, and (4) another distance minimization so the opened ring of the siderophore does not unfold and stays at bonding distance. The complete input is reported in Supporting Information.

After 751,500 evaluations, the resulting solutions proposed 135 models that matched the experimental siderophore structures with an average RMSD value of 0.95 Å (Fig. 3). The best 20 solutions found by GaudiMM are reported in Table 4 and the complete set of 135 structures is available in Supporting Information table ESI.2.

To the best of our knowledge, GaudiMM is the first 3D builder to account in a relative straightforward manner with large multidentate ligand with metals.

Multi-objective protein-ligand docking

To further assess the interest of exploring the conformational space of a molecular system by competitive objectives optimization, we decided to test our approach in recognition processes. One of the most widely spread in chemical biology,

and more particularly drug design projects, is protein-ligand docking, which is aimed at predicting the geometric and energetic poses that a given ligand adopts within the binding cavity of a host. Although standard protein-ligand docking is not the target application of GaudiMM and many excellent programs provide excellent yields in this field,^[12,44,45] this is still a natural way forward to test how its multi-objective capabilities would behave on this kind. Indeed, as one can decide which objectives and genes to use for each system, different combinations or *recipes* will provide different levels of accuracy that can be adapted to the problem at hand.

The correct recipe depends on the user's interest but results are already quite impressive by simply using a fast filtering objective (typically clashes), a placement objective (i.e., hydrogen bonds) and a field-tested scoring function (i.e., DrugScoreX or LigScore).

The two scoring functions implemented in GaudiMM have been previously validated and compared with other functions in several other works published the literature. DrugScoreX was successfully tested on the set of 195 protein-ligand complexes prepared by Cheng et al.^[46] obtaining the best results with respect to docking power with a success ratio close to the 95%.^[27] LigScore was validated on the Wang_AutoDock testing set^[47] of 100 protein-ligand complexes obtaining a success ratio close to the 90%.^[48]

Calculations performed under our competitive recipe provide very interesting results that show the versatility of GaudiMM at facing different problems. For example, fast docking jobs could be carried out using rigid bodies and simplistic terms such as minimization of steric clashes, maximization of

Table 4. Best solutions[†] found the prediction of metal bound form of siderophore.

Coordination ^[a]	Clashes ^[b]	Dist.Cyc. ^[c]	Dist.M ^[c]	RMSD1 ^[d]
1.210	5.394	2.680	0.330	0.965
1.361	7.580	1.558	0.253	0.871
1.406	7.527	1.559	0.213	0.929
1.440	5.312	2.196	0.195	0.986
1.441	3.424	2.354	0.308	0.993
1.451	3.417	2.392	0.180	0.923
1.453	1.253	1.956	0.387	0.903
1.454	4.153	1.883	0.269	0.903
1.456	2.215	2.418	0.193	0.856
1.457	3.418	1.915	0.282	0.925
1.457	2.171	1.936	0.250	1.075
1.458	11.479	1.607	0.124	0.900
1.462	5.340	1.942	0.135	0.873
1.476	3.459	2.487	0.172	0.941
1.477	3.459	2.487	0.169	0.964
1.480	3.471	2.723	0.159	0.902
1.480	12.700	1.539	0.223	0.924
1.481	9.656	1.658	0.172	0.936
1.481	3.466	2.492	0.166	0.920
1.481	8.923	1.552	0.214	1.022
RMSD _{mean}				0.951

[†] Table obtained using a coordination cutoff of 3.000. [a] RMSD between the ideal polyhedron and the coordination geometry found in the calculation. [b] Clashes measured in Å³. [c] Distance in Å. [d] RMSD of the complete cluster calculated on the siderophore's heavy atoms via UCSF Chimera software.

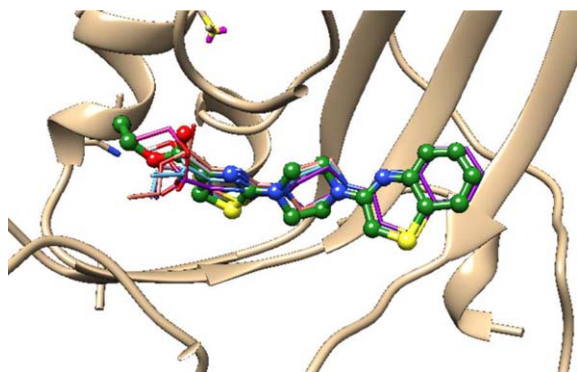


Figure 4. Docking of PDB structure 1VRH. Crystallographic structure is depicted in forest green with ball and stick representation, best solutions in other colors and thin sticks. [Color figure can be viewed at wileyonlinelibrary.com]

hydrophobic interactions, and optimization of H-bonds networks, while more accurate dockings could be performed with normal modes analysis, rotameric side chains and force field energy minimization. Non-conventional docking studies are also possible in GaudiMM thanks to its plurigenetic features, such as systems where multiple ligands compete for the same binding site simultaneously, or problems that require to optimize the size of certain substituents of a ligand within the cavity of a protein. However, exotic challenges like those are outside the scope of this publication and will be further discussed in submissions to come.

A first attempt to benchmark a simple LigScore recipe against the original GOLD dataset^[12] successfully reproduces 57.6% of the 100 crystallographic structures, which is comparable to well-established scoring functions already tested in several works published in the literature (GOLD reported a 69.7% success rate in its original publication).^[12,27,46–49] The benchmark is described in detail in the Supporting Information and the results are summarized in Supporting Information table ESI.6. While this is not the target usage of GaudiMM, we firmly believe that with directed efforts toward optimizing the sampling stage the number of hits would surely increase and rival those docking methods with higher accuracy. The following

Table 5. Best solutions[†] found for the docking calculation of the structure 1VRH.

DSX ^[a]	Clashes ^[b]	Contacts ^[c]	RMSD ^[d]
–146.928	14.109	–61.964	1.260
–144.814	10.462	–57.694	1.244
–144.567	29.241	–64.708	1.256
–143.170	6.463	–63.871	1.641
–138.634	3.704	–65.444	1.280
–137.029	2.428	–75.161	0.682
–131.682	4.317	–76.180	1.177
RMSD _{mean} ^[d]			1.220

[†] Table obtained using a scoring cutoff of DSX < 0. [a] DSX score value. [b] Clashes measured in Å³. [c] Sum of L.J.-like potential obtained for vdW hydrophobic contacts per the following formula: $U = \sum_{i,j} 4\epsilon \left[\left(\frac{\sigma}{r_{ij}} \right)^{12} - \left(\frac{\sigma}{r_{ij}} \right)^6 \right]$ with $\sigma = 1$ and $\epsilon = 0.25$. [d] RMSD (in Å) calculated on the heavy atoms via UCSF Chimera software.

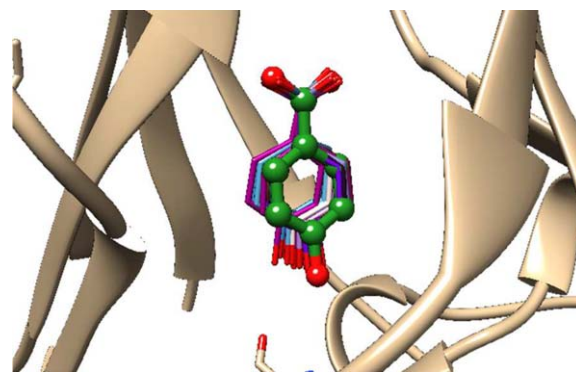


Figure 5. Docking of PDB structure 2PHH. Crystallographic structure is depicted in forest green with ball and stick representation, best solutions in other colors and thin stick representation. [Color figure can be viewed at wileyonlinelibrary.com]

paragraphs describe the results of three showcase docking calculations.

All of them were analyzed with full torsion flexibility on the non-minimized ligand, which could move and rotate within 12 Å of the crystallographic position. All of them successfully reproduced the experimental structure, but using different sets of objectives. The parameters of the GA guaranteed an average sample size of 60,600 different poses before proposing the best candidate solutions.

1VRH. Binding of a piperazine to *Rhinovirus B* HRV14/SDZ 880–061 complex^[50] was correctly reproduced with minimization of clashes, maximization of hydrophobic interactions and

Table 6. Best solutions[†] found for the docking calculation of the structure 2PHH.

LigScore ^[a]	H _{bonds} ^[b]	Clash. ^[c]	Contacts ^[d]	RMSD ^[e]
–21.89	3	11.716	–30.057	1.772
–21.20	3	5.209	–26.851	0.802
–20.26	6	3.321	–30.641	0.869
–18.34	3	25.337	–33.067	1.321
–17.51	1	9.167	–30.919	1.275
–16.54	3	10.737	–31.106	1.005
–16.08	3	4.936	–31.378	1.152
–15.39	5	9.356	–32.467	1.178
–14.91	1	46.005	–36.124	1.679
–14.29	4	7.107	–30.781	1.159
–14.24	2	6.647	–31.594	1.118
–14.21	2	6.043	–32.433	1.110
–13.89	2	31.899	–33.437	1.803
–13.10	2	42.476	–40.087	1.964
–13.08	1	40.404	–36.300	1.637
–12.87	2	39.935	–41.143	1.731
–12.62	2	29.729	–41.298	1.601
–12.46	1	35.954	–43.426	1.605
–12.25	0	25.041	–36.522	1.474
–11.99	1	39.028	–47.753	1.800
RMSD _{mean} ^[e]				1.552

[†] Table obtained using a scoring cutoff of LigScore < 0. [a] LigScore score value. [b] Number of H bonds. [c] Clashes measured in Å³. [d] Sum of L.J.-like potential obtained for vdW hydrophobic contacts per the following formula: $U = \sum_{i,j} 4\epsilon \left[\left(\frac{\sigma}{r_{ij}} \right)^{12} - \left(\frac{\sigma}{r_{ij}} \right)^6 \right]$ with $\sigma = 1$ and $\epsilon = 0.25$. [e] RMSD (in Å) calculated on the heavy atoms via UCSF Chimera software.

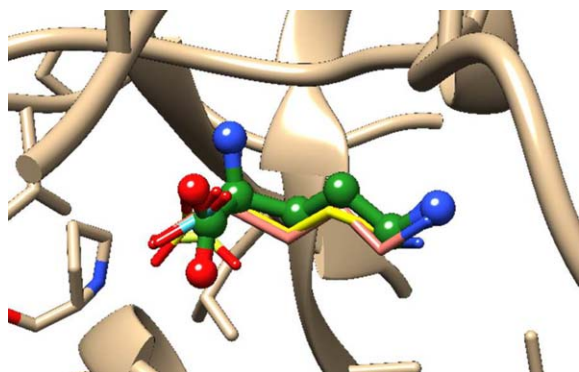


Figure 6. Docking of PDB structure 1LAH. Crystallographic structure is depicted in forest green with ball and stick representation, best solutions in other colors and thin sticks. [Color figure can be viewed at wileyonlinelibrary.com]

optimized DrugScoreX analysis. The complete input is reported in Supporting Information. The top seven solutions according to DrugScoreX reported less than 15 \AA^3 of volumetric overlap (clashes). The experimental structure is reproduced with high accuracy reporting the best RMSD of 0.682 \AA . The best solutions are shown in Figure 4 and reported in Table 5. A larger sample of solution is reported in Supporting Information table ESI.3.

2PHH. Crystallographic structure of adenosine 5-diphosphoribose in *Pseudomonas fluorescens* p-hydroxybenzoate hydroxylase^[51] was correctly reproduced in the first cluster. The experimental structure is predicted with the best RMSD of 0.802 \AA . In this case, clashes

minimization and hydrophobic interactions maximization were kept, but hydrogen bond network optimization was added and the scoring function was replaced with LigScore. The best poses are shown in Figure 5 and the first 20 solutions of the first cluster are reported in Table 6. A more complete sample can be found in Supporting Information table ESI.4.

1LAH. The same recipe used for 2PHH worked excellently for L-ornithine in *Salmonella enterica* periplasmic ornithine-binding protein.^[52] In this case, the best RMSD between the experimental structure and the simulated one results in 0.648 \AA . The top 20 solutions are reported in Table 7 and the best are shown in Figure 6. A more complete sample of solutions is reported in Supporting Information table ESI.5.

Our initial tests on the viability of GaudiMM to deal with protein-ligand docking problems have been particularly encouraging, illustrate its versatility and how multi-objective recipes provide this framework with a wide range of applications.

Conclusions

As molecular modelers strive to push the limits of the software they use, a shift from descriptive to predictive applications is occurring. The final goal of that quest is the generation of accurate 3D models from scratch. But to get there one must conquer a number of smaller milestones.

In principle, GaudiMM could be expected to be as accurate as major molecular modeling programs, as its accuracy mainly depends on which quality of the objectives and genes could be reached. While nothing prevents from developing, for example, a DFT objective to evaluate single-point energies of the candidates along the simulation, its main interest pivots toward getting molecular guesses in a reasonable amount of time to present physically and chemically sound models as the starting points of a multiscale protocol involving additional software: with a first set of GaudiMM solutions, one could set a long MD simulation, which in turn would return suitable models for QM/MM optimizations. However, this should not discourage from programming this hypothetical QM objective, as it could be indeed very useful in other computational chemistry fields, such as catalysis and surface science.

In a way, GaudiMM is more about answering the question: "if a molecule could respond to these restrictions, what geometry could be acceptable in that case." As long as the adequate genes and objectives are provided, the answer is guaranteed. We firmly believe that it can become a particularly important asset in nowadays molecular modeling community.

Further Work

The current state of GaudiMM can be considered the reference implementation of our proof-of-concept and we are now hardly working in getting the code to highest level of performance.

For example, the evaluation stage, typically the most time-consuming step, is performed individually for each candidate solution by design, which allows for easy parallelization with a multi-process strategy. This can be coupled with the idea that GaudiMM jobs can be thought as a series of evaluations of

Table 7. Best solutions[†] found for the docking calculation of the structure 1LAH.

LigScore ^[a]	H _{bonds} ^[b]	Clash. ^[c]	Contacts ^[d]	RMSD ^[e]
-21.00	6	2.743	-17.028	1.110
-20.07	6	0.000	-19.446	0.648
-20.02	6	0.000	-20.717	1.016
-19.15	6	5.364	-23.648	1.502
-18.60	7	4.595	-21.407	1.443
-18.19	4	5.984	-31.974	1.475
-17.85	6	7.070	-25.711	1.407
-17.34	6	4.274	-24.801	1.487
-17.14	6	4.624	-25.082	1.321
-17.14	5	3.560	-24.825	1.486
-17.07	5	3.565	-24.831	1.487
-16.11	7	3.382	-19.726	1.340
-16.09	5	4.402	-29.654	1.369
-16.07	6	3.855	-26.164	1.472
-16.04	7	6.693	-23.937	1.455
-15.40	8	14.366	-21.368	1.493
-15.30	4	2.395	-21.151	1.138
-14.92	6	3.966	-26.294	1.522
-14.77	5	5.984	-30.172	1.533
RMSD _{mean} ^[e]				1.486

[†] Table obtained using a scoring cutoff of LigScore < 0. [a] LigScore score value. [b] Number of Hbonds. [c] Clashes measured in \AA^3 . [d] Sum of L.J.-like potential obtained for vdW hydrophobic contacts per the following formula: $U = \sum_{i,j} 4\epsilon \left[\left(\frac{\sigma}{r_{ij}} \right)^{12} - \left(\frac{\sigma}{r_{ij}} \right)^6 \right]$ with $\sigma = 1$ and $\epsilon = 0.25$. [e] RMSD mean (in \AA) of the complete cluster calculated on the heavy atoms via UCSF Chimera software.

different set of coordinates generated by the exploration stage for one or more given molecular topologies, which means that some objectives could be re-engineered to make use of highly optimized trajectory analysis tools, such as MDTraj^[53] or PyTraj.^[54] Of course, if after these changes some parts are still slow, the standard procedures of optimization could still be applied: wrapping those methods with Numba^[55] or rewriting them as C extensions.

Acknowledgment

Support of COST Action CM1306 is kindly acknowledged.

Keywords: molecular modeling · protein-ligand docking · multi-objective optimization · genetic algorithms · metallopeptides

How to cite this article: J. Rodríguez-Guerra Pedregal, G. Sciortino, J. Guasp, M. Municoy, J.-D. Maréchal. *J. Comput. Chem.* **2017**, *38*, 2118–2126. DOI: 10.1002/jcc.24847



Additional Supporting Information may be found in the online version of this article.

- [1] M. G. Martin, *Mol. Simul.* **2013**, *39*, 1212.
- [2] W. L. Jorgensen, J. Tirado-Rives, *J. Comput. Chem.* **2005**, *26*, 1689.
- [3] K.-C. Huang, R. J. White, *J. Am. Chem. Soc.* **2013**, *135*, 12808.
- [4] D. Hohl, R. Idaszak, R. O. Jones, In Proceedings of Supercomput '90; IEEE Computer Society Press: New York, **1990**; pp. 816–825.
- [5] I. D. Kerr, R. Sankaramakrishnan, O. S. Smart, M. S. Sansom, *Biophys. J.* **1994**, *67*, 1501.
- [6] D. Baker, D. Weinfurter, T. Maurer, W. Gronwald, H. R. Kalbitzer, D. Baker, R. Bonneau, D. Baker, C. Hardin, T. Pogorelov, Z. Luthey-Schulten, A. Murzin, S. Brenner, T. Hubbard, C. Chothia, L. L. Conte, S. Brenner, T. Hubbard, C. Chothia, A. Murzin, C. Chothia, A. Lesk, C. Sander, R. Schneider, A. Martin, M. MacArthur, J. Thornton, D. Vitkup, E. Melamund, J. Moul, C. Sander, M. Martin-Renom, R. Stuart, A. Fiser, R. Sanchez, F. Melo, A. Sali, B. Al-Lazikani, J. Jung, Z. Xiang, B. Honig, J. Westbrook, Z. Feng, L. Chen, H. Yang, H. Berman, C. Orengo, J. Bray, D. Buchan, A. Harrison, D. Lee, F. Pearl, W. Pearson, S. Altschul, W. Gish, W. Miller, E. Myers, D. Lipman, S. Altschul, T. Madden, A. Schaffer, J. Zhang, Z. Zhang, W. Miller, R. David, M. Korenberg, I. Hunter, J. Thompson, T. Gibson, F. Plewniak, F. Jeanmougin, D. Higgins, W. Browne, A. North, D. Phillips, K. Brew, T. Vanaman, R. Hill, T. Blundell, B. Sibanda, M. Sternberg, J. Thornton, J. Greer, T. Jones, S. Thirup, R. Unger, D. Harel, S. Wherland, J. Sussman, M. Claessens, C. V. Cutsem, I. Lasters, S. Wodak, M. Levitt, T. Havel, M. Snow, S. Srinivasan, C. March, S. Sudarsanam, A. Sali, T. Blundell, S. Brocklehurst, R. Perham, A. Aszodi, W. Taylor, A. Kolinski, M. Betancourt, D. Kihara, P. Rotkiewicz, J. Skolnick, P. Güntert, C. Mumenthaler, K. Wüthrich, A. Brünger, P. Adams, G. Clore, W. DeLano, P. Gros, R. Grosse-Kunstleve, R. Döker, T. Maurer, W. Kremer, K.-P. Neidig, H. Kalbitzer, C. Zhang, J. Hou, S.-H. Kim, N. v. Nuland, I. Hangyi, R. v. Schaik, H. Berendsen, W. v. Gunsteren, R. Scheek, Z. Jia, J. Quail, E. Waygood, L. Delbaere, J. Uppenberg, C. Svensson, M. Jaki, G. Bertilsson, L. Jendeborg, A. Berkenstam, W. Gronwald, H. Kalbitzer, P. Postma, J. Lengeler, G. Jacobson, R. Laskowski, J. Rullmann, M. MacArthur, R. Kaptein, J. Thornton, H. Xu, M. Lambert, V. Montana, K. Plunket, L. Moore, J. Collins, H. Xu, T. Stanley, V. Montana, M. Lambert, B. Shearer, J. Cobb, P. Cronet, J. Petersen, R. Folmer, N. Blomberg, K. Sjöblom, U. Karlsson, H. Xu, M. Lambert, V. Montana, D. Parks, S. Blanchard, P. Brown, W. Gronwald, R. Kirchhofer, A. Görler, W. Kremer, B. Gansmeier, K. Neidig, C. Dominguez, R. Boelens, A. Bonvin, J. Linge, M. Habeck, W. Rieping, M. Nilges, J. Linge, M. Williams, C. Spronk, A. Bonvin, M. Nilges, S. Nabuurs, A. Nederveen, W. Vranken, J. Doreleijers, A. Bonvin, G. Vuister, C. Gröger, A. Möglich, M. Pons, B. Koch, W. Hengstenberg, H. Kalbitzer, A. Möglich, B. Koch, W. Gronwald, W. Hengstenberg, E. Brunner, H. Kalbitzer, N. Tjandra, A. Bax, E. Brunner, R. Koradi, M. Billeter, K. Wüthrich, A. Görler, H. Kalbitzer, A. Görler, W. Gronwald, K. Neidig, H. Kalbitzer, Z. Jia, M. Vandonselaar, W. Hengstenberg, J. Quail, L. Delbaere, T. Maurer, R. Döker, A. Görler, W. Hengstenberg, H. Kalbitzer, A. Görler, W. Hengstenberg, M. Kravanja, W. Beneicke, T. Maurer, H. Kalbitzer, B. Jones, P. Rajagopal, R. Klevit, *Nature* **2000**, *405*, 39.
- [7] D. Douguet, E. Thoreau, G. Grassy, *J. Comput. Aided. Mol. Des.* **2000**, *14*, 449.
- [8] G. M. Verkhivker, P. A. Rejto, D. K. Gehlhaar, S. T. Freer, *Proteins Struct. Funct. Bioinform.* **1996**, *25*, 342.
- [9] D. E. Goldberg, *Genetic Algorithms in Search, Optimization, and Machine Learning*, Addison-Wesley Longman Publishing Co., Inc.: Boston, MA, 1989.
- [10] G. Jones, P. Willett, R. C. Glen, *J. Comput. Aided. Mol. Des.* **1995**, *9*, 532.
- [11] R. D. Brown, G. Jones, P. Willett, R. C. Glen, *J. Chem. Inf. Comput. Sci.* **1994**, *34*, 63.
- [12] G. Jones, P. Willett, R. C. Glen, A. R. Leach, R. Taylor, *J. Mol. Biol.* **1997**, *267*, 727.
- [13] G. M. Morris, D. S. Goodsell, R. S. Halliday, R. Huey, W. E. Hart, R. K. Belew, A. J. Olson, *J. Comput. Chem.* **1998**, *19*, 1639.
- [14] N. Nair, J. M. Goodman, *J. Chem. Inf. Comput. Sci.* **1998**, *38*, 317.
- [15] M. L. Beckers, L. M. Buydens, J. A. Pikkemaat, C. Altona, *J. Biomol. NMR* **1997**, *9*, 25.
- [16] N. Srinivas, K. Deb, *Evol. Comput.* **1994**, *2*, 221.
- [17] K. Deb, A. Pratap, S. Agarwal, T. Meyarivan, *IEEE Trans. Evol. Comput.* **2002**, *6*, 182.
- [18] K. Deb, H. Jain, *IEEE Trans. Evol. Comput.* **2014**, *18*, 577.
- [19] E. Zitzler, L. Thiele, TIK-Report 43. Computer Engineering and Networks Laboratory (TIK), Department of Electrical Engineering, Swiss Federal Institute of Technology (ETH), Zurich ETH Zentrum, Zurich, **1998**.
- [20] E. Zitzler, M. Laumanns, L. Thiele, TIK-Report 103. Computer Engineering and Networks Laboratory (TIK), Department of Electrical Engineering, Swiss Federal Institute of Technology (ETH), Zurich ETH Zentrum, Zurich, **2001**.
- [21] M. Kim, T. Hiroyasu, M. Miki, S. Watanabe, In *Parallel Problem Solving from Nature - PPSN VIII*. PPSN 2004. Lecture Notes in Computer Science, X. Yao, et al., Eds.; Springer: Berlin, Heidelberg, **2004**; pp. 742–751.
- [22] M. J. Vainio, M. S. Johnson, *J. Chem. Inf. Model.* **2007**, *47*, 2462.
- [23] A. Grosdidier, V. Zoete, O. Michielin, *Proteins Struct. Funct. Bioinform.* **2007**, *67*, 1010.
- [24] F.-A. Fortin, F.-M. De Rainville, M.-A. Gardner, M. Parizeau, C. Gagné, *J. Mach. Learn. Res.* **2012**, *13*, 2171.
- [25] E. F. Pettersen, T. D. Goddard, C. C. Huang, G. S. Couch, D. M. Greenblatt, E. C. Meng, T. E. Ferrin, *J. Comput. Chem.* **2004**, *25*, 1605.
- [26] P. Eastman, M. S. Friedrichs, J. D. Chodera, R. J. Radmer, C. M. Bruns, J. P. Ku, K. A. Beauchamp, T. J. Lane, L.-P. Wang, D. Shukla, T. Tye, M. Houston, T. Stich, C. Klein, M. R. Shirts, V. S. Pande, *J. Chem. Theory Comput.* **2013**, *9*, 461.
- [27] G. Neudert, G. Klebe, *J. Chem. Inf. Model.* **2011**, *51*, 2731.
- [28] D. Russell, K. Lasker, B. Webb, J. Velázquez-Muriel, E. Tjioe, D. Schneidman-Duhovny, B. Peterson, A. Sali, *PLoS Biol.* **2012**, *10*, e1001244.
- [29] A. Bakan, L. M. Meireles, I. Bahar, *Bioinformatics* **2011**, *27*, 1575.
- [30] H. M. Berman, *Nucleic Acids Res.* **2000**, *28*, 235.
- [31] O. Ben-Kiki, C. Evans, B. Ingerson, YAML Ain't Markup Language (YAML™) Version 1.2. Available at: <http://www.yaml.org/spec/1.2/spec.html>, accessed on January 23, **2017**.
- [32] J. Rodríguez-Guerra, **2017**. doi:10.5281/ZENODO.556352.
- [33] P. Vanhee, A. M. van der Sloot, E. Verschueren, L. Serrano, F. Rousseau, J. Schymkowitz, *Trends Biotechnol.* **2011**, *29*, 231.
- [34] Z. Wang, G. Wang, *Nucleic Acids Res.* **2004**, *32*, 590D.
- [35] W. A. Powell, C. M. Catranis, C. A. Maynard, *Mol. Plant. Microbe. Interact.* **1995**, *8*, 792.
- [36] G. Wang, X. Li, Z. Wang, *Nucleic Acids Res.* **2009**, *37*, D933.
- [37] S. V. Krivov, M. Karplus, *Proc. Natl. Acad. Sci. USA* **2004**, *101*, 14766.
- [38] M. Yemini, M. Reches, E. Gazit, J. Rishpon, **2005**, *77*, 5155.
- [39] J. Wang, E. Palecek, P. E. Nielsen, G. Rivas, X. Cai, H. Shiraishi, N. Dontha, D. Luo, P. A. M. Farias, *J. Am. Chem. Soc.* **1996**, *118*, 7667.

- [40] K. Lindorff-Larsen, S. Piana, K. Palmo, P. Maragakis, J. L. Klepeis, R. O. Dror, D. E. Shaw, *Proteins* **2010**, 78, 1950.
- [41] S. Zirah, S. A. Kozin, A. K. Mazur, A. Blond, M. Cheminant, I. Ségalas-Milazzo, P. Debey, S. Rebuffat, *J. Biol. Chem.* **2006**, 281, 2151.
- [42] CID=34231, **2017**. Available at: <https://pubchem.ncbi.nlm.nih.gov/compound/enterobactin>, accessed on January 23.
- [43] N. Li, L. Gu, To be Publ. n.d. doi:10.2210/PDB3TLK/PDB.
- [44] O. Trott, A. J. Olson, *J. Comput. Chem.* **2010**, 31, 455.
- [45] B. Kramer, M. Rarey, T. Lengauer, *Proteins Struct. Funct. Genet.* **1999**, 37, 228.
- [46] T. Cheng, X. Li, Y. Li, Z. Liu, R. Wang, *J. Chem. Inf. Model.* **2009**, 49, 1079.
- [47] R. Wang, Y. Lu, S. Wang, *J. Med. Chem.* **2003**, 46, 2287.
- [48] H. Fan, D. Schneidman-Duhovny, J. J. Irwin, G. Dong, B. K. Shoichet, A. Sali, *J. Chem. Inf. Model.* **2011**, 51, 3078.
- [49] J. W. M. Nissink, C. Murray, M. Hartshorn, M. L. Verdonk, J. C. Cole, R. Taylor, *Proteins* **2002**, 49, 457.
- [50] D. A. Oren, A. Zhang, H. Nesvadba, B. Rosenwirth, E. Arnold, *J. Mol. Biol.* **1996**, 259, 120.
- [51] J. M. van der Laan, H. A. Schreuder, M. B. Swarte, R. K. Wierenga, K. H. Kalk, W. G. Hol, J. Drenth, *Biochemistry* **1989**, 28, 7199.
- [52] B. H. Oh, G. F. Ames, S. H. Kim, *J. Biol. Chem.* **1994**, 269, 26323.
- [53] R. T. McGibbon, K. A. Beauchamp, M. P. Harrigan, C. Klein, J. M. Swails, C. X. Hernández, C. R. Schwantes, L.-P. Wang, T. J. Lane, V. S. Pande, *Biophys. J.* **2015**, 109, 1528.
- [54] H. Nguyen, D. R. Roe, J. Swails, D. A. Case, **2016**. doi:10.5281/ZENODO.44612.
- [55] S. K. Lam, A. Pitrou, S. Seibert, In Proceedings of Second Work. LLVM Compil. Infrastruct. HPC - LLVM '15, ACM Press: New York, New York, **2015**; pp 1–6.

Received: 2 February 2017
Revised: 27 April 2017
Accepted: 10 May 2017
Published online on 12 June 2017