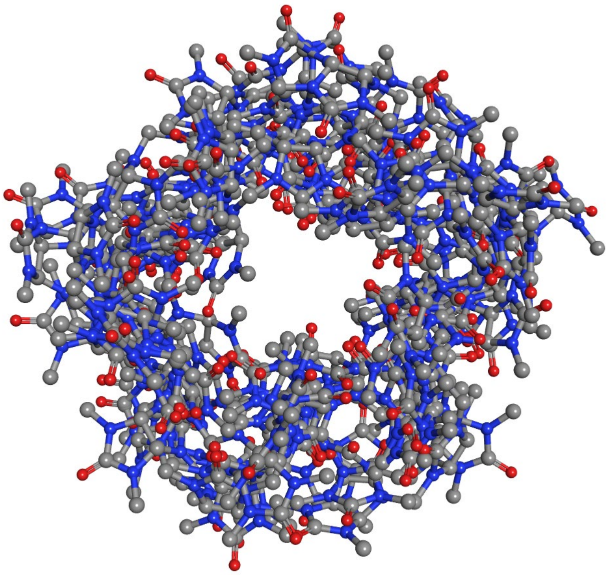
Development and Validation of Algorithms for the Generation of Conformer Ensembles Representing Protein-Bound Ligand Conformations



Cumulative Dissertation
with the aim of achieving the degree

Dr. rer. nat.

at the Faculty of Mathematics, Informatics and Natural Sciences
Department of Informatics
of Universität Hamburg

submitted by

Nils-Ole Friedrich
born in Hanau

Hamburg, June 2020

1st Reviewer: Prof. Dr. Johannes Kirchmair
2nd Reviewer: Prof. Dr. Peter Kolb

Date of thesis defense: 27.10.2020

The title page shows an ensemble of eight superposed conformations of the macrocyclic compound tetraicosamethylbambus[12]uril (PubChem ID 53233270), generated with Conformator (the algorithm developed for this thesis). The figure was generated with MOE;[41] hydrogens are not depicted.

Acknowledgements

I would like to sincerely thank my supervisor, Prof. Dr. Johannes Kirchmair, for making this dissertation possible, for his guidance, patience, motivation and advice. My thanks also go to him for the opportunity to present my work at national and international conferences.

I also thank Prof. Dr. Matthias Rarey for his interesting lectures, his commitment to his students and for introducing me to the main topic of my research, as well as his continued expertise and interesting discussions during my thesis.

Special thanks go to my current and former colleagues in the ACM team, especially Christina de Bruyn Kops and Conrad Stork for the comfortable, casual atmosphere in our office, for proofreading and all the helpful comments and invitations. I am grateful to have had the opportunity to work in such a pleasant and productive environment.

I would like to thank the entire ZBH Algorithmic Molecular Design working group for kindly providing the NAOMI software library, continuously improving it and creating a pleasant working atmosphere. I would particularly like to thank Kai Sommer for the introduction at the beginning of my master thesis and for his always available advice. Special thanks are due to Florian Flachsenberg for his helpful knowledge and support during software development, as well as our successful cooperation in the course of my doctorate. I thank Dr. Christin Schärfer for helping me get started from Cambridge and for providing CONFECT. Many thanks also to Jörn Adomeit and Gerd Embruch for the technical support, the productive discussions and the many entertaining lunch breaks. I would like to thank Melanie Geringhoff for all her support in bureaucratic matters.

Last but definitely not least, my deepest gratitude goes to my family for their love and unconditional support throughout my life and my studies. Special thanks to my wife, Christina, for her love and all the late nights and for keeping me sane over the past few years.

Kurzfassung

Die systematische Suche nach neuen Wirkstoffen für die Medizin ist teuer und zeitaufwändig. In zunehmendem Maße wird sie durch computergestütztes Wirkstoffdesign unterstützt. Anwendungen wie Docking, Pharmakophormodelle, die Suche in 3D-Datenbanken und die Erstellung von 3D-QSAR-Modellen sind dabei auf Ensembles von Konformationen angewiesen, um die Flexibilität kleiner Moleküle angemessen zu berücksichtigen. Bei der Konformations-Ensemble-Generierung handelt es sich um ein komplexes Problem, da die Anzahl der Freiheitsgrade selbst bei kleinen Molekülen sehr groß sein kann. Aufgrund seiner zentralen Bedeutung für das computergestützte Wirkstoffdesign ist die Konformations-Ensemble-Generierung seit mehr als drei Jahrzehnten Gegenstand intensiver Forschung. Im Lauf der Jahre wurden viele Algorithmen zur Konformations-Ensemble-Generierung entwickelt, die Qualität und der Umfang der Benchmarking-Datensätze für ihre Validierung nahmen allerdings nur langsam zu.

In der hier präsentierten Arbeit wird eine Methode für die vollautomatische Zusammenstellung großer, qualitativ hochwertiger Datensätze vorgestellt. Sie bewertet die Gültigkeit und Genauigkeit der 3D-Strukturen kleiner Moleküle anhand einer Reihe von Kriterien, einschließlich ihrer physikochemischen und strukturellen Eigenschaften aber besonders daran, inwieweit sie der experimentell bestimmten Elektronendichte gerecht werden. Die Methode wurde verwendet, um proteingebundene Konformationen von Liganden in über 350.000 per Röntgenkristallstrukturanalyse generierten Proteinstrukturmodellen aus der größten öffentlich zugänglichen Proteindatenbank zu filtern und zu bewerten. Das Ergebnis war der Sperrylite-Datensatz aus 10.936 hochwertigen Strukturen von 4.548 verschiedenen Molekülen. Er wurde verwendet, um die Variabilität der bioaktiven Konformationen kleiner Moleküle näher zu analysieren. Der Platinum-Datensatz enthält für jedes der 4.548 Moleküle im Sperrylite-Datensatz die Konformation aus dem bestbewerteten Proteinstrukturmodell. Das “Platinum Diverse Dataset” wiederum besteht aus 2.859 Strukturen von Konformationen verschiedenartiger Moleküle aus dem Platinum-Datensatz. Neben seiner hohen Qualität und bemerkenswerten Größe, verhindert es so systematische Fehler, die durch die Häufung einiger weniger Arten von Molekülen auftreten könnten. Alle drei Datensätze sind öffentlich frei verfügbar. Das Platinum Diverse Dataset ist dabei der erste Datensatz dieser Art, der qualitativ hochwertig und gleichzeitig groß genug ist, um Benchmark-Studien mit Algorithmen zur Konformations-Ensemble-Generierung durchzuführen, die relevante Aussagen zur statistischen Signifikanz von Unterschieden in der Leistung der verschiedenen Algorithmen ermöglichen. Im Rahmen dieser Arbeit wurde dieser Datensatz verwendet, um die bislang umfassendste Benchmark-Studie zur Konformations-Ensemble-Generierung durchzuführen. Die Leistung von sieben frei verfügbaren und acht kommerziellen Algorithmen wurde miteinander verglichen. Die Analyse zeigte, dass kommerzielle Algorithmen eine signifikant höhere Genauigkeit erzielten. Außerdem wurde deutlich, dass kommerzielle Algorithmen deutlich weniger geometrische Fehler in Molekülen produzierten.

Die in dieser Arbeit gewonnenen Erkenntnisse und Erfahrungen wurden für die Entwicklung von Conformator verwendet, einem neuen wissensbasierten Algorithmus zur Erzeugung von Ensembles von Konformationen. Conformator ist für nichtkommerzielle Zwecke und zum Einsatz in der akademischen Forschung frei verfügbar. Die von Conformator generierten Ensembles erreichen eine signifikant höhere Genauigkeit als die aller getesteten frei verfügbaren Algorithmen und es konnte kein signifikanter Unterschied zum leistungsstärksten kommerziellen Algorithmus festgestellt werden. Conformator ist in der Lage, in kurzer Zeit fehlerfreie Ensembles von Konformationen mit sehr hoher Genauigkeit für kleine Moleküle und Makrozyklen zu produzieren und schließt damit die Lücke zwischen kommerziellen und frei verfügbaren Algorithmen.

Abstract

The systematic search for new drugs is expensive and time-consuming. The process of drug discovery is more and more supported by computer aided drug design. Applications such as docking, pharmacophore search, 3D database searching and the creation of 3D-QSAR models are dependent on conformational ensembles to adequately represent the flexibility of small molecules. The generation of conformational ensembles is a complex problem because of the high number of degrees of freedom, even in small molecules. Because of its importance to the field, conformer ensemble generation has been the subject of intensive research for more than three decades. While there have been many intriguing ideas for algorithms for conformer ensemble generation, the quality and size of the benchmarking datasets to test their validity have improved very slowly.

To compile a large dataset of high-quality protein-bound ligand conformations from X-ray structural data, a fully automated cheminformatics pipeline for their selection and extraction was developed during this thesis. The pipeline evaluates the validity and accuracy of the 3D structures of small molecules according to multiple criteria, including their physicochemical and structural properties and, most importantly, their fit to the experimentally determined electron density. Extracted from a total of over 350,000 structures of co-crystallized ligands stored in the Protein Data Bank, the resulting Sperrylite and Platinum datasets are the largest publicly available datasets of such high quality. The Sperrylite Dataset consists of 10,936 high-quality structures of 4,548 unique ligands. It was utilized to assess the variability of the bioactive conformations of small molecules. The Platinum Dataset contains the 4,548 unique protein-bound ligands with the smallest diffraction-component precision index in the Sperrylite Dataset. The Platinum Diverse Dataset is a diversified subset of the Platinum Dataset of 2,859 compounds. In addition to its high quality and remarkable size, the Platinum Diverse Dataset is unbiased, diverse, and easily updatable. The Platinum Diverse Dataset is the first publicly available dataset from X-ray structural data in the Protein Data Bank of adequately high quality and sufficient size for thorough benchmark studies of conformer ensemble generators, which allow statements on the statistical significance of differences in performance between algorithms. In the course of this thesis, the Platinum Diverse Dataset was utilized to conduct the most comprehensive benchmark study of conformer ensemble generators to date. The performance of seven freely available and eight commercial conformer ensemble generators were compared to each other. The tests showed that commercial algorithms generally obtain higher accuracy and robustness with respect to input formats and molecular geometries.

The findings and experience gained during the benchmarking studies and the analysis of the variability of bioactive conformations was used for the development of Conformator, a new knowledge-based algorithm for generating conformer ensembles. Conformator is freely available for noncommercial use and academic research. The conformer ensembles generated by Conformator are significantly more accurate than

those of all free tools tested, and there is no significant difference to the best performing commercial algorithm. It could be demonstrated that Conformator, with its high accuracy and speed, as well as its robustness with respect to input formats, molecular geometries, and its handling of macrocycles, effectively closes the gap between commercial and freely available algorithms.

Contents

1 Introduction 1

1.1 Conformer Generation - Relevance to Drug Discovery 6

1.2 State of the Art in Conformer Ensemble Generation 7

1.3 Conflict of Objectives 13

1.4 Definitions 14

2 Development of a Method for the Automated Generation of High-Quality Benchmark Datasets 15

2.1 Sources for Data on Small Molecule Conformations 15

2.2 Evaluation Studies 17

2.3 Sperrylite and Platinum Datasets 23

2.3.1 Dataset Compilation 23

2.3.2 Analysis of the Sperrylite and Platinum Datasets 25

2.3.3 Usage and Analysis of the Platinum Datasets by Others 28

3 Development of a Novel Method for the Generation of Conformer Ensembles 31

3.1 Conformator 31

3.2 Benchmarking Conformer Ensemble Generators 34

3.2.1 Measures of Similarity 35

3.2.2 Validation Tool 36

3.2.3 Conformer Ensemble Generation 37

3.2.4 Hardware Setup 38

4 Statistical Analysis and Additional Methods 39

5 Conclusions and Further Directions 41

5.1 Improvement of Benchmark Datasets 41

5.2 Future Development of Conformator 43

References 50

Bibliography of this Dissertation’s Publications 70

Publications Related to This Cumulative Thesis 70

Further Authored Publications 70

Abbreviations 71

Appendix A Publication and congress contributions 74

A.1 Contributions to Publications of the Cumulative Dissertation 74

A.2 Contributions to Further Publications 76

A.3 Conference Contributions 79

Appendix B Software Architecture and Application 79

B.1 Libraries and Functions of Conformator 80

B.2 Conformator User Guide 82

B.3 Conformer Generation with RDKit 84
B.4 Conformer Generation with MOE 85

Appendix C Journal Articles 90
C.1 Published Journal Articles 90
C.2 Supporting Information of Published Journal Articles 144

Structure of This Thesis

This thesis gives an introduction to two closely linked topics in the field of drug discovery and cheminformatics, it describes the flexibility of small molecules in general and conformer ensemble generation, as a way to represent this flexibility, in particular. After the concepts of molecular flexibility and conformer generation are introduced, the relevance to the field is outlined. The thesis then goes on to describe the state of the art in conformer ensemble generation, giving an overview of the wide variety in approaches and popular algorithms. The first part of the thesis is concluded with a brief outline of the conflict of objectives in conformer ensemble generation.

The aim of this thesis was to analyze the flexibility of protein-bound conformations of small molecules and their representation by conformer ensemble generators and make progress in this field, if necessary and possible. The performance of conformer ensemble generators is commonly evaluated by comparing their ability to reproduce experimentally determined ligand conformations. The second part of this thesis therefore describes sources for data on small molecule conformations and goes into detail about evaluation studies on conformer ensemble generators that were done in the past. The latter reveals a clear lack of quality and size of the datasets used for the evaluation studies. In most cases the datasets were too small to obtain results of any statistical value.

In order to remedy this situation and to evaluate the state of the art in conformer ensemble generation in a manner that is statistically accurate and meaningful, a new large dataset of high-quality structures of protein-bound ligand conformations had to be compiled. To this end, a fully automated cheminformatics pipeline was developed that automatically evaluates the validity and accuracy of the 3D structures of small molecules according to numerous criteria, including their fit to the electron density. The cheminformatics pipeline is described in detail in ref D1. With it more than 350,000 crystal structures from the PDB were filtered, resulting in the Sperrylite and Platinum datasets. The Sperrylite Dataset is a complete collection of 10,936 high-quality conformations of protein-bound ligands (with up to 16 rotatable bonds) in the PDB. The Platinum Dataset consists of the 4,548 unique protein-bound ligands in the Sperrylite Dataset. A diversified subset of 2,859 structures, the Platinum Diverse Dataset, is by far the most suitable publicly available dataset for benchmarking conformer ensemble generators to date. The evaluation studies conducted with the Platinum Diverse Dataset, presented in ref D1 and D2, constitute the most comprehensive benchmark study of conformer ensemble generators so far. The performance of seven freely available and eight commercial conformer ensemble generators were compared to each other, including different ensemble sizes and, in some cases, various force fields. The tests showed significant differences in the performance of the tested algorithms and revealed that the commercial conformer ensemble generators generally obtain higher accuracy and robustness with respect to input formats and molecular geometries.

Based on the Sperrylite Dataset the variability of the bioactive conformations of 91 small molecules, each represented by a minimum of ten structures, was analyzed. Surprisingly, the variability was found to be largely independent of the number of rotatable bonds. A clear trend for the formation of few clusters of highly similar conformers was observed for a representative subset of 17 approved drugs and cofactors. Ligands were regularly found to adopt similar conformations, even when bound to vastly different proteins. The publication of this analysis, ref D3, also serves as a general overview and introduction to the topic of diversity of conformations of protein-bound ligands.

The knowledge gained during these studies was utilized to develop a new knowledge-based algorithm for generating conformer ensembles, named Conformator. It is freely available for non-commercial use and academic research and is described in detail in ref D4. The conformer ensembles generated by Conformator are significantly more accurate than those of all free tools tested. In fact, with the help of the Platinum Diverse Dataset it could be demonstrated that there is no significant difference in the accuracy of Conformator and the best performing commercial algorithm. Effectively closing the previously identified gap between commercial and freely available algorithms. Additionally, Conformator stands out with its speed, its robustness with respect to input formats, molecular geometries, and a novel algorithm for macrocycle conformer generation.

The present manuscript contains an appendix with reprints of the four publications that form the core of this thesis. Additionally, the appendix includes the supporting information of these publications, as well as a description of the architecture and application of the developed software. The bibliography is accompanied by statements of authorship.

1

Introduction

Many diseases which were life threatening half a century ago can be treated today, but countless diseases remain without a cure. Prominent examples include aging-associated diseases like cancer, autoimmune disease like multiple sclerosis and systemic lupus erythematosus (SLE) and re-emerging diseases that were thought to have been eradicated, like tuberculosis or malaria. An additional complication is the development of tolerance and resistance in bacteria, viruses and parasites against currently available drugs.[5,6] In view of the worldwide burden of diseases the discovery and development of new drugs remains absolutely necessary.

The systematic search for new drugs that are effective and safe, is expensive and time-consuming. Depending upon the therapy, the developing firm, the type of estimate and the expenditures included in the calculation (e.g. cost of failed developments) estimates for the cost of the development of a new drug range from US$ 0.5 billion to more than US$ 2.5 billion.[7–11] Classical approaches in drug discovery depend on screening large compound collections to identify potential candidates and their stepwise synthesis. Increasingly the drug discovery process is facilitated or sped up by computer-based techniques.[7,12,13] Only the most promising compounds found in computational experiments are tested in vitro or in vivo and, if successful, progress to clinical trials.

The macromolecule inherently associated with a specific disease process is almost always a protein. When a drug exists or is sought that interacts with it, the macromolecule is termed a drug target. Human drug targets mainly belong to four protein families: receptors, enzymes, transporters and ion channels.[14–17] Hence, the interactions between proteins and small molecules is of great scientific and medical interest. It is a central issue in drug design. Small molecules that can bind to macromolecules and form complexes with them are called ligands. In most cases a ligand is sought that specifically binds to a protein of interest. The area of interaction is usually a groove in the protein and is called binding pocket. The binding pocket where catalysis occurs in an enzyme is its active site. Binding a ligand can change the three-dimensional structure of the protein, or block the active site, thereby affecting its function. However, it

is especially the smaller and more flexible ligand that changes its three-dimensional structure in order to bind a macromolecule.[18]
The binding of a small ligand and a protein depends on the compatibility of their shapes (at the binding pocket), their specific interactions and solvent effects.[19–21] They usually interact via hydrogen bonds, ionic bonds, hydrophobic effects and van der Waals (vdW) forces. Covalent bonding between a ligand and a protein is also observed but is left out in many studies and programs because it is complicated by the reaction between the ligand and the receptor. Which interactions can take place between ligand and protein is determined by the atoms forming the binding pocket and the atomic makeup of the ligand and to a considerable degree by the flexibility of both molecules.

The conformation of a molecule is the arrangement of its atoms in three-dimensional space. Multiple conformations of the same molecule can be converted into each other by rotations about single bonds. The angle that exists in a chain of four atoms (A-B-C-D) between the plane passing through the first three atoms (A-B-C) and the plane passing through the last three atoms (B-C-D) is called torsion angle or dihedral angle (Figure 1). It can take values between 180° and -180°.

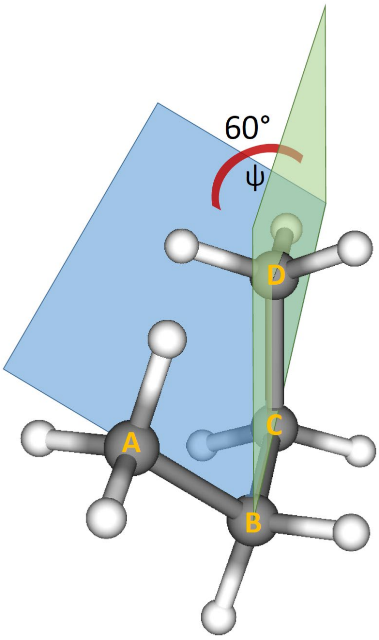


Figure 1: 3D model of a conformation of n-butane, the torsion angle ψ is defined by 4 consecutive covalently bound atoms (A-B-C-D) between the planes spanned by the first three atoms (A-B-C) and the last three atoms (B-C-D). Figure was generated using TorsionAnalyzer.[22]

Although the rotations about a single bond are often called “free“, they have to overcome an energy barrier between different conformations and there is usually an energy difference between diverse conformations. Conformations can therefore be regarded as points on a continuous energy landscape. For example, if the rotations of the terminal methyl groups are ignored (as is common practice), the relative energies of the simple molecule n-butane are a function of the central torsion angle. The global energy minimum (anti conformation), a local energy minimum (gauche conformation) and a transition state between the two, an eclipsed conformation that represents a local energy maximum, are depicted in Figure 2. All possible conformations of a molecule form its conformational space. Conformations that correspond to local minima on the potential energy surface are called conformers**.**[23]

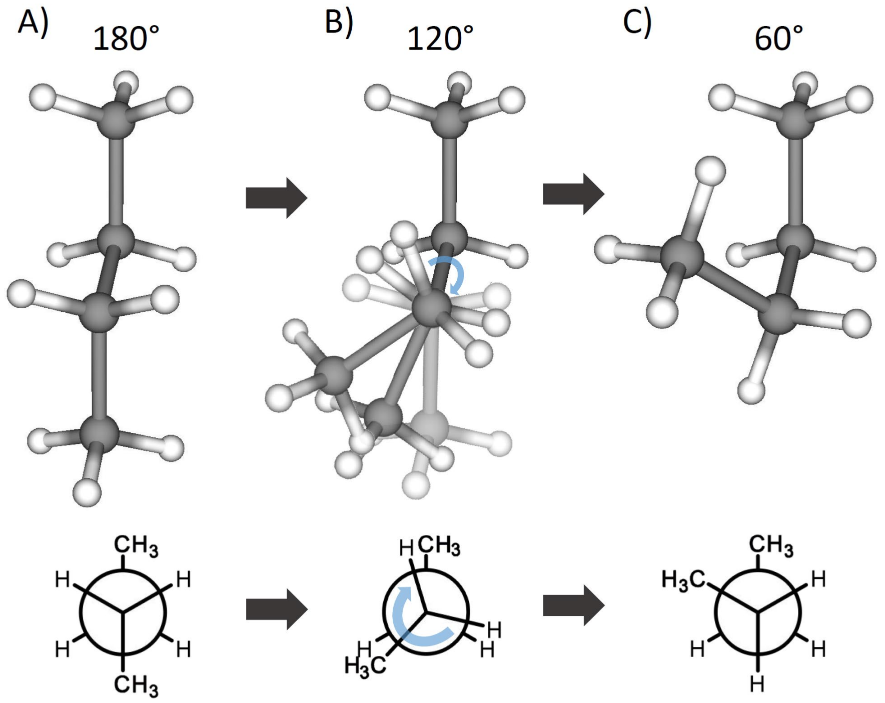


Figure 2: 3D models (top, with the torsion angle ψ) and Newman projections (bottom) of n-butane conformations. The anti (A, 180°) and gauche (C, 60°) conformations are connected by a rotation about the central single bond, passing the eclipsed conformation (B, 120°). The anti conformation is the lowest energy conformation of n-butane, since here the steric repulsion of the methyl groups is minimized. While there is significant steric repulsion between the two methyl groups in the gauche conformation, it is still lower in energy than the eclipsed conformation.[24,25] An energy maximum is reached at a torsion angle of 0° when the methyl groups are in an eclipsed position (cis conformation, not depicted). Figures were generated using TorsionAnalyzer.[22]

Rings and steric hindrances can restrict the rotation about single bonds. Steric hindrances occur when more bulky groups limit rotation. Rings change conformations by coupled rotations about several single bonds. A process often involving minor deformations of bond angles. For cyclohexane, the most stable conformation is called chair conformation, because of its folded shape. A “ring flip” inverts the ring and rapidly converts it to a different chair conformation. In this process it passes through different, less stable conformations of higher energy.[26]

Molecules indeed have a preferred conformation, but it cannot be assumed that this conformation is constantly present, as was first shown by D. H. R. Barton in 1950.[27] The conformation of a molecule constantly changes within certain limits as long as the energy barrier between the different conformations is small. This applies both for the conformation of proteins as well as that of ligands. The protein-ligand complex resulting from the binding process is a dynamic system and is subject to constant small or large changes. On the one hand two conformations of a molecule may differ by only a small rotation about a single bond, on the other hand transporter proteins can undergo large conformational changes to transfer molecules across a membrane.[28,29] Hence, conformational changes range from a fraction of Å to nm and happen in time frames of ns to s. Apart from the surrounding molecules, they can be influenced by temperature, pH, light and many other factors.[30] The ligand is usually very flexible in solution and is upon binding often forced into a strained (energetically unusual) conformation.[31,32] The amount of distortion of a molecule from its relaxed state is called strain energy.

The number of theoretically possible conformations of a given molecule is a function of the number of rotatable bonds, flexible rings, and atom angles. To a certain extent it also depends on the specific types of atoms and their arrangement (e.g. atom angles and intramolecular steric hindrances). Another factor is the minimum measurable difference or the (arbitrary) cutoff where two conformations are considered different. The theoretical number of spatial states of a molecule with at least one rotatable bond is therefore almost infinite.[33] However, most of the theoretically possible conformations of a molecule are energetically unfavorable, are only occupied for a very short time and are generally not observed experimentally.[34] In practice, for computational chemistry only realistically measurable distances are of interest. Other limitations are useful, e.g. restricting rotations to a few energetically very favorable torsion angles. Even then, the number of conformations to be considered is relatively large because their number rapidly increases with the number of rotatable bonds, this is termed a combinatorial explosion. For example, if only three angles per rotatable bond are allowed for a molecule with ten rotatable bonds 59,049 conformations are possible, allowing five angles per bond already leads to 9,765,625 theoretically possible conformations (without taking steric hindrance into account). The conformation of a small molecule bound to a protein often differs from the conformations found in solution, gas phase or small-molecule crystal structures.[31,35,36] In many cases it also does not correspond to the global energy minimum or even any local energy minimum.[18,31,35,37,38] Large collections of solid-state structures like the Protein Data Bank (PDB)[39] managed by the Research Collaboratory for Structural Bioinformatics (RCSB) and the Cambridge Structural Database (CSD)[40] offered by the Cambridge Crystallographic Data Center (CCDC), demonstrate that drug-like molecules can adopt a variety of conformations, even in the crystalline state.[39,40] An example of the conformational diversity of a single molecule found in crystal structures from the PDB is shown in Figure 3.

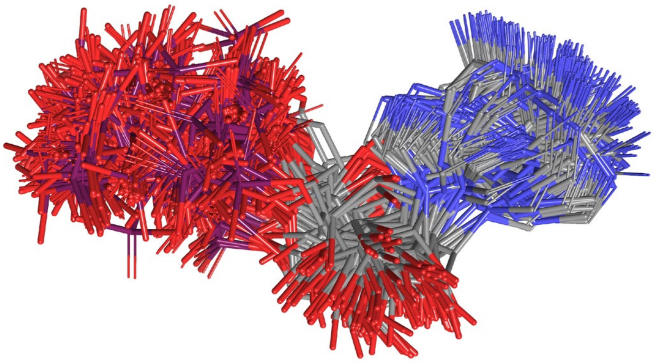


Figure 3: 218 superposed conformers of adenosine triphosphate (ATP) from crystal structures in the PDB. ATP is the most important molecule for intracellular transport and storage of chemical energy and takes part in many metabolic processes. As such it was cocrystallized quite often over the last decades. ATP is very flexible with eight rotatable bonds but is mostly found in elongated conformations. Figure was generated using MOE;[41] hydrogens are not depicted. Reprinted with permission from (Friedrich et al., 2018).[3] Copyright 2018 Frontiers in Chemistry.

Of all these conformations one is usually only interested in the bioactive conformation that binds to a specific protein. This conformation is often only stable while binding but is maintained within relatively narrow limits. The deciding factors in the binding of small molecules to proteins are interacting functional groups of both molecules, the displacement of water molecules, the reduction of the entropic degrees of freedom of both molecules and the mutual correspondence of their surfaces. Changes in the conformation of a molecule alter the accessibility of its functional groups and its surface. Therefore, different conformations of the same molecule can bind to different proteins and the flexibility of a molecule has a particularly strong influence on the number and variety of proteins it can bind to.[42] For example, the flexibility of an odorant affects by how many different olfactory receptors it can be detected.[43]

For many small molecules no experimental data on their bioactive conformation bound to a macromolecule of interest are available and therefore must be predicted. Conformer ensembles are collections of different conformers that are used to represent the flexibility of small molecules (Figure 4) and can be used to predict the protein-bound conformation. The prediction of protein-bound ligand conformations is of enormous importance to drug discovery, especially because the bioactive conformation can differ substantially from the low energy conformation that we observe in the gas phase or in solvent. Thus, conformer ensemble generation by algorithms is necessary for understanding the chemistry of small molecules and to infer possible biological roles of drug candidates, including their druggable targets, mode of action (MoA) and affinity, as well as off-target effects.[44,45]

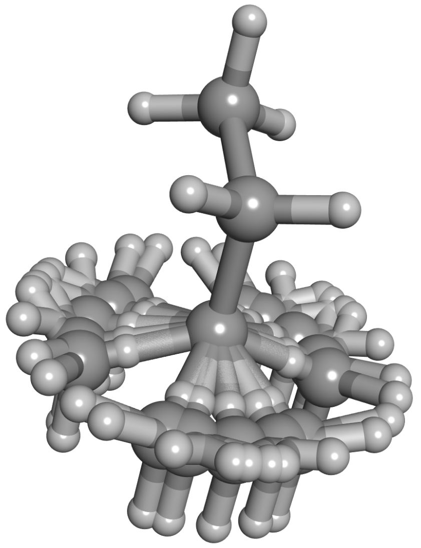


Figure 4: 3D model of a conformer ensemble of n-butane, consisting of 15 superposed conformers. Ensemble generated with Conformator,[4] the torsion angle ψ at 180°, 60° and 60° (“peak angles”), each with tolerances of 20° and 30° in both rotation directions, as defined by the torsion angle library.[46] Figure was generated using MOE.[41]

1.1 Conformer Generation - Relevance to Drug Discovery

Detailed and correct information of protein-bound ligand conformations is an essential precondition for many analyses in computational chemistry, that rely on the application of 3D computational approaches such as docking, 3D-QSAR (quantitative structure-activity relationship) and pharmacophore modeling, virtual screening or shape-based similarity searches.[34] Depending on the use case the necessary conformations can either be generated on-the-fly or calculated in advance and stored in databases.

Molecular docking algorithms attempt to predict protein-ligand complexes to discover potential ligands or gain further insight into known interactions. Relevant results of docking studies are the conformation of a ligand, the relative orientation in which it binds to the protein (binding mode) and intermolecular interactions (e.g. hydrogen bonds and hydrophobic contacts).[47,48] Molecular docking algorithms also address the challenging task of quantifying the binding affinity of the ligand, which is an important optimization parameter in drug design. Molecular docking has proven to be an effective instrument in drug discovery.[49] Changes in protein conformation upon ligand binding range from small local adjustments to large-scale rearrangements.[50,51] While ligand flexibility is usually taken into consideration by using conformer ensembles, protein flexibility is often ignored and remains challenging for docking algorithms.

Nevertheless, various attempts were made to model the flexibility of the protein while docking.[52–55] Docking is the fundamental technique in structure-based virtual screening. In a classical virtual screening protocol docking is employed to distinguish between biologically active and inactive compounds. Like for many computational approaches large ring structures (macrocycles) are difficult to address with molecular docking.[56]

QSAR models in cheminformatics aim to describe a correlation between a collection of molecular descriptors or physico-chemical properties and bioactivity of the compounds.[57,58] 3D-QSAR utilizes molecular descriptors generated from 3D molecular structures. One of its biggest challenges is that the bioactive conformation of a compound is usually unknown. Most proposed solutions add one or multiple dimensions to map variations in conformation, orientation, solvent effects or adaptation to a receptor.[59–61] It is not uncommon for QSAR models to be based on questionable chemical structures.[62] 3D- and 4D-QSAR applications may derive molecular descriptors from conformer ensembles of ligands or protein-ligand complexes for creating more robust models.[61,63,64]

Pharmacophore models describe the requirements that are necessary for a ligand to bind to a target protein in a particular mode. The ligand-receptor interaction can be represented by steric and electronic pharmacophores features that include information from the ligand and the amino acids surrounding it in the binding pocket.[65,66] Molecular shape or size can be taken into account through exclusion volumes. A pharmacophore may also only describe a small fragment of a molecule and different conformers of the same molecule may fit the pharmacophore restraints. To take flexibility into account pharmacophore models can be derived from conformer ensembles or be made flexible.[65,67]

3D virtual screening and shape-based similarity searches in molecular databases for target prediction and investigation of polypharmacology can include or support docking, 3D-QSAR and pharmacophores. They usually rely on conformer ensembles to sample the conformational space. Most of the time the main goal of 3D similarity searches is to significantly reduce the search space for consecutive experiments. Pure shape-based methods allow for so-called scaffold hopping, where compounds with different core structures but similar bioactivity are sought.

1.2 State of the Art in Conformer Ensemble Generation

Because of its importance to the field of computational drug discovery and cheminformatics many algorithms have been developed that sample the low energy conformational space of (small) drug-like molecules and compose conformer ensembles. Since

the 1970s, a variety of attempts have been made to use conformer ensembles to describe the flexibility of small molecules. However, there is still no consensus on the best strategy for conformer ensemble generation. Therefore, a manifold of methods is used today, including evolutionary algorithms, molecular dynamics simulations, geometric distance and knowledge-based approaches, as well as random and systematic searches. Furthermore, there exist mixtures of these approaches, e.g. it is common practice to evaluate or minimize randomly generated conformations with short molecular dynamics simulations.
For conformer ensemble generation the torsion angles in a molecule are in many cases considered to be independent of each other. This simplification allows programs an incremental buildup of conformations or the combination of conformations of parts of a molecule. As already mentioned, rings can also adopt different conformations, as they are flexible as well. However, their bonds cannot be considered independent of each other and rings are therefore usually treated as a unit.

There are two main categories of approaches for conformer ensemble generation: systematic and stochastic search algorithms. Systematic searches change torsion angles of all rotatable bonds by a set amount. Stochastic searches use random algorithms such as distance geometry, Monte Carlo simulations and genetic algorithms to sample torsion angles. There exist hybrid forms and both approaches principally face the same challenges. The main problem in conformer ensemble generation is that the conformational space grows roughly exponentially with increasing degrees of freedom. The number of degrees of freedom in a molecule depends mainly on the number of rotatable bonds and flexible rings and, to a lesser extent, on possible bond lengths and bond angles. These depend on the atom types involved and their chemical environment. There are simply too many potential conformations for any kind of real exhaustive search, especially for large molecules but in theory even for the smallest molecule with one rotatable bond an almost infinite number of conformations is possible. Possible solutions include reasonable minimum interconformer differences, restriction to low energy conformations and the rigid rotor approximation, where the molecule is considered rigid, with fixed bond lengths and bond angles. With few exceptions (e.g. in some acyclic bonds), this assumption is considered not harmful.[68,69] Due to these restrictions, it is possible to estimate the conformational space of a molecule and to sample it for conformer ensemble generation.[70,71]
Additionally, it is frequent practice in conformer generation to limit the application domain to covalent bonds and to limit the allowed atom types, e.g. exclude metals and neglect ionic bonding. Also, usually only non-covalently bound ligands are processed. Statistically derived data from databases (like PDB or CSD) can be utilized to determine the most common angles between different atom types. Another way to greatly reduce the search space, is the use of fixed precomputed parts of molecules that are

not especially flexible and occur often. An example for this practice are libraries of small ring conformations.

Another common theme is the clustering of generated conformations. To generate a conformer ensemble of small enough size, identical and similar conformations are identified and removed. Many different distance measures are used for clustering but by far the most common measure is the minimum heavy-atom root-mean-square deviation (RMSD). More often than not, clustering and force field minimization are the most time-consuming steps in conformer ensemble generation.

Quantum chemical (QC) computations can be used to explore large parts of the conformational and the reactional space.[72] Ab initio methods, like the Hartree-Fock method, are derived directly from theoretical principles of quantum mechanics (QM), without including experimental data.[73] Semiempirical QC methods, derived from either Hartree-Fock or density functional theory by applying systematic approximations, are faster and much more computationally efficient than ab initio calculations.[74] Density functional theory (DFT) can be applied as an ab initio or semiempirical method to investigate the ground state (lowest-energy state) of quantum-mechanical systems.[75] More traditional DFT approaches are known in the field of cheminformatics for their long runtimes, high computational cost and the very limited number of atoms they can handle, but more recent methods (e.g. GFN2-xTB) can handle up to 1000 atoms in a relatively short time.[76] These methods are usually applied to gain a deeper understanding of one or few molecules but can also be utilized for conformer ensemble generation.[77,78] In this case the conformer ensemble is compiled from significantly populated minimum energy structures.

Originally the generation of conformer ensembles took a long time and depending on the complexity of the molecule, the conformer generation method and the desired accuracy still results in significant runtimes. Therefore, many programs that use conformer ensembles require precalculated ensembles. Multiple modern conformer generation algorithms are fast enough that it is possible to deal with the flexibility of small molecules for some applications on the fly instead of storing conformer ensembles in a database.

As stated above, many algorithms have been developed for conformer ensemble generation. Brief descriptions of the algorithms used in the benchmark studies of this work can be found in ref D1 and D2. The following section and Table 1 give a short and general overview of available algorithms, for more detailed descriptions of the individual algorithms please refer to the associated publications.

All conformer ensemble generators use knowledge about atoms, bonds, and molecules to some degree. An algorithm is termed “knowledge-based” when it makes intensive use of experimentally or theoretically obtained knowledge of conformations. In most cases, experimentally derived knowledge is used to keep the number of conformations in the ensemble as low as possible without losing too much useful information. Either rules for the construction of conformations are defined or templates for parts of molecules are stored in libraries. Most commonly, the latter is used in the form of predefined libraries of torsional angles and ring conformations. Examples of knowledge-based conformer ensemble generators that utilize a torsion angle library and a library of ring conformations are the latest algorithm in RDKit[79] termed Experimental-Torsion basic Knowledge Distance Geometry (ETKDG)[80] and CONFECT[81], as well as the algorithm developed in this work, Conformator.[4] The torsion angle library stores typical torsion angles for specific patterns of rotatable bonds. The torsion libraries used in CONFECT and Conformator contain hundreds of rules derived from the CSD encoded as SMARTS patterns. For each bond defined by these patterns they provide the most frequently occurring torsion angles ("peaks") and the two most frequently occurring deviations ("tolerances") from the peak angles (cf. Figure 4). [22,46] Similar knowledge-based approaches can also be found in the programs ROTATE,[82] MIMUMBA,[83] iCon (Inte:Ligand)[84] and OMEGA (OpenEye).[85,86] The OMEGA algorithm generates energetically accessible combinations of molecular fragment templates and scores them with a modified version of the MMFF94s force field.[84,87] Most conformer ensemble generators are also to some extent graph-based, Frog2 is classified as graph-based, because it is particularly dependent on the decomposition of the molecular graph.[88,89]

Random search algorithms involve random changes to coordinates of the atoms or torsional angles. The resulting conformation is usually optimized, compared with the rest of the ensemble and kept if it is dissimilar enough. Both, the Molecular Operating Environment (MOE, Chemical Computing Group)[90] and MacroModel (Schrödinger)[91] are able to do a random search. Apart from the use of chemical knowledge, the most common approach to generate coordinates of conformations is distance geometry. Here, the description of a molecule consists of a list of distance and chirality constraints. These function as lower and upper bounds on the distances between pairs of atoms and the chirality of its rigid quadruples of atoms.[92,93] Distance geometry algorithms generate conformations from the distance bounds matrix of the molecule based on the connection table and a set of rules. They randomly generate distance matrices which satisfy these bounds and produce 3D coordinates from the resulting distances. This process is often called embedding. Different random distance matrices result in different conformers that form the ensemble. Examples of distance geometry approaches for conformer ensemble generation are DG-AMMOS[94,95] and the distance geometry algorithm in RDKit (RDKit DG).[79] A special case of distance geometry are "self-organizing" algorithms, like stochastic proximity embedding (SPE)[96] and self-organizing superimposition (SOS).[97] A detailed introduction to distance geometry and its principles for generating conformations can be found in ref 98.

Other methods for conformer generation, like molecular dynamics (MD) simulations or evolutionary algorithms are less frequently used. The LowModeMD method, one of the conformer ensemble generators in MOE and Frog2 use MD simulations for the generation of conformer ensembles. Evolutionary algorithms, also called genetic algorithms (GA), are based on the basic principles of biological evolution. They generate a random ensemble (population) of possible solutions, i.e. conformations, and evaluate them through a fitness function. The best rating conformations are changed randomly (mutation) or combined (recombination), while the rest is discarded (selection). Thus, the ensemble changes over time and develops to better solutions. Approaches that apply evolutionary algorithms are e.g. Balloon[99] and Cyndi.[94]

Conformer ensemble generators that perform a systematic search assign each rotatable bond of a ligand an angle between 0° and 360° at regular distances. If this distance would be chosen small enough all possible conformations of a molecule could be created. However, this is generally not useful because too large ensembles are created, which cannot be generated or processed by downstream tools in any reasonable time frame. Examples of conformer ensemble generators that are able to perform a systematic search are Catalyst,[70,100] CAESAR (part of the Catalyst Component Collection)[101] and MOE.

Molecular mechanics (MM) calculations with classical force fields attempt to represent the potential energy of a molecule with simple functions. Many conformer ensemble generators minimize computed conformations with a force field to avoid steric clashes and high strain in the structures. The reliability of these force field methods for conformer generation remains a matter of debate.[80,102]

TABLE 1. Overview of some widely used conformer ensemble generators

	Systematic search	Distance geometry	Knowledge-based	MD simulation	Graph-based	Evolutionary
Balloon DG		✓				
Balloon GA		✓				✓
Catalyst	✓		✓			
Confab	✓		✓			
Confgen	✓		✓			
ConfgenX			✓			
Cxcalc			✓			
Cyndi						✓
DG-AMMOS		✓				
Frog2				✓	✓	
iCon	✓		✓			
MacroModel				✓		
MIMUMBA			✓			
MOE LowModeMD				✓		
MOE Stochastic		✓				
MS-DOCK	✓		✓			
Multiconf-Dock	✓					
OMEGA	✓		✓			
RDKit DG		✓				
RDKit ETKDG		✓	✓			
ROTATE			✓			
Rubicon		✓	✓			

In recent years, the exploration of macrocycles for drug discovery has emerged as one of the most actively pursued research fields in cheminformatics.[103–106] This is a relatively new development, because macrocycles with their structural complexity do not fit the usual paradigm of "drug-likeness" and are difficult to synthesize. Indeed, it is not unlikely for a macrocyclic compound to violate most of the criteria of the famous Lipinski's rule of five or similar systems.[107,108]

The most common source of macrocycles are natural products, in fact the more than 100 marketed macrocycle drugs are almost exclusively derived from natural products.[103] It could be shown that a large number of natural products are readily obtainable and that they are highly diverse and populate regions of chemical space that are of high relevance to drug discovery.[109] Marketed macrocyclic drugs include e.g. the gastrointestinal prokinetic agent ulimorelin (TZP-101)[110] and the antibiotic murepavadin (POL7080)[111] (Figure 5). Macrocyclic drugs often function differently than small molecule drugs and can interact with target proteins that are highly challenging for smaller molecules. Macrocycles interact with a broad spectrum of targets, including ATPases, kinases, GPCRs and proteases.[105]

There is no universally accepted precise definition of macrocycles. Concepts differ in the number of atoms required and meaning either a cyclic molecule or a macromolecular cyclic part of a molecule. In this thesis and the accompanying publications macrocycles are defined as compounds including at least one ring formed by 10 or more atoms. Macrocycles are particularly hard to handle for conformer ensemble generators, since they are by definition large and often contain many coupled rotatable bonds, leading to a mix of high flexibility and various conformational restrictions. Because of this, most algorithms for conformer ensemble generation (and many other applications) will skip macrocycles completely.

As part of this thesis, a new knowledge-based conformer ensemble generator was developed, called Conformator. One of the major conceptual advancements of Conformator include a novel approach to sampling the conformational space of macrocycles. It also features a new clustering algorithm for the assembly of conformer ensembles and an extended set of rules for sampling torsion angles. Conformator further stands out by its robustness with respect to molecular geometries and input formats.

In the course of this thesis the most comprehensive benchmark study of conformer ensemble generators to date was conducted. The knowledge gained during these studies was taken into account during the development of Conformator. As a result, Conformator provides significantly higher accuracy in the reproduction of bioactive conformations from crystal structures than all freely available tools tested and is on par with the best performing commercial algorithm.

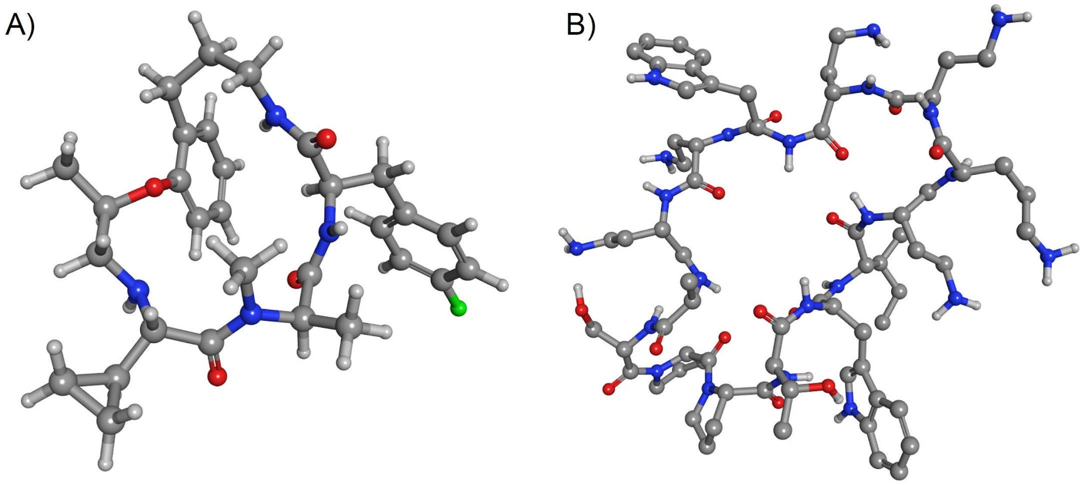


Figure 5: 3D models of two examples for macrocyclic drugs, the gastrointestinal prokinetic agent ulimorelin (A) and the antibiotic murepavadin (B). Figures were generated using MOE.[41]

1.3 Conflict of Objectives

The goal in the generation of conformer ensembles for computer-aided drug design is to quickly provide a small conformer ensemble (meaning a low number of structures), which widely covers the conformational space and accurately reproduces the protein-bound ligand conformations. This creates a conflict of objectives in conformer ensemble generation between accuracy, ensemble size and computing time. Accuracy of a conformer ensemble is usually measured as the minimum root-mean-square deviation (RMSD) in Å between the experimentally determined bioactive conformation and any computed conformers of an ensemble. The ensemble size is equal to the number of structures contained in the ensemble.

Different algorithms put varying priorities on each of these parameters, and thus the algorithm applied should depend on the specific use case. A highly sophisticated and computationally expensive algorithm that generates potentially large ensembles of high-quality conformers, should be selected if top priority is given to accuracy. Smaller ensembles, especially those containing fewer irrelevant conformations, are of high interest for most downstream applications. The runtime of many downstream applications grows exponentially with the number of conformers to process, because of this, many conformer ensemble generators reject too similar conformers or those of extremely high energy by default. However, smaller ensembles generally result in a loss of shape diversity and increase the chance to "miss" the desired bioactive conformer. If a loss of quality is acceptable, e.g. if a large number of molecules are to be screened, especially if they are to be screened repeatedly, a smaller ensemble size is recommended. When time is of the essence, computationally efficient algorithms are preferable, even though this choice can be accompanied by further loss of quality.

Theoretically, there exists a runtime-quality trade-off. With high runtimes ensembles of high accuracy and well-adjusted size could be generated. And this is true, to a certain extent, e.g. force field minimization costs time and without it the quality of the ensemble decreases. However, in reality and if we view the algorithms as a whole (including e.g. force field minimization and clustering), the accuracy and ensemble size are largely dependent on the algorithm used and no amount of additional runtime would better the quality of the conformer ensemble substantially. Examples of this can be found in comparisons done for this work. The performance of ConfGenX[112] was tested with the OPLS_2005 (optimized potentials for liquid simulations) and OPLS3 force fields[113] and the new cluster algorithm developed for Conformator was compared to the K-Medoids cluster algorithm.[2,4] Additionally the best performing algorithms in terms of accuracy and ensemble size were found to be among the fastest algorithms.

Another possible metric for the performance of a conformer ensemble generator is the performance of downstream applications that depend on their input. This however is problematic for multiple reasons. Cappel et al. found virtual screening results to be insensitive to the conformer ensemble generator used. As a consequence algorithms that generate smaller conformer ensembles might be preferred for virtual screening, but for 3D-QSAR modeling they found that models based on larger ensembles and with energy optimization performed better.[114] Others have found the opposite to be true for their 3D-QSAR models.[115] Overall this might indicate that these performance tests are (at least sometimes) dominated by special characteristics of the downstream application.

1.4 Definitions

Some terms that are widely used in the field of cheminformatics are often not clearly defined. In this thesis “high-quality structures” refers to any structures matching the quality criteria defined in ref D1. Unless stated otherwise, this term only refers to the quality of the protein-bound ligand, not the overall structure of the protein-ligand complex. Accuracy of a conformer ensemble was defined as the minimum RMSD in Å measured between the experimentally determined protein-bound conformation and any conformer of the computed ensemble. Intramolecular clashes are defined as overlaps of more than 30 % of the van der Waals radii of 1−4-connected (or more distant) heavy atom pairs that are not part of the same ring system.

As introduced above, macrocycles are defined as compounds that include at least one ring formed by 10 or more atoms. Three-letter codes in italics refer to PDB ligand identifiers and four-letter codes refer to PDB entries.

2

Development of a Method for the Automated Generation of High-Quality Benchmark Datasets

2.1 Sources for Data on Small Molecule Conformations

The conformation of small molecules can experimentally be determined in a protein-ligand complex or a small-molecule crystal structure. Two major methods provide three-dimensional structures of protein-ligand complexes close to or at atomic resolution, X-ray diffraction and nuclear magnetic resonance (NMR) spectroscopy. Further methods like cryogenic electron microscopy (cryo-EM) exist but are less frequently used. Small-molecule crystal structures are determined by X-ray and neutron diffraction analyses.

Three-dimensional X-ray crystal structures are a main source of our understanding of molecular interactions between proteins and ligands, as well as the relationship between structure and biological function overall. X-ray crystal structures of protein-ligand complexes have been produced since the late 1950s.[116] Here, a crystal of tightly packed molecules is rotated through an X-ray beam to obtain an X-ray diffraction pattern. The patterns mainly consist of arrays of spots at regular distances called reflections. From the obtained diffraction patterns, the amplitude, and the phase of the X-rays in each reflection an electron density map can be computed. The amplitudes are usually obtained by measuring the reflection intensities, the phases by isomorphous or molecular replacement. The electron density map is interpreted to generate a 3D model of the molecular structure.[117] It has to be emphasized that the X-ray diffraction experiment is the last experimental step when a 3D structure of a protein is determined. The electron density maps contain only approximate positions of the corre-

sponding atoms in the crystal. The quality of X-ray crystal structures ultimately depends on the quality and quantity of the underlying experimentally determined diffraction data.[118–120]
Crystals of protein-ligand complexes can be hard to obtain, since the crystallization process is complicated, the resulting crystals can have cracks, show unfavorable morphologies or simply be too small. Additionally the mechanisms behind organic crystallization are not well-understood.[121] Size and purity of a crystal are directly correlated with the quality of an X-ray structure determination.[122] Optimization of crystallization conditions is therefore the first and most important step on the way to obtaining a new high-quality crystal structure. The process is considered by many in the field a "form of art" that even after intensive training requires a certain amount of "luck". Many X-ray crystal structures today are cooled down to cryogenic temperatures to slow radiation damage during data collection, but it has been shown that this technique introduces bias to the conformational distribution of the protein and leads to smaller, over-packed models.[123]

Methods for determination of the structure and dynamics of small- to medium-sized organic molecules by NMR were developed since the late 1960s.[124] But the first publication of an NMR structure did not take place until 1985.[125] NMR is based on the resonant interaction between the magnetic moment of atomic nuclei. One of the main advantages of NMR spectroscopy compared to X-ray crystallography is that it does not necessarily require the protein in crystallized form. NMR provides structural information on the local conformation and distances between neighboring atoms but also on the dynamics and chemical kinetics of these systems at the atomic level.[126,127] NMR made it possible to determine molecular conformation in solution and to study conformational exchanges.[128,129]

Cryo-EM is often discussed as a promising alternative to X-ray crystallography, especially since the Nobel Prize in Chemistry was awarded to Jacques Dubochet, Joachim Frank and Richard Henderson "for developing cryo-electron microscopy for the high-resolution structure determination of biomolecules in solution" in 2017.[130] Cryo-EM can be used to determine the structure of protein-ligand complexes that are very hard to crystallize or not at all, including highly dynamical systems.[131] In X-ray crystallography the protein is usually forced into one conformation, cryo-EM allows for the mapping of molecules in different conformations. Already hundreds of cryo-EM structures have been published, but the main challenge for the method remains the validation of their quality.[132,133]

Neutron diffraction analysis applies neutron scattering to determine structures similar to X-ray diffraction. Both methods complement each other because of the different scattering properties of neutrons and X-rays, but only very few structures were solved by neutron diffraction so far.[134,135] X-ray crystallography is still by far the most used method for determining protein structures. The largest publicly available collection of protein structures, the PDB, contains over 130,000 X-ray crystal structures of proteins, over 11,000 NMR structures and more than 3,000 that were determined with cryo-EM (as of January 2020). The CSD is a large repository for small-molecule crystal structures and contains more than one million 3D structures from X-ray and neutron diffraction analyses.

2.2 Evaluation Studies

Due to the importance of conformer ensembles to the field of computer-aided drug design many algorithms for their generation have been developed. Most evaluation studies compared a new algorithm to one or two established algorithms, therefore there is a similar number of evaluation studies and algorithms. Usually conformer ensemble generators have been tested for their ability to reproduce experimentally determined bioactive conformations from crystal structures of protein-ligand complexes. In rare cases ensembles have been tested for their coverage of bioactive conformational space while keeping computational efficiency in mind.[80,136] In very few evaluation studies conformer ensembles were systematically checked for geometrical errors.
For more than 30 years conformer ensemble generators have been developed and compared.[137] Many studies only analyzed the conformational space of a single molecule of interest. Dataset size, i.e. the number of molecular structures in a dataset, increased over the years but generally small datasets were used to make statements on the statistical significance of differences in performance between algorithms. In most cases, the level of statistical significance of the results, the uncertainty of the 3D coordinates in the underlying data and, above all, the minimum number of data points required for accurate conclusions were ignored.

Evaluation studies of conformer ensemble generators originally used more or less randomly chosen CSD or PDB structures. Over time manual curation of datasets by experts became more common, to ensure some level of quality. Nearly all of these studies used different datasets and different algorithms for comparison, rendering it impossible to compare different studies directly. A list of datasets used for evaluation of conformer ensemble generators can be found in Table 2. For most datasets, the resolution of the X-ray structures served as an important or even exclusive quality criterion for selection of the molecules. However, this approach is not appropriate, since resolution is not a measure of the quality of a model but the quantity of the underlying data.[118,138] It also provides no information about the completeness, reproducibility or signal-to-noise ratio of the data on which it is based. Nevertheless, filtering crystal structures for resolution is useful, since it indicates the quantity of data gathered and only models with a resolution of at least 2.7 Å can have a ratio of experimental data points to parameters greater than 1.[85]

TABLE 2. Overview of datasets used for validation of conformer ensemble generators[a]

Dataset name	CSD	PDB	NofMC[b]	Year	Reference
Ghose et al.	76			1993	139
Boström		32		2001	35
Original GOLD (Nissink et al.)		134		2002	140
Diller and Merz		65		2002	141
Boström et al.		36		2003	142
Perola and Charifson		100		2004	31
Kirchmair et al.		510		2005	143
Izrailev et al.		68		2006	144
Kirchmair et al. (expanded set)		778	12	2006	145
Liu et al.		329		2006	94
Agrafiotis et al.		59		2006	146
Astex Diverse Set (Hartshorn et al.)		85		2007	147
Li et al.		918		2007	101
Vernalis (Chen and Foloppe)		256		2008	148
Bonnet et al.		19	19	2009	149
Bai et al.		742		2010	150
Hawkins et al.	480	197		2010	85
Ebejer et al.	469	239	9	2012	151
Iridium-HT (Warren et al.)		121	1	2012	152
Chen and Foloppe		333	30	2013	153
Shelley (Watts et al.)	83	67	150	2014	154
Riniker and Landrum	1,290	238	24	2015	80
Prime-MCS (Sindhikara et al.)	60	148	208	2017	155
Platinum Diverse (Friedrich et al.)		2,859	29	2017	1,2

[a]Most of the datasets are derived from PDB structures, some include structures from the CSD; the Prime-MCS dataset includes 18 structures downloaded from the Biologically Interesting Molecule Reference Dictionary (BIRD).[156] BIRD is a subset of the PDB and the structures were counted as such. The set by Perola and Charifson includes 50 structures from the publicly unavailable Vertex structure collection (not included in the table).[40]

[b]Number of macrocycles as defined by the associated publication, which is sometimes not identical to the definition used in this thesis (compounds including at least one ring formed by 10 or more atoms) and only appearing in the table when explicitly addressed by the authors or otherwise known.

In 1990, Saunders et al. analyzed conformer ensemble generators for their ability to generate large ring structures on cycloheptadecane.[137] They tested systematic and random search methods, as well as molecular dynamics and a distance geometry method. Rings were "frozen" in their original X-ray conformations during the conformational searches, to keep computation time at reasonable levels. The authors concluded that cycloheptadecane was lying close to the boundary of what could be adequately addressed with the methods and computational resources at that time.

Inspired by this and similar publications Ghose et al. were the first to acknowledge a general lack of a “standardized set of molecules” for the validation of conformer ensemble generators.[139] Their attempt to provide such a dataset consisted of 76 molecules from the CSD that had previously been used by other groups to evaluate the performance of force fields.[157,158] Seventy-two of these 76 molecules were used for the validation of the Sybyl search method by Judson et al.[159] From the already small set of molecules three were removed because they had no torsion angles outside of rings and one for lack of force field parameters. Jaeger et al. validated the algorithm termed conformational energy downward driver (CEDD) on 74 molecules from the dataset by Ghose et al.[160] For their method to consider rings in the conformations these had to be supplied as starting points. Shortly after, the group compared the performance of Sybyl and MacroModel[91] on the same dataset.[161] Again ring conformations were omitted from the search in conformer generation.

The first evaluation of multiple conformer ensemble generators in the 21st century was conducted by J. Boström in 2001, comparing Catalyst, Confort,[162] Flo99,[163] MacroModel and OMEGA for their ability to reproduce bioactive conformations on 32 protein-bound ligand structures.[35] For the selection of structures a cutoff for resolution of 2.0 Å was applied as the primary quality criterion. The performance of OMEGA was tested on a set of 36 ligands in a follow-up study by Boström et al. two years later.[142]

Diller and Merz investigated 65 protein-ligand complexes from the PDB to find 3D-descriptors that separate random conformations from active conformations. They found bioactive conformations trending towards more extended conformations than randomly generated ones. Extended conformations have more solvent accessible, surface area and fewer internal interactions.[141]

Perola and Charifson studied conformational changes of drug-like molecules upon binding to proteins and compared the performance of a Monte Carlo conformational search published by Abagyan and Totrov[164] to that of Catalyst and MacroModel.[31] They highlighted that the usefulness of studies is limited by the size and composition of the datasets used. Their dataset consisted of 150 compounds from protein-ligand complexes with known binding affinities and a focus on pharmaceutically relevant structures. Perola and Charifson selected 100 structures with available binding constants from the PDB with a resolution of less than 3 Å and 50 structures from the Vertex structure collection. They concluded qualitatively that ligands tend to bind in an extended conformation, even when a folded conformation is more stable in solution. The authors discuss this characteristic as a potential criterion for the assessment of the biological relevance of generated conformations.

Izrailev et al. tested a simple heuristic to bias conformational sampling toward more extended or more compact conformations on a slightly updated version of the dataset by Diller and Merz, now containing 68 structures. They concluded that the heuristic significantly improved the chances of finding bioactive conformations.[144]

In 2005 Kirchmair et al. published the first study of a conformer ensemble generator (Catalyst) with a benchmark dataset of several hundred protein-bound ligand conformations extracted from the PDB.[143] For a comparison of OMEGA (version 2.0 at the time) and Catalyst Kirchmair et al. manually expanded the dataset to a size of 778 drug molecules and pharmacologically relevant structures.[145] For both datasets the resolution of the crystal structure was used as the primary quality criterion.

The conformer ensemble generator CAESAR, that is part of the Catalyst Component Collection, was validated and compared to different settings of Catalyst on a related dataset containing 918 molecules extracted from the PDB.[101] A subset of this dataset (742 structures) was later used to compare multiple molecular force fields for the evolutionary algorithm Cyndi.[150] Cyndi was also compared to Balloon (version 0.6.6.4641) and Catalyst at different settings on a combined dataset of a subset of the Astex dataset and the subset used by Izrailev et al. containing 329 structures extracted from the PDB.[94]

Chen and Foloppe compiled the Vernalis test set of 256 chemically diverse drug-like ligands, including the 32 structures from the original Boström set (resolution < 2 Å), 94 of the 100 publicly available structures from the set of Perola and Charifson (resolution < 3 Å) that are not in the first set and 130 additional structures extracted from the PDB (resolution < 2.5 Å).[148] The complete set was utilized to compare three algorithms from MOE, Systematic Search, Stochastic Search, and Conformation Import to Catalyst. They analyzed the coverage and diversity of the conformational space covered by the ensembles with pharmacophores and concluded that both algorithms were well suited for their intended task and that MOE performed at least as well as Catalyst.

Chen and Foloppe later benchmarked different low-mode based approaches from MacroModel against Stochastic Search and LowModeMD from MOE.[153] They also investigated different force fields and different settings for some of the algorithms. This time the authors explored three different datasets of X-ray structures from the PDB: 253 drug-like ligands from the Vernalis test set, a set of 50 diverse and more flexible compounds with 12 to 20 rotatable bonds and 30 macrocycles (defined as a ring of at least 9 atoms) with 9 to 30 rotatable bonds in the cycle. Both the flexible and the macrocycle set were clustered and filtered with multiple criteria, including a maximum resolution of 2 Å. The authors concluded that compared to the default settings much better results can be obtained by adopting enhanced search parameters, regarding the energy window, the maximum ensemble size, and the maximum total number of iterations. For MOE they found much better performance with the generalized Born (GB) model[165] over the distance-dependent dielectric constant (Diel) in the treatment of solvation.

Bonnet et al. compared the performance of Catalyst, CAESAR, MacroModel, MOE, Omega, Rubicon and the two self-organizing algorithms SPE and SOS.[149] To investigate the effect of ring size on the performance of the different algorithms they compiled a dataset of 19 structures: eight cyclopeptides, five cyclodextrins and six naturally occurring macrocycles with known biological activity. They found the three distance geometry methods (SOS, SPE, and Rubicon) to be the most robust and universally applicable, and SOS to be preferable over SPE for its superior speed.

OMEGA was tested on a dataset of 480 druglike molecules from the CSD and 197 ligands from the PDB.[85] Notably Hawkins et al. gathered data from different datasets and filtered them with a set of quality criteria, including the real-space R-value (RSR),[166] the real-space correlation coefficient (RSCC),[167] the occupancy-weighted B-factor (OWAB) and the diffraction-component precision index (DPI; Goto). [168,169] The dataset curation approach was later refined by Warren et al. and used to select the 121 high-quality structures for the dataset Iridium-HT.[152] Both publications inspired the development of the automated dataset generation process in this thesis.

Ebejer et al. examined the performance of the four freely available algorithms Balloon, Confab, Frog2, and RDKit DG against MOE on a dataset of 708 drug-like molecules.[151] 469 structures were taken from the CSD based on the work of Hawkins et al. and 239 structures from the PDB, with 85 of those being from the Astex Diverse Set.[147] The authors found an overall trend of increasing RMSD with an increasing number of rotatable bonds, and that RDKit DG and Confab performed statistically better than the other methods. They also found Confab to be more suitable for molecules with a large number of rotatable bonds.

Riniker and Landrum used two datasets to test how well conformations from crystal structures can be reproduced by generated ensembles of the older purely distance geometry method (RDKit DG) and ETKDG, using RMSD and TFD. Additionally, they examined the different force fields for conformer minimization in RDKit, the Universal Force Field (UFF) and the Merck Molecular Force Field (MMFF), as well as the diversity of the generated ensembles. The first set consisted of 1,290 distinct small molecules from the CSD of which 469 were used before by Ebejer et al. and 821 additional structures from the CSD were extracted following the same procedure as described by Hawkins et al. Their second test set consisted of 238 crystal structures of drug-like molecules bound to proteins from the PDB, 79 of those were taken from the Astex Diverse Set. Like Ebejer et al. they found the same overall trend of increase of RMSD with an increase in the number of rotatable bonds. They concluded that ETKDG outperformed the older method, but also found different results for the two datasets and remarked that “in order to rule out effects from the smaller size of the PDB data set, the comparison should be repeated with a larger data set of biologically active conformations”.[80]

Watts et al. compiled a set of 150 macrocycles (the “Shelley Set”) for a benchmark study of MacroModel, with 67 ligands from the PDB and 83 ligands from the CSD.[154]

The set contains a larger fraction of peptidic macrocycles. The authors analyzed mean and median RMSD values for the ring atoms only and found overall good performance of MacroModel. However, especially for larger ring structures they observed larger RMSD values and elaborate on ideas for improving the performance of conformer ensemble generators.

A similar dataset was used for the validation of a novel algorithm for macrocycle conformer ensemble generation, Prime macrocycle conformational sampling (Prime-MCS) by Sindhikara et al.[155] It was compared to MOE (LowModeMD) and MacroModel (Baseline Search). The Prime-MCS dataset includes 208 molecular structures: 60 structures from the CSD, 130 obtained from crystal structures in the PDB and 18 structures downloaded from the Biologically Interesting Molecule Reference Dictionary (BIRD),[156] which is a subset of the PDB. The authors evaluated the algorithms in terms of accuracy, diversity of the ensembles and computational speed and concluded that Prime-MCS was the fastest and produced the most accurate and diverse ensembles of the methods tested.

The performance of the recently developed macrocycle generation for OMEGA was tested by Poongavanam et al. on a test set of the 60 available conformations in the PDB (resolution < 3 Å) and CSD of 10 flexible molecules, including 8 macrocycles. For roxithromycin the authors generated 9 conformations by NMR spectroscopy and compared those to the ensembles generated by the conformer ensemble generators. They concluded that OMEGA performed "somewhat better than MOE and MacroModel" on these compounds.[170]

During the first three decades of intensive research on conformer ensemble algorithms the quality and size of the benchmarking datasets had improved substantially but was still not at a point where sound statistical analysis of multiple algorithms was possible. In the course of this thesis the most comprehensive benchmark study of conformer ensemble generators to date was conducted on a dataset of 2,859 diverse and unique high-quality protein-bound ligand conformations from the PDB. Overall 16 algorithms were benchmarked, seven freely available and eight commercial conformer ensemble generators were compared to each other and later to the algorithm developed during this work.[1,2,4] For this purpose a fully automated cheminformatics pipeline for the selection and extraction of high-quality protein-bound ligand conformations from X-ray structural data was developed. The pipeline evaluates the validity and accuracy of the 3D structures of small molecules according to multiple criteria, including their fit to the electron density.

2.3 Sperrylite and Platinum Datasets

For the comparison of the performance of multiple conformer ensemble generators a large dataset of high-quality structures of protein-bound ligand conformations was essential for statistically meaningful results. To this end the Sperrylite and Platinum Datasets were compiled. The goal was to assemble a complete collection of all high-quality structures of small molecules (up to 16 rotatable bonds) in the PDB and to create a subset suitable for conformer ensemble generator benchmark studies.

2.3.1 Dataset Compilation

The workflow for the compilation of the Sperrylite and Platinum Datasets is described in ref D1 with some small improvements described in ref D2. A simplified overview of the cheminformatics pipeline for selecting high-quality X-ray structures of protein-bound ligand conformations from the PDB is depicted in Figure 6. Filtering over 350,000 ligand conformations from the PDB with the Platinum quality criteria resulted in the Sperrylite Dataset. The definition of these selection criteria was in line with those of the Iridium-HT dataset. The criteria included the fit of the 3D structure to the electron density, as well as physicochemical and structural properties.

After a simple query to the PDB web service,[171] all further steps were fully automated using shell and Python scripts. The sequence of the individual steps was optimized for short runtimes and can be found in Figure 1 of ref D1. Lists of "unwanted ligands" and "organo-metallic complexes" were obtained from the sc-PDB and used as filters to remove those compounds.[172] DPI values were calculated with DPICalc[173] according to the definition by Goto.[169] RDKit was used for Butina clustering,[174] computing canonical SMILES and structural similarity, as well as the number of rotatable bonds and heavy atoms. Only ligands with a minimum of 10 heavy atoms and 1 to 16 rotatable bonds were selected for the dataset.

The electron density support of the atom positions of all ligands was examined on electron density maps downloaded from the Uppsala Electron Density Server (EDS).[175] The automation of the evaluation of the fit to the electron density was made possible through the recently developed Electron Density score for Multiple Atoms (EDIA$_m$).[176] The EDIA$_m$ of the complete molecule results from the combination of the scores for its individual atoms (Electron Density scores for Individual Atoms, EDIA). EDIA was originally developed for the analysis of electron density around water molecules by Nittinger et al.[177] Only ligands with EDIA$_m$ greater than 0.8 were incorporated into the dataset.

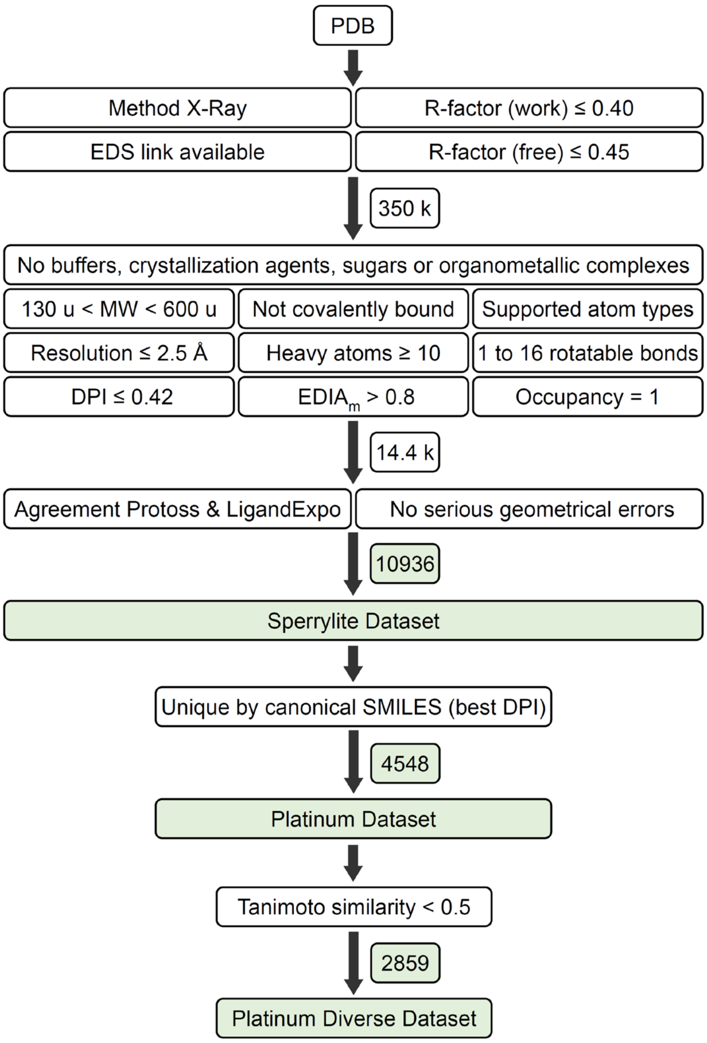


Figure 6: Simplified outline of the cheminformatics pipeline for selecting high-quality X-ray structures of protein-bound ligand conformations from the PDB and the resulting Sperrylite and Platinum datasets (green). Numbers of ligands that passed each collection of filtering steps (and which constitute the three datasets) are reported on the right of the arrows. The sequence of the individual steps that was optimized for short runtimes can be found in ref D1.

The Sperrylite Dataset resulting from the filtering process is a complete set of 10,936 high-quality structures of 4,548 unique protein-bound ligands.[3] The stereochemistry of the compounds was checked by generating isomeric smiles with UNICON.[178] To identify the approved drugs present in the Sperrylite Dataset the ligands were compared to the Approved Drugs subset of DrugBank.[179] Ninety-one ligands in the Sperrylite Dataset are represented by at least 10 structures, and these served as the basis of the analysis of the diversity of protein-bound conformations of small-molecule drugs and cofactors in ref D3.

The Platinum Dataset is a subset of the Sperrylite Dataset and consists of the 4,548 unique protein-bound ligands with the smallest DPI (Figure 6). Note that maximum DPI of 0.42 Å was allowed during the compilation of the Sperrylite and Platinum datasets, following the work of Hawkins et al.,[85] which leads to a maximum average positional uncertainty of 0.6 Å in the structures.[168] For each ligand with the same PDB ligand ID only the isomer with the most occurrences was kept from the Sperrylite Dataset to compile the Platinum Dataset. At the time DPI was used to select one conformation for the Platinum Dataset of each molecule from the Sperrylite Dataset, mainly because it was a well-established value in the literature for similar cases. As a global structure quality measure, it is not perfectly suited for this task. Today the EDIA_m should be used instead. This also allows proper prioritization between multiple ligands in different binding pockets of the same protein-ligand complex.

The Platinum Diverse Dataset in turn is a subset of the Platinum Dataset selected by Butina clustering with ECFP6-like Morgan fingerprints and a Tanimoto similarity cutoff of 0.5 (computed with RDKit). Because of this, the Platinum Diverse Dataset is the least biased of the generated datasets since it does not contain accumulations of similar compounds. If it is slightly biased, then by special interest in certain molecules (e.g. HIV proteases) by the scientific community and industry. And since it is compiled from X-ray structures from the PDB it is also biased by ease of crystallization. Nevertheless, this is true for all datasets of protein-ligand structures and the Platinum Diverse Dataset is by far the most suitable dataset for benchmarking conformer ensemble generators to date.

There exists a large overlap of 2,763 compounds between the Platinum Diverse Dataset 2016_01 (2,912 compounds, used in ref D1) and the Platinum Diverse Dataset 2017_01 (2,859 compounds, used in ref D2 and D4). Identical mean and median RMSD values with both versions of the Platinum Diverse Dataset were obtained for the RDKit DG algorithm, further indicating that the results produced with both datasets can be directly compared.

2.3.2 Analysis of the Sperrylite and Platinum Datasets

Ref D3 focuses on the analysis of the bioactive conformational space of a representative set of 17 approved drugs and cofactors extracted from the Sperrylite Dataset. As such it analyzed a part of the Sperrylite Dataset in great detail, but it can also serve as a general overview and introduction to the topic of diversity of conformations of protein-bound ligands.

The Platinum Dataset was analyzed with StructureProfiler, detailed results can be found in the supporting information of ref 180. StructureProfiler is a software tool for automated profiling of X-ray protein structures based on customizable criteria catalogues. These criteria include B factor checks, the search for uncommon torsion angles, an inhouse intra- and intermolecular clash criterion and the EDIA$_m$ to check the electron density support. One of the preconfigured criteria catalogues is the Platinum set of criteria. With this option StructureProfiler contains most of the steps in the workflow for generating the Platinum Dataset from a complete list of PDB structures, including DPI (Goto), R-factors and EDIA$_m$. StructureProfiler, like Conformator, is part of the NAOMI ChemBio Suite.[181] StructureProfiler also includes dataset configurations for the Astex and Iridium datasets, as well as the option to combine all three test set criteria. When tested with these combined criteria, intermolecular clashes with ligands not detected by NAOMI were found for 19 ligands and intramolecular clashes were reported for the ligands *VVV* from the PDB complex 3nhf and *1T4* from complex 4kky ligands respectively (Figure 7). These can be attributed to the different clash criteria of the Astex and Platinum datasets. For 240 of the 4626 structures in the Platinum Dataset EDIA$_m$ violations were detected. They can be attributed to the fact that EDIA$_m$ was updated in the meantime and that electron density maps are now retrieved from the PDBe[182] instead of the Uppsala Electron Density server (EDS).[175]

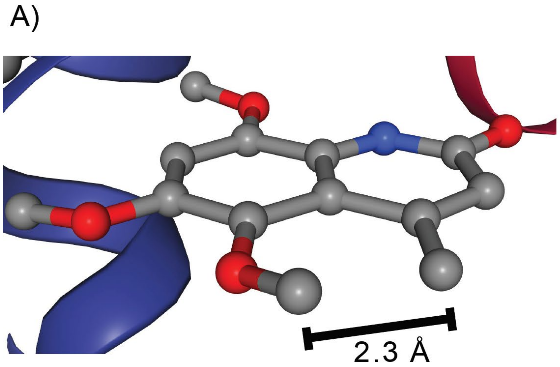


Figure 7: The ligands *VVV* in the PDB complex 3nhf (A) and *1T4* in the PDB complex 4kky (B). Distance values indicate intramolecular clashes according to the Astex Diverse Set clash criterion. Figures were generated using the ProteinsPlus Server;[183] hydrogens are not depicted.

Manual inspection of structures in the Platinum dataset revealed ligands closely interacting with metal ions. The PDB and NAOMI treat these metal ions as separate components and thus the molecules interacting with them are not filtered out during dataset generation, but they could be considered part of the ligand in some cases.[184] Examples include the ligands *ECA* in the PDB complex 2xv1 and *SE8* in 3mwf, each wrapping around an Fe3+ ion (Figure 8). Both occupy somewhat unusual conformations that are unlikely but possible to be produced by conformer ensemble generators.

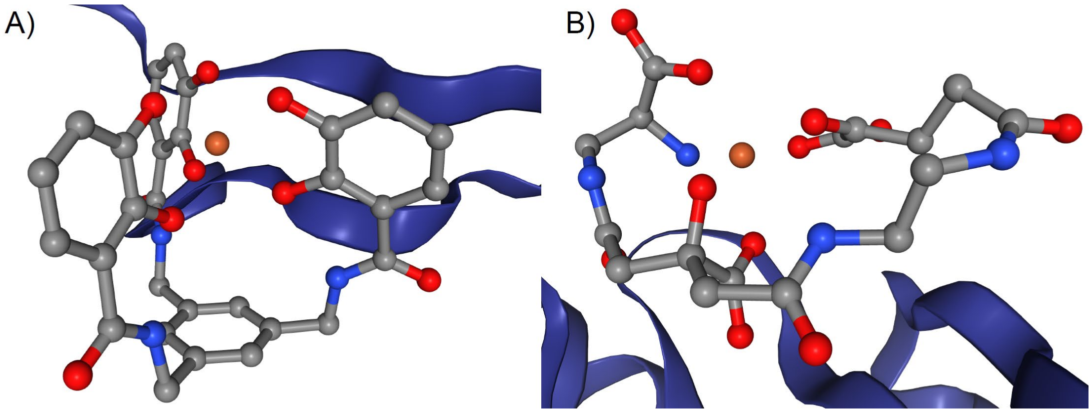


Figure 8: The ligands *ECA* in the PDB complex 2xv1 (A) and *SE8* in the PDB complex 3mwf (B) closely interact with an Fe^{3+} ion (brown). Because of this they occupy somewhat unusual conformations that are less likely to be produced by conformer ensemble generators. Figures were generated using the ProteinsPlus Server;[183] hydrogens are not depicted.

Systematic differences between the solid phase and the gas phase cannot always be excluded. An example of this is the preference of biphenyl (*BNL*) for a coplanar geometry when crystallized and a torsion angle of 44° in gas phase.[185] In the Platinum Dataset however a structure of biphenyl from the PDB complex 3gzx (biphenyl dioxygenase) is found with a dihedral angle of 124°, well supported by the electron density ($EDIA_m$ of 0.81) and in 4.5 Å distance from an Fe^{2+} ion.[186]

Three-membered rings lead to different and strained geometries in molecular structures. For example, each of the three carbon atoms in cyclopropane is connected to four other atoms, but instead of an sp3 hybridization a tetrahedral geometry is present. Each inner bond angle amounts to 60° in place of the energetically most stable 109.5°, but in the "bent bonds" model the interorbital angle is described with, it is 104°.[187] The torsion angle of the hydrogen atoms at neighboring carbons is nearly 0°. The torsion angles involving three-membered rings are not explicitly considered by the torsion library, although they are part of the overall statistic. The number of three-membered rings in the Platinum Dataset is 147, in the Platinum Diverse Dataset it is 110. Only 27 of those in the Platinum Dataset and 22 of those in the Platinum Diverse Dataset are non-terminal three-membered rings, only these are of higher interest for conformer generation.

The Platinum Diverse Dataset was further analyzed with NP-Scout, a machine learning approach that quantifies natural product-likeness of small molecules.[188] Out of the 2,859 structures 127 (4.4 %), including nine macrocycles, were assigned a natural product class probability of 1.0, and therefore likely are natural products. Of the 463 structures (16.2 %) that were assigned a probability of at least 0.8 to be natural products, 15 were macrocyclic compounds.

2.3.3 Usage and Analysis of the Platinum Datasets by Others

In addition to the analysis of the Platinum datasets in ref D1 and later with StructureProfiler, the Platinum datasets were already used for a number of benchmark studies by other groups and were analyzed further in the process.

Cole et al. used the Platinum Diverse Dataset (2,859 structures) for evaluation of a new knowledge-based conformer ensemble generator based on CSD data.[184] To avoid bias they used the Platinum dataset for evaluation only and not for training of the algorithm. Conformer generation was carried out with the maximum number of conformers set to 50 and 250 for easy comparison with results obtained in ref D1 and D2, that are part of this thesis. They were provided with the initial 3D inputs that were used in ref D2 by the author of this thesis. Cole et al. reproduced the evaluation of the ETKDG conformer generator and found small differences in mean and median RMSD. These differences can be attributed to the use of an analogous but slightly different algorithm for computation of RMSD values. Additionally, they analyzed the structures in the Platinum dataset with the knowledge-based library of molecular geometry derived from the CSD called Mogul.[189]

Jain et al. compared the performance of seven modes of ForceGen (version 4.4) directly to the results obtained in ref D2 on the Platinum Diverse Dataset.[190] The authors focused on speed and parallelization of their method that relies on a modified version of the MMFF94s force field.[87] They concluded that ForceGen was as accurate as OMEGA and faster than all other methods in the comparison, however, without reproducing runtimes or RMSD values for any of the other algorithms with their hardware setup. Jain et al. also explored conformer ensemble generation for macrocycles with ForceGen on the 29 macrocyclic ligands (about 1 %) of the Platinum Diverse Dataset. They pointed out that these macrocyclic compounds were less complex than those in macrocycle-focused datasets.

Wahl et al. benchmarked randomly generated conformers minimized with the MMFF94s force field against conformers minimized with the MM2-derived force field[191] implemented in the open-source software DataWarrior[192] and the OPLS3 force field on the structures in the Platinum Diverse Dataset.[193] Only 2,581 of the 2,859 molecules from the dataset could successfully be processed by all three force fields. This was due to the large failure rate of the MM2 force field of 9.7 %. The MMFF94s and OPLS3 force fields had a failure rate of 0.8 % each. They concluded that the conformers minimized with the MMFF94s are of similar accuracy to those minimized with the OPLS3 force field and that it represents a clear improvement over the MM2 force field for this task.

Yoshikawa and Hutchison analyzed the Platinum Dataset and divided the 4,548 compounds into 9,741 fragments with at least five atoms.[194] They found 7,852 (80.6 %) of these fragments in a rigid fragment database generated from the Crystallography Open Database (COD).[195] They used the coordinates of fragments from COD, Ligand

Expo[196] and the Platinum Dataset for implementing a fragment-based coordinate generation in the open source cheminformatics toolkit Open Babel. Yoshikawa and Hutchison also employed the Platinum Dataset to benchmark their new approach for coordinate generation and to compare its performance to that of the former version in Open Babel and to that of RDKit (release 2018.09.1) with the ETKDG method. To avoid bias only COD fragments were used for training in these tests. They found their method to be twice as fast as the old implementation and resulting in a greatly increased success rate, but also found RDKit ETKDG to be slightly more accurate.

Chan et al. used a bivariate von Mises distribution to analyze correlated torsions in small molecules.[197] To benchmark the performance of their new method, Bayesian Optimization with Knowledge-based Expected Improvement (BOKEI), they assembled a dataset of 533 unique molecules from the Platinum Dataset and the dataset assembled by Ebejer et al.[151] They also used their new tool to analyze the COD, ChEMBL 25[198] and the Platinum Dataset for the number of molecules with the presence of correlated torsions. They found 9.2 % of the 4,548 compounds in the Platinum Dataset, 13.5 % of the 110,623 compounds in the COD and 14.6 % of the 1,870,461 molecules in ChEMBL 25 to contain correlated torsions.

Wang et al. recently improved the conformer generation of RDKit ETKDG for molecules containing small or large aliphatic (i.e., non-aromatic) rings.[199] They added additional torsional-angle potentials to describe small aliphatic rings and revised the potentials for acyclic bonds to now also facilitate the sampling of macrocycles. In addition, recently updated vdW radii matching those in the Blue Obelisk data repository,[200] were utilized in the calculation of the distance bounds matrix. To restrict the enormous search space of macrocycles and bias it towards conformations found in experimentally derived structures, they introduced different heuristics based on elliptical geometry and customizable Coulombic interactions. The authors demonstrated the performance of the new algorithm on datasets of diverse macrocycles and cyclic peptides. Two datasets were used to test structures with small rings; 600 molecules from the CSD and 1,401 molecules from the Platinum Diverse Dataset with at least one aliphatic ring of a maximum size of eight and with a molecular weight below 600 g/mol. For the evaluation of the macrocycle conformer generation a dataset of 636 experimental structures for 482 unique single-macrocycle molecules was assembled based on the work by Hawkins et al. by filtering multiple datasets for structures with a single large ring; with 40 structures from BIRD, 262 from the CSD, 53 from the LigandExpo 2016, 261 from the Prime-MCS dataset by Sindhikara et al. and 19 from the most recent D3R Grand Challenge 4.[201] Additionally the NMR structure of a cyclic decapeptide was investigated. The authors analyzed RMSD values for complete molecules and for macrocyclic ring structures only, and used the two-sided paired *t*-test to compare the two versions of the algorithm. They conclude that the additions improved the ability of the ETKDG conformer generator to efficiently sample relevant conformations of small and large rings, but that the effect is only detectable for small ensemble sizes; on the structures in the Platinum Dataset it is visible for an ensemble size of 10, but not for an ensemble size of 100.

3

Development of a Novel Method for the Generation of Conformer Ensembles

3.1 Conformator

A novel knowledge-based algorithm for generating conformer ensembles called Conformator was developed during this thesis. It is based on the software library NAOMI[202] and the previously introduced CONFECT algorithm. Conformator is built on established concepts of incremental construction of conformers with a torsion driver at its core. These simple concepts are augmented by an elaborate algorithm for the assignment of torsion angles from a torsion angle library (revised and extended version by Guba et al.)[46] to rotatable bonds. The algorithm is described in detail in ref D4. As part of Conformator a new clustering algorithm for the assembly of conformer ensembles was developed. This cluster algorithm takes advantage of the fact that the list of initially generated conformers is partially presorted, to deduce individual RMSD thresholds for molecules and substantially reduce the number of necessary comparisons between pairs of conformers. The new clustering algorithm is based on sphere exclusion clustering,[203] it is described in detail and with a visual representation in Figure S1 in the supporting information of ref D4. The performance of the new clustering algorithm implemented in Conformator was tested against the performance of the k-medoids clustering algorithm (partitioning around medoids method)[204,205] and proved to be more than ten times faster, while reaching the same accuracy.

Conformator ensures chemically correct bond lengths and bond angles, as well as the planarity of conjugated systems (including rings) in its generated conformations. The algorithm offers two modes for conformer ensemble generation, “Fast” focuses on computational efficiency and “Best” on accuracy. In the validation study of Conformator was able to generate conformer ensembles for 99.9 % of all tested molecules. It showed remarkably high accuracy with no significant difference to the highest-ranked commercial algorithm OMEGA and significantly higher accuracy than the seven free

algorithms tested, including the RDKit algorithms. With a maximum ensemble size of 250, Conformator reached a median RMSD of 0.47 Å to the 2859 protein-bound ligand conformations in the Platinum Diverse Dataset. The median runtimes of Conformator (Fast 1 s, Best 3 s) were also very similar to those of OMEGA (2 s). Interestingly, OMEGA and Conformator performed best on different sets of molecules. OMEGA shows higher performance in sampling molecules with fewer than five rotatable bonds, which account for more than half of all molecules of the Platinum Diverse Dataset, but Conformator performs better on molecules with five or more rotatable bonds. This reveals a potential for further development in both algorithms.
Conformator is available as part of the NAOMI ChemBio Suite and as a standalone tool free for non-commercial use and academic research (at https://uhh.de/naomi). Conformator reads molecular structures from SMILES and InChI notations, as well as SD and MOL2 files. Note that NAOMI and with it Conformator is additionally able to process older standards of MOL files, but at the time of this thesis the corresponding parser was not tested thoroughly and hence it was not included in the published list of formats.

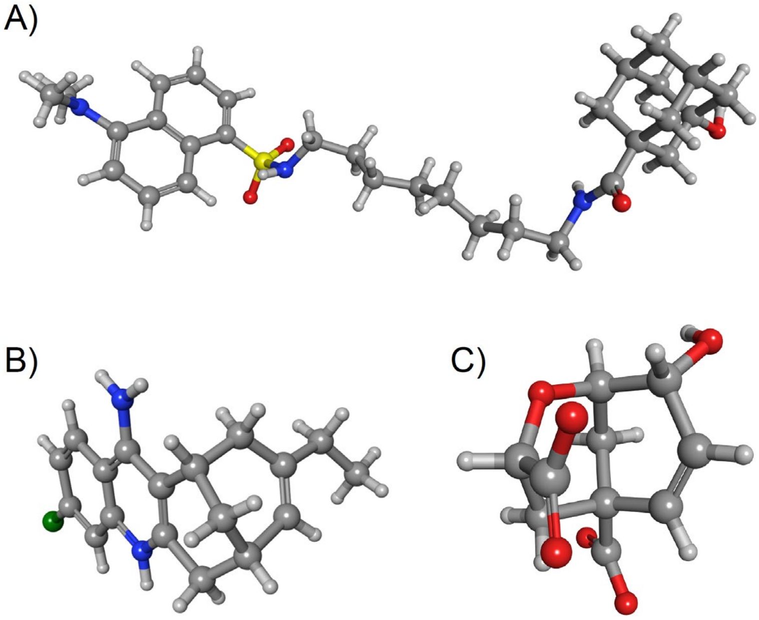


Figure 9: 3D models of the three molecules in the Platinum Diverse Dataset, for which Conformator cannot generate coordinates from 2D input; *SAW* (A), *HUX* (B) and *TSA* (C). Figures were generated using MOE.[41]

When given 2D input (SMILES) only, Conformator failed to produce ensembles for three out of the 2859 molecules in the Platinum Diverse Dataset. These three molecules contain small, bridged ring systems (Figure 9). The three molecules can be successfully processed, if valid input coordinates are given and the option to generate new 3D coordinates is not set.

Intensive testing of Conformator was done only for molecules with up to 16 rotatable bonds. Thus, most larger compounds might be outside the applicability domain of the model. For a small sample size (e.g. cut macrocyclic structures) Conformator successfully produced ensembles for molecules with up to 36 rotatable bonds. The most extreme test case (suggested by a reviewer of ref D4) was antiamoebin I with 56 rotatable bonds. It proved to be too large for the underlying model (and some common computer systems). There are $27{\cdot}10^{21}$ theoretically possible conformations in the underlying model for antiamoebin I. In cases like this Conformator still produces an ensemble (on systems with at least 16 GB of RAM) but is not able to properly execute its described workflow. A warning is given to the user that the input molecule is highly flexible and that insufficient sampling is to be expected. And while the general workflow and algorithms (especially the reduction of the rotatable bond angles) can handle input of arbitrary size in theory, it is not advised to use the program for molecules with more than 16 rotatable bonds. Not only because of the technical limitations but also because the theoretical support for representing the flexibility of larger molecules with conformer ensembles produced by torsion driving is questionable.

One of the key features of Conformator is its ability to generate conformer ensembles for macrocyclic structures. The macrocycle conformer generation algorithm of Conformator is described in detail in ref D4 and a visual representation can be found in Figure S2 in the supporting information. Instead of processing individual rings, the concept of unique ring families (URFs)[206,207] is utilized to consider one ring family at a time. All macrocyclic URFs in a ring system are iteratively cut at single bonds (outside of small rings) until no macrocycle remains in the resulting ring system. It is ensured that during this process the molecule remains connected. The open ring structure is then, in an intermediate step, handled with Conformator's standard algorithm for conformer generation. The resulting conformations are clustered and the cut bonds are reintroduced to close the macrocycle conformations again. Therefore the conformations used as starting points for cyclization and subsequent minimization are already valid (with the exception of the part where the macrocyclic bond has to be reintroduced), in a stark contrast to DG approaches, which usually start from randomly generated coordinates. Ref D4 introduced the macrocyclic optimization score (MCOS) that is used to reconstruct the macrocycle. It includes several well-known components from common force fields and some components specific to the optimization of macrocycles. The composition and calculation of the MCOS are described in the associated publication and formulas and graphs of its terms are provided in the Figures S3−S9 of the supporting information. To reduce the number of parameters in the optimization down to at most one bond angle per atom and one torsion angle per bond in the optimization, it is performed employing internal coordinates (the torsion angles and bond angles in the macrocycles). The optimization of the closed macrocycles is based on a modified reimplementation of the BFGS-B algorithm,[208,209] inspired by recent work on the refinement of water placement in protein crystal structures.[177] The main reasoning behind Conformator's macrocycle conformer generation algorithm

was to minimize the number of interfaces between cut parts of the molecule. This is nearly the opposite approach to the solution offered by Bonnet et al. where the number of interfaces between separate components is close to the theoretical maximum.[149]
Conformator's capabilities to produce ensembles for molecules with multiple macrocycles were tested on the 49 macrocyclic structures in the Sperrylite Dataset and for structures obtained from the BIRD library.[4,210] The tests clearly showed that the conformer ensemble generation for molecules with multiple macrocycles is able to generate conformer ensembles for a large variety of macrocyclic systems. However, no meaningful statistics on the accuracy of reproduction was possible, since the available pool of high-quality structures of protein bound macrocycles was too small, e.g. most of the structures in BIRD do not meet the quality criteria of the Platinum datasets.

Known information about bond lengths, bond angles, clashes and planarity was used to check for geometrical errors in conformations that were candidates for the Platinum Datasets. The same geometry checks are implemented in the previously mentioned StructureProfiler and the corresponding NAOMI library. They are used to filter macrocycle conformations in Conformator. The same checks for geometrical errors were also performed on conformers generated with Conformator for all molecules in the Platinum Diverse Dataset (2856 structures) with a maximum ensemble size of 250 and revealed no detectable wrong geometries left in Conformator output. When generating conformer ensembles for large and more complex molecules than those in the Platinum Diverse Dataset, small geometrical errors can be allowed, if otherwise no conformation could be generated.
Optionally Conformator can consider hydrogen clashes during the conformer generation and clustering. In this case the hydrogen clashes are also used during conformer generation and in the RMSD clustering of macrocycles. By default, hydrogen clashes are not utilized.

3.2 Benchmarking Conformer Ensemble Generators

Accuracy of a conformer ensemble was defined in the benchmark studies, like in most evaluations of conformer ensemble generators, as the minimum RMSD in Å measured between the experimentally determined protein-bound conformation and any conformer of the computed ensemble. Conformer ensemble generators usually are designed to generate diverse ensembles, because of this accuracy is, to some extent, a function of ensemble size.[211] The chance for one of the conformers in the ensemble to closely resemble the experimentally observed conformation generally increases with the number of conformers generated.
Two benchmark studies that directly compare the performance of seven free and eight commercial conformer ensemble generators were conducted with the Platinum Diverse Dataset during the development phase of Conformator.[1,2] The freely available

algorithms benchmarked were the RDKit DG and ETKDG algorithms, Confab,[212] Frog2, Multiconf-DOCK[213] and the Balloon DG and GA algorithms.[214] The eight commercial algorithms were ConfGen,[215] ConfGenX, cxcalc,[216] iCon, MOE LowModeMD, MOE Stochastic, MOE Conformation Import and OMEGA were benchmarked against RDKit DG the best performing free algorithm.

The benchmark studies showed that the distance geometry approach of RDKit and its knowledge-based counterpart, ETKDG, were the best freely available conformer generators at the time. OMEGA proved to be the leading commercial algorithm. The tests also showed that commercial algorithms generally obtain higher accuracy and robustness with respect to input formats and molecular geometries.

3.2.1 Measures of Similarity

Quantifying the similarity between molecules or conformations of the same molecule is a non-trivial task. While many measures for molecular similarity exist, the RMSD is the de facto standard for comparison of conformations and thus in benchmarks of conformer ensemble generators. RMSD is generally regarded as an objective, universal and intuitive function, but RMSD results should always be viewed in context and with its disadvantages in mind. The RMSD is not normalized, which can result in very high RMSD values for highly flexible molecules. This is not a problem when comparing the RMSD of conformers of the same molecule, but it is a frequent practice to combine the RMSD values of very different molecules to calculate mean or median values. Then the lack of normalization can severely skew results, especially for small dataset sizes. Another important limitation of the RMSD is that it is context-free and thus neglects potential interactions. Many algorithms for RMSD calculation exist and they usually differ because of different handling of symmetries in molecular structures.

RMSD values for this work were calculated with NAOMI, which determines the RMSD based on the best superposition of a pair of conformers, taking into account molecular symmetry via complete automorphism enumeration. Conformator offers the possibility to calculate the minimum pairwise RMSD between a generated conformer and the input conformer, and the minimum pairwise RMSD between any generated conformers. (The user is advised to use these options only if necessary, since they may lead to substantially longer runtimes.)

There exist a large number of alternatives to RMSD, e.g. PubChem3D[217] (an extension to PubChem) and the shape-optimized similarity search tool ROCS[218] primarily utilizes shape-Tanimoto (ST), color-Tanimoto (CT), and TanimotoCombo (TC).[217] TC measures the complementarity in shape and distribution of chemical features in 3D.

The Generally Applicable Replacement for rmsD (GARD) evaluates the alignment between the atoms of a reference structure and the atoms of a conformation.[219] GARD weights atomic contributions by their relative importance to binding. This weighting is based on statistics by Andrews et al.[220] and is customizable, but this is also a weakness of the approach, since whether a functional group is important for binding always heavily depends on the specific protein and the ligand.

Torsion Fingerprint Deviation (TFD) is another measure for the comparison of small-molecule conformations.[221] The computation of TFD is much faster than that of RMSD, since no superposition of the structures has to be determined. In addition to the RMSD values, TFD calculations were carried out for all experiments with conformer ensembles in this work. RMSD and TFD values are usually correlated and no relevant difference was found between the two for any of the direct comparisons of performance of conformer ensemble algorithms. Thus overall, the results obtained by TFD comparisons confirmed all results found by comparison of RMSDs.

3.2.2 Validation Tool

To validate the results of various conformer ensemble generators a validation tool was developed (see Figure 10). The goal was to make the conditions for comparing the different programs as uniform as possible. The input parameters for the tool are the conformer ensemble generator to be used, the desired maximum ensemble size and the data set to be read. The crystal structure of each molecule of the data set is individually loaded and used, on the one hand, at the end of the validation for the comparison with this same original structure. On the other hand, it was used after erasing and re-calculating the 3D coordinates as input for the conformer ensemble generator. This approach ensured that each of the algorithms started the ensemble generation with the same conformation of the molecule and was necessary because many of the tested conformer ensemble generators need 3D structures as input. The conformer ensemble generator was called with the input parameters and the newly calculated 3D coordinates. For most conformer ensemble generators, including Conformator, the maximum ensemble size should not be considered a hard limit that the algorithm must meet, but a maximum value that may be reached in certain cases.

Conformator and OMEGA were benchmarked with both 3D structures and SMILES as input. No difference was found between the results for both algorithms, which can be seen as proof that in both algorithms the 3D information given by the input is really not used.

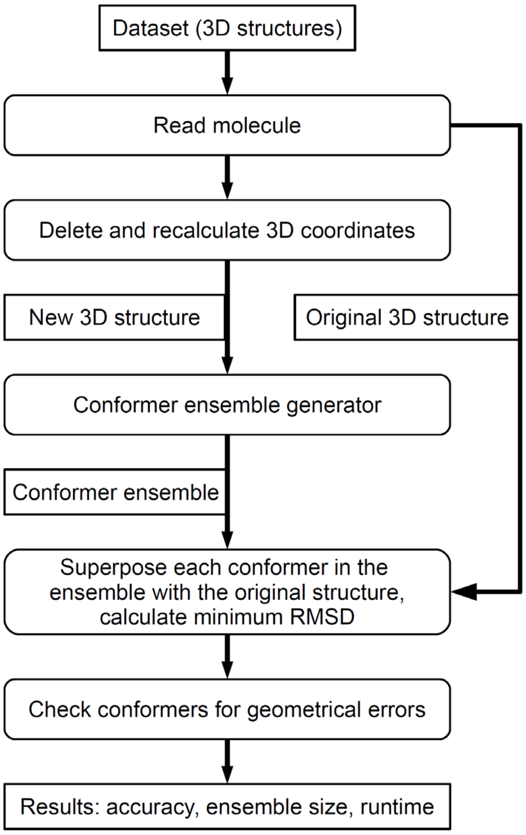

Dataset (3D structures)
Read molecule
Delete and recalculate 3D coordinates
New 3D structure
Original 3D structure
Conformer ensemble generator
Conformer ensemble
Superpose each conformer in the ensemble with the original structure, calculate minimum RMSD
Check conformers for geometrical errors
Results: accuracy, ensemble size, runtime

Figure 10: Workflow for benchmarking conformer ensemble generators. Reprinted with permission from (Friedrich et al., 2017).[1] Copyright 2017 American Chemical Society.

3.2.3 Conformer Ensemble Generation

Conformer ensembles were generated with standard 3D conformation computed for each molecule with NAOMI from its SMILES notation as input. All conformer ensemble generators tested were started with the validation tool described above. With the exception of MOE and RDKit no further scripts were needed. As suggested in the RDKIT User Manual a Python script was written for conformer generation with RDKit DG and RDKit ETKDG. The Python script used is included in the Appendix (see Appendix B3). Functions of MOE can be accessed in the graphical user interface (GUI) or via the built-in programming language Scientific Vector Language (SVL). The SVL-script used to call the function “ConfSearch” from the command-line is shown in the Appendix (see Appendix B4). Non-default settings used for conformer ensemble generation with freely available algorithms were described in ref D1 and with commercial algorithms in ref D2.

Confab requires a 3D input structure with reasonable bond lengths and angles since the algorithm does not currently explore ring conformations. This is an unfair advantage in the tests, especially for larger rings and ring systems. However, since Confab performed poorly in comparison to most other algorithms this fact was not emphasized.

3.2.4 Hardware Setup

All calculations for this thesis were performed single-threaded on Linux workstations equipped with Intel Xeon processors (2.2−2.7 GHz) and 126 GB of main memory running openSUSE 42.2, with the exception of ref D1, here the same workstations were running openSUSE 13.1 at the time of the study.

4

Statistical Analysis and Additional Methods

A thorough statistical analysis was a central part of the three publications that include benchmarks of conformer ensemble generators for this thesis, i.e. ref D1, D2 and D4. The summarized results and methods of the statistical analysis can be found in the associated publications, and detailed results in the supporting information. This sets the work apart from most publications in the field, since it has been, unfortunately, common practice to claim “significant” results, while ignoring errors in crystal structures, bias in the datasets used and the minimum number of data points required (i.e. dataset size). Some authors at least noted that small dataset size was an issue, but the common reaction was to add another small number of structures and again declare significant differences in performance in the results, without any form of statistical analysis.

A notable exception to these practices is the work by Hawkins at al. for the evaluation of OMEGA and related work.[85,119,222] The authors noticed the same general trend in the scientific community in this field and stated: “The usual practice in this area has been to compare an aggregate statistic such as mean or median results, from a number of different tools or parameter sets and to declare one superior, without any account of the errors in these terms.”[85]

As described earlier, during this thesis the Platinum Diverse Dataset was compiled and used for benchmarking studies of conformer ensemble generators. The dataset is of adequately high quality and sufficient size for a statistical evaluation. For the comparison of the performance of the different algorithms for conformer ensemble generation pairwise Mann−Whitney U tests were carried out to test for statistical significance at $\alpha = 0.05$ and $\alpha = 0.01$, with the Holm−Bonferroni method[223] applied to control the familywise error rate (FWER). The p-values were reported for pairwise comparisons at maximum ensemble sizes 250 and 50, for the free algorithms in Table S1 of the supporting information of ref D1, for the commercial algorithms (including different force fields and clustering algorithms) in additional text files ci7b00505_si_002 and ci7b00505_si_003 of the supporting information of ref D2. For comparisons to Conformator please refer to Table S2 and S3 of the supporting information of ref D4.

All RMSDs reported in this work (including ref D1–4) were calculated with NAOMI, which takes molecular symmetry into account via complete automorphism enumeration. It was used to calculate the minimum heavy-atom RMSD between the reference structure and any of the computed conformers of an ensemble and the minimum heavy-atom RMSD for the best superposition of each pair of conformers.
It is meaningless to compute or compare RMSD values with greater precision than the accuracy of the underlying experimental data, e.g. when comparing generated conformers with experimentally determined structures. DPI values for all complex structures were calculated with DPICalc[173] according to the definition by Goto.[169] Following the work of Hawkins et al.,[85] a maximum DPI of 0.42 Å was allowed during the compilation of the Sperrylite and Platinum datasets. This leads to maximum average positional uncertainty of 0.6 Å in the structures.[168] The ligand structures in the complexes were assumed to have average positional uncertainty in the complex structures, but the real uncertainty for the position of the ligand atoms should be even lower, especially when considering the EDIA$_m$ calculations. Nevertheless, an RMSD value of 0.6 Å was used as an important threshold in the comparisons of success rates of the different conformer ensemble generators in representing protein-bound ligand conformations.

NAOMI was also used to detect anomalous geometries in the datasets and in generated conformers. The deviation of atom angles and bond angles from known optimal values as well as the divergence from planarity of aromatic rings and ring systems of up to 6 bonds per relevant cycle was measured.

In the dataset generation and for analysis of conformer ensembles the number of rotatable bonds was calculated with RDKit.[224] The default setting was used, that does not consider amide and ester bonds as rotatable.

NCBI BLAST (basic local alignment search tool)[225–227] was used to calculate the all-against-all sequence identity of proteins and the sequence identity of individual pairs of proteins was measured with MOE[90] based on sequence and structural alignments. R[228] was utilized to generate principal component analysis (PCA)-derived score plots of the alignments with the minimum median RMSDs were generated with R for each ligand in ref D3.

Runtime measurements for each tested algorithm and each molecule in the Platinum Dataset were conducted while processing SD files containing single molecules. The resulting runtimes were rounded to full seconds. In repeated runtime experiments deviations of less than 5 % were observed.

5

Conclusions and Further Directions

5.1 Improvement of Benchmark Datasets

Data quality is of high importance in all scientific fields because low data quality leads to invalid conclusions. Many analyses in cheminformatics, including the evaluation of conformer ensemble generators, heavily depend on atomic resolution 3D structures from X-ray crystallography. The reliability of the crystallographic model is directly affected by the quality of collected X-ray diffraction data. The high quality of the structures contained in the Sperrylite and Platinum datasets is ensured by a multitude of criteria, including global and local fit of the crystallographic model to the X-ray diffraction data. On the one hand, data quality is of extreme importance for benchmarking conformer ensemble generators, e.g. RMSD values can only be reported with high precision if the structures for comparison are of high quality. On the other hand, dataset size is the essential prerequisite for assuring statistical significance and for detecting more subtle differences in the performance of different algorithms. The Sperrylite and Platinum datasets derived for this thesis with the described cheminformatics pipeline are the largest publicly available datasets of such high quality. The Sperrylite Dataset is a complete set of 10,936 high-quality structures of 4,548 unique protein-bound ligands filtered from more than 350,000 crystal structures in the PDB. Its subset the Platinum Dataset consists of precisely these 4,548 unique protein-bound ligands, each of them the conformer (of the same molecule) with the smallest DPI. The Platinum Diverse Dataset is a diversified subset of the Platinum Dataset and still includes 2,859 compounds.

In the compilation of the next version of the datasets the quality of the Platinum Dataset can be enhanced by using a more appropriate measure to select the conformer for each molecule from the Sperrylite Dataset. Here, the EDIA$_m$ should be used instead of the DPI to select the best conformer, since the DPI is a global structure quality measure and the $EDIA_m$ allows checking the local electron density fit of each ligand. Consequently, the $EDIA_m$ would also allow proper prioritization between multiple conformations of the same ligand in different binding pockets of the same protein-ligand complex.

Another way to elevate the quality of the Platinum Dataset, especially for applications which exceed the scope of pure benchmarking of conformer ensemble generators, is further filtering of molecules. A class of molecules that was previously neglected in the dataset compilation is sugars. A preliminary investigation revealed that there are at least 45 sugar-like structures in the Platinum Diverse Dataset. The recently developed algorithm SugarBuster[109] for the removal of sugars and sugar-like moieties could be applied to generate another Platinum Dataset without these compounds. (The first executable version of SugarBuster was developed by the author of this thesis, based on work by K. Sommer. SugarBuster was later refined, completed, and applied by M. Garcia de Lomana.) While some sugars and sugar-like compounds might propose interesting challenges to conformer ensemble generators, they are generally not of great interest for drug discovery.

False positive results in high-throughput screenings are still one of the main problems in the early stages of drug discovery. Assay interference is typically caused by aggregators, reactive compounds, or so-called pan-assay interference compounds (PAINS).[229] Many of these compounds are so-called "frequent hitters", because they interact with a wide variety of target proteins or interfere with the detection method.

Hit Dexter is a machine learning approach that predicts frequent hitters and allows to filter potential PAINS, compounds with undesirable fragments and potential aggregators from the dataset.[230] Together with an analysis of drug-likeness, like the comparison of the Platinum datasets with the Approved Drugs subset of DrugBank in ref D1, Hit Dexter and SugarBuster could be applied to generate a subset of ligands that is even more focused on compounds of interest for drug discovery.

For now the Platinum Diverse Dataset (2,859 compounds) in particular can serve as a standardized set of molecules for the validation of conformer ensemble generators. With this dataset there finally exists an unbiased, diverse, updatable dataset of adequately high quality and sufficient size for this task. After more than 30 years of conformer ensemble generator validations, this is a very valuable result in its own right.

For macrocyclic conformations the situation is entirely different. At the time of the compilation of the Sperrylite Dataset (February 2017) there were only 760 crystal X-ray structures containing macrocycles in the PDB and only 49 of them adhered to the Platinum quality criteria. These 49 conformations of macrocycles in the Sperrylite Dataset were of 36 unique molecules that are part of the Platinum Dataset. In light of the high interest in these compounds for drug discovery, there is a clear need for a large and high-quality benchmark dataset of macrocyclic structures. In many ways the size of macrocycle datasets is nowadays similar to that of the drug-like datasets of small ligands two decades ago, but the PDB is rapidly growing, new structures are of higher quality overall and tools like StructureProfiler greatly simplify quality checks of large compound collections. The logical next step is to apply the Platinum quality criteria to the 636 macrocyclic structures (482 unique molecules) assembled by Wang et al. Additionally, the PDB is rapidly growing and the interest in macrocyclic compounds is continuously high, which will lead to a steep increase in the number of X-ray structures containing them. Thus, the next version of the Sperrylite Dataset could allow for a thorough investigation of macrocycle conformer generation.

5.2 Future Development of Conformator

The development of the new knowledge-based algorithm for generating conformer ensembles called Conformator benefited greatly from the findings and experience gained during the benchmarking studies and the analysis of the variability of bioactive conformations. It could be shown that Conformator is accurate and effective with significantly higher performance than all non-commercial tools, including the RDKit algorithms, and has extremely similar performance (no significant difference) to the highest performing commercial algorithm, OMEGA. Conformator further stands out with its handling of macrocycles, as well as robustness with respect to input formats and molecular geometries. Despite the substantial improvement in performance compared to its predecessors, some additions and refinements for Conformator that might be implemented in the future are presented in the following paragraphs.

Conformator is designed to handle both 2D and 3D input, it showed a 100 % success rate for 3D input and a 99.9 % success rate for 2D input of the molecules in the Platinum Diverse Dataset. For a few small, bridged ring systems (three out of the 2856 structures in the Platinum Diverse Dataset with the PDB IDs *HUX*, *SAW* and *TSA*) Conformator cannot generate coordinates from 2D input at this point. In these rare cases, the algorithm for macrocycle conformer generation could provide effective solutions. Small bridged ring systems where no coordinates can be generated by the standard algorithm would be cut and reconnected after conformer generation. A proof of concept test was successfully conducted on a single molecule (*TSA*). For general applicability, a special minimization for these cases (similar to the macrocycle minimization procedure described in D4) might be necessary, considering ring strain and abnormal angles of bonds to neighboring atoms. This could be supported or replaced by an update of the small ring assembler and the small ring template library in NAOMI. Even then, a generalized version of the macrocycle conformer generation procedure might be used as a fallback mode in case the template library approach fails. This would lead to a 100 % success rate of Conformator even for the 2D input of the Platinum Diverse Dataset.

The present macrocycle conformer generation procedure could be fine-tuned for improved results or shorter runtimes. At the moment a relatively rough selection is obtained during pre-selection of ring-like (cut) macrocycle conformers. Additionally, further rules could be implemented to determine which bond to cut when. Currently, the first bond that fits the priority rule system is cut, prioritizing carbon−carbon and then carbon-incident bonds, first outside then inside of conjugated systems. Also, bonds that are not adjacent to small rings are favored, but optimizing the selection process could include searching for a bond that is as far away as possible from more rigid parts (rings, branches) of the molecule. This might ease the search for solutions in the optimization process. Additionally, the maximum number of macrocycle conformers before and after internal clustering could be more precisely adapted to the respective molecule (e.g. number of rotatable bonds, atom types).
Conformers of macrocycles that were generated and optimized by the macrocycle conformer generation procedure (without geometrical errors detected) could be saved as templates, so they do not have to be calculated again. This would lead to a massive speed-up for the few molecules that are responsible for a large portion of the overall runtime when generating conformer ensembles for the complete Platinum Dataset repeatedly. However, it is questionable whether these templates would generally be useful since the chemical complexity of the macrocycle also increases (almost exponentially) with the ring size.

Another approach that mainly benefits macrocycle conformer generation but could also be useful for other, particularly flexible, molecules, is the biasing of the ensemble for specific boundary conditions. The user might e.g. wish for particularly round macrocycles or might want to target a specific binding pocket or shape. An extreme example is the binding of the macrocyclic ligand *1P1* in the minor-groove of a DNA duplex, in the complex structure 3i5l from the PDB (Figure 11).[231] The resulting conformation is particularly challenging for conformer ensemble generation. It is theorized to disrupt the transcription factor-DNA interface and thus influence gene expression. Since many human diseases are caused by dysregulated gene expression, this form of DNA modulation is a very interesting application for macrocyclic drugs.[232]
The biasing of the conformational search can be implemented in Conformator on the basis of the clash handling during conformer generation. Here, whole branches of the search tree can be rejected early in the process, leading to a more thorough investigation of other regions of the conformational space of the molecule. Exclusion volumes (space that is not allowed to be occupied by parts of the molecule) could simply be implemented as “overly large atoms”, e.g. in the middle of macrocycles or surrounding the ensemble to mimic a binding pocket. However, it should be noted that, similar effects might also be achieved by guiding the conformer generation along a circle or implementing eccentricity constraints as recently described by Wang et al.[199] This would also lead to significantly fewer structures that have to be optimized.

Overall, it is an interesting approach to combine exclusion volumes and other concepts of pharmacophore modeling with conformer ensemble generation, not only for macrocycles but for all molecules. Conformator could be expanded to a pharmacophore-guided conformation generation, that compiles ensembles for specific binding pockets. Similar approaches are being pursued in the development of flexible 3D pharmacophores and have already shown promising results in multiple applications.[65–67]

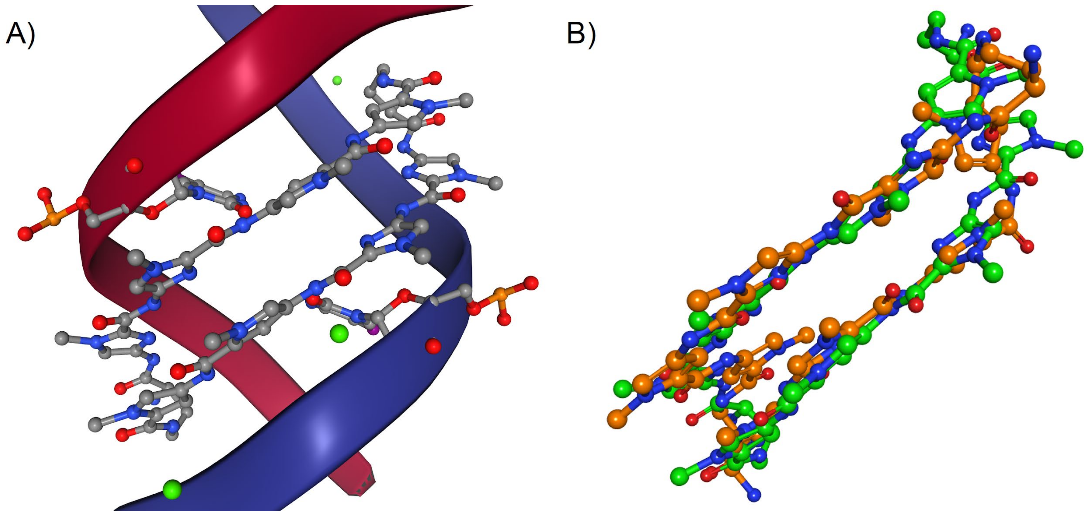


Figure 11: The macrocyclic ligand *1P1* bound in the minor-groove of a DNA duplex of the PDB complex 3i5l (A); and the same ligand conformation (green carbon atoms) superposed with the best fitting conformation generated by Conformator with standard settings (orange carbon atoms). The twisted conformation is extremely challenging to reproduce; the RMSD between the two conformations is nearly 4 Å. The figure was generated with MOE;[41] hydrogens are not depicted.

The user of conformer ensemble generators does not always care for ensembles smaller than the defined maximum ensemble size. Another quality level could easily be implemented in Conformator using already-generated conformations for filling out gaps (the largest differences in RMSD between conformers in the ensemble) as optimally as possible until the maximum ensemble size is reached. This maximizes accuracy for cases where smaller ensemble sizes than the maximum ensemble size are not desired. Additional conformations could be generated, if the maximum ensemble size is not reached, by finer sampling of torsional angles. Finer sampling of the conformational space might also allow to avoid small clashes that are currently allowed if otherwise no conformation could be generated. These intramolecular clashes, however, might be better handled by post-optimization of the generated conformations with the optimization procedure originally developed for macrocycle conformer generation. Note that the detection and handling of clashes is heavily dependent on their definition. While Conformator allows overlaps of up to 30 % of the van der Waals radii of 1−4-connected (or more distant) heavy atom pairs that are not part of the same ring system, other definitions are much more strict (cf. 2.3.2 Analysis of the Sperrylite and Platinum Datasets). An additional option can be implemented into Conformator that gives the user the choice between different clash definitions.

For current and future benchmark studies of conformer ensemble generators standardization is highly important. Only if the same dataset and performance measures are used, results can directly be compared. While this thesis standardized the benchmark of conformer ensemble generators and the creation of large high-quality benchmark datasets, it did not standardize the calculation of the RMSD to the same degree. RMSD results calculated with different algorithms for the same comparison of molecular structures can differ quite severely, most likely through different handling and definition of symmetries. Therefore, it is imperative that in the near future a thorough investigation into the RMSD calculation is conducted and its measurement standardized. For now the RMSD calculation options implemented in Conformator offer one way to compare RMSD values calculated with other methods in future benchmark studies.

In Conformator the flexibility of a molecule is estimated based on the maximum number of possible conformations resulting from the enumeration of all torsion angle values stored in the library, without the consideration of potential clashes. This value is multiplied by a factor (10 for the mode “Fast” and 20 for the mode “Best”) and the result used as the maximum number of candidate conformers to generate before clustering. At the moment this value is computed repeatedly throughout an angle removal procedure that iteratively removes angles from rotatable bonds. The number of rotatable bonds, torsion angles (peaks and tolerances) for each bond and ring templates for small rings is known, as well as the maximum number of conformers for clustering and the maximum number of conformers in the final ensemble. Hence there might be a way to calculate (or at least closely estimate) which rotatable bonds will keep how many (and thus which) torsion angles, to reduce the number of necessary calculations and speed up Conformator even further.

The novel clustering algorithm implemented in Conformator for the compilation of representative conformer ensembles exploits the partial presorting of consecutively generated conformers and allows for high speed, without significant loss of accuracy. On the one hand, the new clustering algorithm should be compared to further clustering algorithms in addition to the already tested K-Medoids algorithm, to ensure it is the most effective for this use case; possible candidates include the Jarvis-Patrick (with fuzzy similarity measure) and the X-Means algorithm.[233,234] On the other hand, it could be interesting to look for further applications of the new clustering algorithm.

The clustering procedure in Conformator might be further sped up by implementing different search algorithms for the ideal cluster radius (e.g. binary search), instead of the iterative process described in the supporting information of D4. The clustering algorithm is hard to speed up by parallelization, because it relies on the fact that the list of initially generated conformers is partially presorted to avoid unnecessary RMSD calculations. Still, it might be beneficial to implement a parallelized version of the clustering algorithm to evaluate the acceleration of the algorithm. Parallelization might be useful outside of the clustering procedure for parts of molecules, when computing conformer ensembles of large and flexible molecules or molecules with multiple macrocyclic systems. Usually conformer ensemble generation is already heavily parallelized on the molecule level, meaning single molecules of potentially large databases being processed in parallel.

More appropriate handling of specific groups of molecules depending on their individual properties could be implemented in Conformator. Different machine learning methods might be applied to determine chemical patterns that correspond to certain settings of internal options of the algorithm. These options could include the maximum number of conformers to generate before clustering, the initial radius and the increase of the radius for the next round of clustering or different clash tolerances (maximum overlap of van der Waals radii or when to take hydrogen clashes into account). This could happen on different levels, e.g. for classes of molecules or specific to types of rotatable bonds, similar to the torsion library already in use. For the classification of molecules, simple descriptors like the number of atoms, rotatable bonds, or rings might be sufficient.

The steadily growing computational power of widespread computer systems will over time enable scientists to work with larger ensembles. This might only be possible for molecules with a certain number of rotatable bonds (the problem of combinatorial explosion remains) and brings with it a danger of worsening the signal-to-noise ratio by burying the relevant conformations, leading to more false-positive results.

Since OMEGA and Conformator perform best on different sets of molecules, these differences should be further analyzed. It is possible that OMEGA performs particularly well on the smallest molecules in the Platinum Diverse Dataset because it uses PDB-derived information for biasing torsion angles.

In the same line, it might be an interesting approach to include all known high-quality conformations of the molecule in question from the Sperrylite dataset into the output ensemble. While it is impossible to test the complete version of this approach that includes all high-quality conformations, a benchmark would still be possible by splitting the Sperrylite dataset with a date cutoff (e.g. cutoff 2015) into an “output set” (conformations the algorithm is allowed to use as output) and a test set. An RMSD cutoff for too-similar conformers should be defined and clustering used to reduce large ensembles, keeping the database size small enough for a desired maximum size. Note that the objective of this approach is not to artificially dominate benchmark studies of conformer ensemble generators but to produce scientifically valuable assistance to downstream programs and users that lack the time or expertise to compile the relevant conformations.

Manual examination of structures in the Platinum dataset revealed that some of the compounds interact closely with metal ions, leading to unusual conformations. In future versions of the Platinum Dataset these could be regarded as “part of the ligand” (as suggested by Cole et al. in ref 184) and filtered out or included in a separate dataset. Usually conformer ensemble generators skip compounds with metals. NAOMI could be extended to handle these cases more precisely in the future, the recently released tool METAlizer can already be used to predict and visualize the coordination geometry of metals in metalloproteins.[183,235] On the one hand, Conformer ensemble generation considering metal ions might be possible on this basis. On the other hand, these cases may be easier handled through the introduction of exclusion volumes as described earlier for the pharmacophore-guided conformer ensemble generation.

A general solution to the problem posed by conformer ensemble generation greatly depends on the exact definition of the problem. For instance, the accuracy needed, the time frame, what level of information aggregation is appropriate and the specific use case (e.g. downstream programs involved) have to be defined. To solve the problem to a degree that is mostly independent of the exact problem definition, it might be necessary to abandon the “classical model of physics” behind conformer ensemble generation today. A closer look at a complete ensemble (e.g. the ensemble of n-butane in Figure 4) reveals two false assumptions: It implies that all visible conformations are equivalent and that the empty areas between conformations are unreachable. These problems could be eliminated if the location of atoms around rotatable bonds in a conformational ensemble were treated as probability density distributions. Computationally very expensive, accurate quantum chemistry calculations, that scale steeply and non-linearly with molecular size, might be excessive (for large datasets). Most information known about the continuous probability distribution around each torsional angle could be used in a simpler way. A more advanced conformer ensemble generator might define the movement between torsion angle peaks (defined in the torsion angle library) as density distributions. These could directly correspond to distributions found in the torsion library. In fact, a continuous torsion angle potential based solely on torsion angle peaks from the torsion angle library is already in use for the rebuilding of macrocycles by numerical optimization in Conformator (using the von Mises function as the kernel for curve approximation with a tailored equation for kappa).[4,236]

The success of the attempted advancement would require considerable adjustments and improvements to downstream programs. Even with this approach, classical conformer ensembles might stay useful for a very long time, for quick and rough computations as well as communication of (intermediate) results, simple visualization for the user and for teaching.

For now, Conformator, with its high accuracy and speed, its robustness with respect to input formats, molecular geometries, and its handling of macrocycles, represents a clear step in the right direction for describing the flexibility of small molecules with conformer ensembles. It completely closes the previously identified gap between commercial and freely available algorithms. The Sperrylite and Platinum datasets may serve as freely available datasets for future research and their compilation as an exemplary model for the generation of large high-quality datasets that enable statistically meaningful benchmarking results.

References

(1) Friedrich, N.-O.; Meyder, A.; de Bruyn Kops, C.; Sommer, K.; Flachsenberg, F.; Rarey, M.; Kirchmair, J. High-Quality Dataset of Protein-Bound Ligand Conformations and Its Application to Benchmarking Conformer Ensemble Generators. *J. Chem. Inf. Model.* **2017**, *57* (3), 529–539.

(2) Friedrich, N.-O.; de Bruyn Kops, C.; Flachsenberg, F.; Sommer, K.; Rarey, M.; Kirchmair, J. Benchmarking Commercial Conformer Ensemble Generators. *J. Chem. Inf. Model.* **2017**, *57* (11), 2719–2728.

(3) Friedrich, N.-O.; Simsir, M.; Kirchmair, J. How Diverse Are the Protein-Bound Conformations of Small-Molecule Drugs and Cofactors? *Front Chem* **2018**, *6*, 68.

(4) Friedrich, N.-O.; Flachsenberg, F.; Meyder, A.; Sommer, K.; Kirchmair, J.; Rarey, M. Conformator: A Novel Method for the Generation of Conformer Ensembles. *J. Chem. Inf. Model.* **2019**, *59* (2), 731–742.

(5) Davies, J.; Davies, D. Origins and Evolution of Antibiotic Resistance. *Microbiology and Molecular Biology Reviews*. 2010, pp 417–433.

(6) Wongsrichanalai, C.; Pickard, A. L.; Wernsdorfer, W. H.; Meshnick, S. R. Epidemiology of Drug-Resistant Malaria. *The Lancet Infectious Diseases*. 2002, pp 209–218.

(7) DiMasi, J. A.; Hansen, R. W.; Grabowski, H. G. The Price of Innovation: New Estimates of Drug Development Costs. *Journal of Health Economics*. 2003, pp 151–185.

(8) Adams, C. P.; Brantner, V. V. Estimating the Cost of New Drug Development: Is It Really 802 Million Dollars? *Health Aff.* **2006**, *25* (2), 420–428.

(9) Paul, S. M.; Mytelka, D. S.; Dunwiddie, C. T.; Persinger, C. C.; Munos, B. H.; Lindborg, S. R.; Schacht, A. L. How to Improve R&D Productivity: The Pharmaceutical Industry's Grand Challenge. *Nat. Rev. Drug Discov.* **2010**, *9* (3), 203–214.

(10) Mullard, A. New Drugs Cost US$2.6 Billion to Develop. *Nature Reviews Drug Discovery*. **2014**, pp 877–877.

(11) DiMasi, J. A.; Grabowski, H. G.; Hansen, R. W. Innovation in the Pharmaceutical Industry: New Estimates of R&D Costs. *J. Health Econ.* **2016**, *47*, 20–33.

(12) Kapetanovic, I. M. Computer-Aided Drug Discovery and Development (CADDD): In Silico-Chemico-Biological Approach. *Chem. Biol. Interact.* **2008**, *171* (2), 165–176.

(13) Yu, W.; MacKerell, A. D., Jr. Computer-Aided Drug Design Methods. *Methods Mol. Biol.* **2017**, *1520*, 85–106.

(14) Hopkins, A. L.; Groom, C. R. The Druggable Genome. *Nat. Rev. Drug Discov.* **2002**, *1* (9), 727–730.

(15) Zheng, C. J.; Han, L. Y.; Yap, C. W.; Ji, Z. L.; Cao, Z. W.; Chen, Y. Z. Therapeutic Targets: Progress of Their Exploration and Investigation of Their Characteristics. *Pharmacol. Rev.* **2006**, *58* (2), 259–279.

(16) Overington, J. P.; Al-Lazikani, B.; Hopkins, A. L. How Many Drug Targets Are There? *Nat. Rev. Drug Discov.* **2006**, *5* (12), 993–996.

(17) Rask-Andersen, M.; Almén, M. S.; Schiöth, H. B. Trends in the Exploitation of Novel Drug Targets. *Nat. Rev. Drug Discov.* **2011**, *10* (8), 579–590.

(18) Nicklaus, M. C.; Wang, S.; Driscoll, J. S.; Milne, G. W. A. Conformational Changes of Small Molecules Binding to Proteins. *Bioorganic & Medicinal Chemistry*. **1995**, pp 411–428

(19) Janin, J. Protein-Protein Recognition. *Prog. Biophys. Mol. Biol.* **1995**, *64* (2-3), 145–166.

(20) Steinbrecher, T.; Labahn, A. Towards Accurate Free Energy Calculations in Ligand Protein-Binding Studies. *Curr. Med. Chem.* **2010**, *17* (8), 767–785.

(21) Baron, R.; Setny, P.; Andrew McCammon, J. Water in Cavity−Ligand Recognition. *Journal of the American Chemical Society*. **2010**, pp 12091–12097.

(22) Schärfer, C.; Schulz-Gasch, T.; Ehrlich, H.-C.; Guba, W.; Rarey, M.; Stahl, M. Torsion Angle Preferences in Druglike Chemical Space: A Comprehensive Guide. *Journal of Medicinal Chemistry*. **2013**, pp 2016–2028.

(23) Moss, G. P. Basic Terminology of Stereochemistry (IUPAC Recommendations 1996). *Pure and Applied Chemistry*. **1996**, pp 2193–2222.

(24) Mo, Y. A Critical Analysis on the Rotation Barriers in Butane. *J. Org. Chem.* **2010**, *75* (8), 2733–2736.

(25) Stojanović, M.; Aleksić, J.; Baranac-Stojanović, M. The Effect of Steric Repulsion on the Torsional Potential of N-Butane: A Theoretical Study. *Tetrahedron*. **2015**, pp 5119–5123.

(26) Jensen, F. R.; Bushweller, C. H. Separation of Conformers. II. Axial and Equatorial Isomers of Chlorocyclohexane and Trideuteriomethoxycyclohexane. *J. Am. Chem. Soc.* **1969**, *91* (12), 3223–3225.

(27) Barton, H. R. The Conformation of the Steroid Nucleus. *Experientia* **1950**, *6* (8), 316–320.

(28) Martonosi, A. *The Enzymes of Biological Membranes: Volume 3 Membrane Transport*; Springer Science & Business Media, **2012**.

(29) Forrest, L. R.; Zhang, Y.-W.; Jacobs, M. T.; Gesmonde, J.; Xie, L.; Honig, B. H.; Rudnick, G. Mechanism for Alternating Access in Neurotransmitter Transporters. *Proc. Natl. Acad. Sci. U. S. A.* **2008**, *105* (30), 10338–10343.

(30) McMahon, B. H.; Müller, J. D.; Wraight, C. A.; Nienhaus, G. U. Electron Transfer and Protein Dynamics in the Photosynthetic Reaction Center. *Biophys. J.* **1998**, *74* (5), 2567–2587.

(31) Perola, E.; Charifson, P. S. Conformational Analysis of Drug-like Molecules Bound to Proteins: An Extensive Study of Ligand Reorganization upon Binding. *J. Med. Chem.* **2004**, *47* (10), 2499–2510.

(32) Liebeschuetz, J.; Hennemann, J.; Olsson, T.; Groom, C. R. The Good, the Bad and the Twisted: A Survey of Ligand Geometry in Protein Crystal Structures. *J. Comput. Aided Mol. Des.* **2012**, *26* (2), 169–183.

(33) Robyt, J. F. Essentials of Carbohydrate Chemistry. *Springer Advanced Texts in Chemistry*. **1998**.

(34) Brameld, K. A.; Kuhn, B.; Reuter, D. C.; Stahl, M. Small Molecule Conformational Preferences Derived from Crystal Structure Data. A Medicinal Chemistry Focused Analysis. *J. Chem. Inf. Model.* **2008**, *48* (1), 1–24.

(35) Boström, J. 10.1023/A:1015930826903. *Journal of Computer-Aided Molecular Design*. **2001**, pp 1137–1152.

(36) Seeliger, D.; de Groot, B. L. Conformational Transitions upon Ligand Binding: Holo-Structure Prediction from Apo Conformations. *PLoS Comput. Biol.* **2010**, *6* (1), e1000634.

(37) Boström, J.; Norrby, P.-O.; Liljefors, T. 10.1023/A:1008007507641. *Journal of Computer-Aided Molecular Design*. **1998**, pp 383–383.

(38) Günther, S.; Senger, C.; Michalsky, E.; Goede, A.; Preissner, R. Representation of Target-Bound Drugs by Computed Conformers: Implications for Conformational Libraries. *BMC Bioinformatics* **2006**, *7*, 293.

(39) Berman, H. M.; Westbrook, J.; Feng, Z.; Gilliland, G.; Bhat, T. N.; Weissig, H.; Shindyalov, I. N.; Bourne, P. E. The Protein Data Bank. *Nucleic Acids Res.* **2000**, *28* (1), 235–242.

(40) Groom, C. R.; Bruno, I. J.; Lightfoot, M. P.; Ward, S. C. The Cambridge Structural Database. *Acta Crystallogr B Struct Sci Cryst Eng Mater* **2016**, *72* (Pt 2), 171–179.

(41) Molecular Operating Environment (MOE), Version 2019.0102; Chemical Computing Group: Montreal, QC, **2019**.

(42) Münz, M.; Hein, J.; Biggin, P. C. The Role of Flexibility and Conformational Selection in the Binding Promiscuity of PDZ Domains. *PLoS Comput. Biol.* **2012**, *8* (11), e1002749.

(43) Peterlin, Z.; Li, Y.; Sun, G.; Shah, R.; Firestein, S.; Ryan, K. The Importance of Odorant Conformation to the Binding and Activation of a Representative Olfactory Receptor. *Chem. Biol.* **2008**, *15* (12), 1317–1327.

(44) Groom, C. R.; Allen, F. H. The Cambridge Structural Database in Retrospect and Prospect. *Angew. Chem. Int. Ed Engl.* **2014**, *53* (3), 662–671.

(45) Kuhn, B.; Guba, W.; Hert, J.; Banner, D.; Bissantz, C.; Ceccarelli, S.; Haap, W.; Körner, M.; Kuglstatter, A.; Lerner, C.; et al. A Real-World Perspective on Molecular Design. *J. Med. Chem.* **2016**, *59* (9), 4087–4102.

(46) Guba, W.; Meyder, A.; Rarey, M.; Hert, J. Torsion Library Reloaded: A New Version of Expert-Derived SMARTS Rules for Assessing Conformations of Small Molecules. *J. Chem. Inf. Model.* **2016**, *56* (1), 1–5.

(47) Leach, A. R.; Shoichet, B. K.; Peishoff, C. E. Prediction of Protein-Ligand Interactions. Docking and Scoring: Successes and Gaps. *J. Med. Chem.* **2006**, *49* (20), 5851–5855.

(48) Lorber, D. M.; Shoichet, B. K. Flexible Ligand Docking Using Conformational Ensembles. *Protein Sci.* **2008**, *7* (4), 938–950.

(49) Phillips, M. A.; Stewart, M. A.; Woodling, D. L.; Xie, Z.-R. Has Molecular Docking Ever Brought Us a Medicine? *Molecular Docking*. **2018**.

(50) Ciemny, M. P.; Debinski, A.; Paczkowska, M.; Kolinski, A.; Kurcinski, M.; Kmiecik, S. Protein-Peptide Molecular Docking with Large-Scale Conformational Changes: The p53-MDM2 Interaction. *Sci. Rep.* **2016**, *6*, 37532.

(51) Orellana, L. Large-Scale Conformational Changes and Protein Function: Breaking the in Silico Barrier. *Frontiers in Molecular Biosciences*. **2019**.

(52) Claussen, H.; Buning, C.; Rarey, M.; Lengauer, T. FlexE: Efficient Molecular Docking Considering Protein Structure Variations. *J. Mol. Biol.* **2001**, *308* (2), 377–395.

(53) Fischer, M.; Coleman, R. G.; Fraser, J. S.; Shoichet, B. K. Incorporation of Protein Flexibility and Conformational Energy Penalties in Docking Screens to Improve Ligand Discovery. *Nat. Chem.* **2014**, *6* (7), 575–583.

(54) Fischer, M.; Fraser, J. S. Predicting Protein Conformational Response in Prospective Ligand Discovery. **2014**.

(55) Antunes, D. A.; Devaurs, D.; Kavraki, L. E. Understanding the Challenges of Protein Flexibility in Drug Design. *Expert Opin. Drug Discov.* **2015**, *10* (12), 1301–1313.

(56) Anighoro, A.; de la Vega de León, A.; Bajorath, J. Predicting Bioactive Conformations and Binding Modes of Macrocycles. *J. Comput. Aided Mol. Des.* **2016**, *30* (10), 841–849.

(57) Hansch, C.; Fujita, T. P-σ-π Analysis. A Method for the Correlation of Biological Activity and Chemical Structure. *Journal of the American Chemical Society*. **1964**, pp 1616–1626.

(58) Kubinyi, H. *3D QSAR in Drug Design: Volume 1: Theory Methods and Applications*; Springer Science & Business Media, **1993**.

(59) Vedani, A.; Dobler, M. 5D-QSAR: The Key for Simulating Induced Fit? *J. Med. Chem.* **2002**, *45* (11), 2139–2149.

(60) Vedani, A.; Dobler, M.; Lill, M. A. Combining Protein Modeling and 6D-QSAR. Simulating the Binding of Structurally Diverse Ligands to the Estrogen Receptor. *J. Med. Chem.* **2005**, *48* (11), 3700–3703.

(61) Dreher, J.; Scheiber, J.; Stiefl, N.; Baumann, K. xMaP-An Interpretable Alignment-Free Four-Dimensional Quantitative Structure-Activity Relationship Technique Based on Molecular Surface Properties and Conformer Ensembles. *J. Chem. Inf. Model.* **2018**, *58* (1), 165–181.

(62) Fourches, D.; Muratov, E.; Tropsha, A. Trust, but Verify: On the Importance of Chemical Structure Curation in Cheminformatics and QSAR Modeling Research. *J. Chem. Inf. Model.* **2010**, *50* (7), 1189–1204.

(63) Vainio, M. J.; Johnson, M. S. McQSAR: A Multiconformational Quantitative Structure−Activity Relationship Engine Driven by Genetic Algorithms. *Journal of Chemical Information and Modeling*. **2005**, pp 1953–1961.

(64) Khedkar, V. M.; Joseph, J.; Pissurlenkar, R.; Saran, A.; Coutinho, E. C. How Good Are Ensembles in Improving QSAR Models? The Case witheCoRIA. *Journal of Biomolecular Structure and Dynamics*. **2015**, pp 749–769.

(65) Wolber, G.; Langer, T. LigandScout: 3-D Pharmacophores Derived from Protein-Bound Ligands and Their Use as Virtual Screening Filters. *J. Chem. Inf. Model.* **2005**, *45* (1), 160–169.

(66) Nettles, J. H.; Jenkins, J. L.; Williams, C.; Clark, A. M.; Bender, A.; Deng, Z.; Davies, J. W.; Glick, M. Flexible 3D Pharmacophores as Descriptors of Dynamic Biological Space. *J. Mol. Graph. Model.* **2007**, *26* (3), 622–633.

(67) Binns, M.; de Visser, S. P.; Theodoropoulos, C. Modeling Flexible Pharmacophores with Distance Geometry, Scoring, and Bound Stretching. *J. Chem. Inf. Model.* **2012**, *52* (2), 577–588.

(68) Taylor, R. Short Nonbonded Contact Distances in Organic Molecules and Their Use as Atom-Clash Criteria in Conformer Validation and Searching. *J. Chem. Inf. Model.* **2011**, *51* (4), 897–908.

(69) Cottrell, S. J.; Olsson, T. S. G.; Taylor, R.; Cole, J. C.; Liebeschuetz, J. W. Validating and Understanding Ring Conformations Using Small Molecule Crystallographic Data. *J. Chem. Inf. Model.* **2012**, *52* (4), 956–962.

(70) Smellie, A.; Kahn, S. D.; Teig, S. L. Analysis of Conformational Coverage. 1. Validation and Estimation of Coverage. *Journal of Chemical Information and Modeling*. **1995**, pp 285–294.

(71) Borodina, Y. V.; Bolton, E.; Fontaine, F.; Bryant, S. H. Assessment of Conformational Ensemble Sizes Necessary for Specific Resolutions of Coverage of Conformational Space. *J. Chem. Inf. Model.* **2007**, *47* (4), 1428–1437.

(72) Grimme, S. Exploration of Chemical Compound, Conformer, and Reaction Space with Meta-Dynamics Simulations Based on Tight-Binding Quantum Chemical Calculations. *J. Chem. Theory Comput.* **2019**, *15* (5), 2847–2862.

(73) Stephens, P. J.; Devlin, F. J.; Chabalowski, C. F.; Frisch, M. J. Ab Initio Calculation of Vibrational Absorption and Circular Dichroism Spectra Using Density Functional Force Fields. *The Journal of Physical Chemistry*. **1994**, pp 11623–11627.

(74) Christensen, A. S.; Kubař, T.; Cui, Q.; Elstner, M. Semiempirical Quantum Mechanical Methods for Noncovalent Interactions for Chemical and Biochemical Applications. *Chem. Rev.* **2016**, *116* (9), 5301–5337.

(75) Parr, R. G.; Weitao, Y. *Density-Functional Theory of Atoms and Molecules*; Oxford University Press, **1994**.

(76) Bannwarth, C.; Ehlert, S.; Grimme, S. 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.* **2019**, *15* (3), 1652–1671.

(77) Cavasin, A. T.; Hillisch, A.; Uellendahl, F.; Schneckener, S.; Göller, A. H. Reliable and Performant Identification of Low-Energy Conformers in the Gas Phase and Water. *Journal of Chemical Information and Modeling*. **2018**, pp 1005–1020.

(78) Pracht, P.; Bohle, F.; Grimme, S. Automated Exploration of the Low-Energy Chemical Space with Fast Quantum Chemical Methods. *Phys. Chem. Chem. Phys.* **2020**.

(79) Tosco, P.; Stiefl, N.; Landrum, G. Bringing the MMFF Force Field to the RDKit: Implementation and Validation. *J. Cheminform.* **2014**, *6* (1).

(80) Riniker, S.; Landrum, G. A. Better Informed Distance Geometry: Using What We Know To Improve Conformation Generation. *J. Chem. Inf. Model.* **2015**, *55* (12), 2562–2574.

(81) Schärfer, C.; Schulz-Gasch, T.; Hert, J.; Heinzerling, L.; Schulz, B.; Inhester, T.; Stahl, M.; Rarey, M. CONFECT: Conformations from an Expert Collection of Torsion Patterns. *ChemMedChem* **2013**, *8* (10), 1690–1700.

(82) Molecular Networks - ROTATE Classic. Https://www.mn-Am.com/products/rotate (accessed Dec 14, 2019).

(83) Klebe, G.; Mietzner, T. A Fast and Efficient Method to Generate Biologically Relevant Conformations. *J. Comput. Aided Mol. Des.* **1994**, *8* (5), 583–606.

(84) Poli, G.; Seidel, T.; Langer, T. Conformational Sampling of Small Molecules With iCon: Performance Assessment in Comparison With OMEGA. *Front Chem* **2018**, *6*, 229.

(85) Hawkins, P. C. D.; Skillman, A. G.; Warren, G. L.; Ellingson, B. A.; Stahl, M. T. Conformer Generation with OMEGA: Algorithm and Validation Using High Quality Structures from the Protein Databank and Cambridge Structural Database. *J. Chem. Inf. Model.* **2010**, *50* (4), 572–584.

(86) OMEGA, Version 2.5.1.4; OpenEye Scientific Software: Santa Fe, NM, **2017**.

(87) Halgren, T. A. MMFF VI. MMFF94s Option for Energy Minimization Studies. *Journal of Computational Chemistry*. **1999**, pp 720–729.

(88) Leite, T. B.; Gomes, D.; Miteva, M. A.; Chomilier, J.; Villoutreix, B. O.; Tufféry, P. Frog: A FRee Online druG 3D Conformation Generator. *Nucleic Acids Res.* **2007**, *35* (Web Server issue), W568–W572.

(89) Miteva, M. A.; Guyon, F.; Tufféry, P. Frog2: Efficient 3D Conformation Ensemble Generator for Small Compounds. *Nucleic Acids Res.* **2010**, *38* (Web Server issue), W622–W627.

(90) Molecular Operating Environment (MOE), Version 2016.08; Chemical Computing Group: Montreal, QC, **2017**.

(91) Macromodel, Version 2016-3; Schrödinger, LLC: New York, NY, **2016**.

(92) Davison, D. B. Distance Geometry and Molecular Conformation. G. M. Crippen, T. F. Havel. *The Quarterly Review of Biology*. **1989**, pp 487–487.

(93) Havel, T. F. Distance Geometry: Theory, Algorithms, and Chemical Applications. *Encyclopedia of Computational Chemistry*. **2002**.

(94) Liu, X.; Bai, F.; Ouyang, S.; Wang, X.; Li, H.; Jiang, H. Cyndi: A Multi-Objective Evolution Algorithm Based Method for Bioactive Molecular Conformational Generation. *BMC Bioinformatics* **2009**, *10*, 101.

(95) Lagorce, D.; Villoutreix, B. O.; Miteva, M. A. Three-Dimensional Structure Generators of Drug-like Compounds: DG-AMMOS, an Open-Source Package. *Expert Opin. Drug Discov.* **2011**, *6* (3), 339–351.

(96) Xu, H.; Izrailev, S.; Agrafiotis, D. K. Conformational Sampling by Self-Organization. *J. Chem. Inf. Comput. Sci.* **2003**, *43* (4), 1186–1191.

(97) Zhu, F.; Agrafiotis, D. K. Self-Organizing Superimposition Algorithm for Conformational Sampling. *J. Comput. Chem.* **2007**, *28* (7), 1234–1239.

(98) Crippen, Gordon & Havel, Timothy. Distance Geometry and Molecular Conformation. Chemometrics;. 15. **1988**.

(99) Vainio, M. J.; Puranen, J. S. Balloon, Version 1.5.0.1143, **2015**.

(100) Smellie, A.; Teig, S. L.; Towbin, P. Poling: Promoting Conformational Variation. *Journal of Computational Chemistry*. **1995**, pp 171–187.

(101) Li, J.; Ehlers, T.; Sutter, J.; Varma-O'Brien, S.; Kirchmair, J. CAESAR: A New Conformer Generation Algorithm Based on Recursive Buildup and Local Rotational Symmetry Consideration. *ChemInform*. **2007**.

(102) Kanal, I. Y.; Keith, J. A.; Hutchison, G. R. A Sobering Assessment of Small-Molecule Force Field Methods for Low Energy Conformer Predictions. *International Journal of Quantum Chemistry*. **2018**, p e25512.

(103) Driggers, E. M.; Hale, S. P.; Lee, J.; Terrett, N. K. The Exploration of Macrocycles for Drug Discovery — an Underexploited Structural Class. *Nature Reviews Drug Discovery*. **2008**, pp 608–624.

(104) Marsault, E.; Peterson, M. L. Macrocycles Are Great Cycles: Applications, Opportunities, and Challenges of Synthetic Macrocycles in Drug Discovery. *J. Med. Chem.* **2011**, *54* (7), 1961–2004.

(105) Levin, J. *Macrocycles in Drug Discovery*; Royal Society of Chemistry, **2014**.

(106) Abdelraheem, E. M. M.; Shaabani, S.; Dömling, A. Macrocycles: MCR Synthesis and Applications in Drug Discovery. *Drug Discov. Today Technol.* **2018**, *29*, 11–17.

(107) Lipinski, C. A.; Lombardo, F.; Dominy, B. W.; Feeney, P. J. Experimental and Computational Approaches to Estimate Solubility and Permeability in Drug Discovery and Development Settings. *Adv. Drug Deliv. Rev.* **1997**, *23* (1-3), 3–25.

(108) Doak, B. C.; Over, B.; Giordanetto, F.; Kihlberg, J. Oral Druggable Space beyond the Rule of 5: Insights from Drugs and Clinical Candidates. *Chem. Biol.* **2014**, *21* (9), 1115–1142.

(109) Chen, Y.; Garcia de Lomana, M.; Friedrich, N.-O.; Kirchmair, J. Characterization of the Chemical Space of Known and Readily Obtainable Natural Products. *J. Chem. Inf. Model.* **2018**, *58* (8), 1518–1532.

(110) Hoveyda, H. R.; Marsault, E.; Gagnon, R.; Mathieu, A. P.; Vézina, M.; Landry, A.; Wang, Z.; Benakli, K.; Beaubien, S.; Saint-Louis, C.; et al. Optimization of the Potency and Pharmacokinetic Properties of a Macrocyclic Ghrelin Receptor Agonist (Part I): Development of Ulimorelin (TZP-101) from Hit to Clinic. *J. Med. Chem.* **2011**, *54* (24), 8305–8320.

(111) Luther, A.; Moehle, K.; Chevalier, E.; Dale, G.; Obrecht, D. Protein Epitope Mimetic Macrocycles as Biopharmaceuticals. *Curr. Opin. Chem. Biol.* **2017**, *38*, 45–51.

(112) ConfGenX, Version 2016−2, Part of the Schrödinger Small-Molecule Drug Discovery Suite; Schrödinger: New York, NY, **2016**.

(113) Harder, E.; Damm, W.; Maple, J.; Wu, C.; Reboul, M.; Xiang, J. Y.; Wang, L.; Lupyan, D.; Dahlgren, M. K.; Knight, J. L.; et al. OPLS3: A Force Field Providing Broad Coverage of Drug-like Small Molecules and Proteins. *J. Chem. Theory Comput.* **2016**, *12* (1), 281–296.

(114) Cappel, D.; Dixon, S. L.; Sherman, W.; Duan, J. Exploring Conformational Search Protocols for Ligand-Based Virtual Screening and 3-D QSAR Modeling. *J. Comput. Aided Mol. Des.* **2015**, *29* (2), 165–182.

(115) Wilkes, J. G.; Stoyanova-Slavova, I. B.; Buzatu, D. A. Alignment-Independent Technique for 3D QSAR Analysis. *J. Comput. Aided Mol. Des.* **2016**, *30* (4), 331–345.

(116) Kendrew, J. C.; Bodo, G.; Dintzis, H. M.; Parrish, R. G.; Wyckoff, H.; Phillips, D. C. A Three-Dimensional Model of the Myoglobin Molecule Obtained by X-Ray Analysis. *Nature* **1958**, *181* (4610), 662–666.

(117) Read, R. J.; Adams, P. D.; Arendall, W. B., 3rd; Brunger, A. T.; Emsley, P.; Joosten, R. P.; Kleywegt, G. J.; Krissinel, E. B.; Lütteke, T.; Otwinowski, Z.; et al. A New Generation of Crystallographic Validation Tools for the Protein Data Bank. *Structure* **2011**, *19* (10), 1395–1412.

(118) Kleywegt, G. J. Validation of Protein Crystal Structures. *Acta Crystallogr. D Biol. Crystallogr.* **2000**, *56* (Pt 3), 249–265.

(119) Hawkins, P. C. D.; Warren, G. L.; Geoffrey Skillman, A.; Nicholls, A. How to Do an Evaluation: Pitfalls and Traps. *Journal of Computer-Aided Molecular Design*. **2008**, pp 179–190.

(120) Arkhipova, V.; Guskov, A.; Slotboom, D.-J. Analysis of the Quality of Crystallographic Data and the Limitations of Structural Models. *J. Gen. Physiol.* **2017**, *149* (12), 1091–1103.

(121) Tsarfati, Y.; Rosenne, S.; Weissman, H.; Shimon, L. J. W.; Gur, D.; Palmer, B. A.; Rybtchinski, B. Crystallization of Organic Molecules: Nonclassical Mechanism Revealed by Direct Imaging. *ACS Cent Sci* **2018**, *4* (8), 1031–1036.

(122) McPherson, A.; Cudney, B. Optimization of Crystallization Conditions for Biological Macromolecules. *Acta Crystallogr. Sect. F Struct. Biol. Cryst. Commun.* **2014**, *70* (Pt 11), 1445–1467.

(123) Fraser, J. S.; van den Bedem, H.; Samelson, A. J.; Lang, P. T.; Holton, J. M.; Echols, N.; Alber, T. Accessing Protein Conformational Ensembles Using Room-Temperature X-Ray Crystallography. *Proc. Natl. Acad. Sci. U. S. A.* **2011**, *108* (39), 16247–16252.

(124) Wüthrich, K. 10.1038/nsb1101-923. *Nat. Struct Biol.* November 1, **2001**, pp 923–925.

(125) Williamson, M. P.; Havel, T. F.; Wüthrich, K. Solution Conformation of Proteinase Inhibitor IIA from Bull Seminal Plasma by 1H Nuclear Magnetic Resonance and Distance Geometry. *J. Mol. Biol.* **1985**, *182* (2), 295–315.

(126) Kay, L. E. Protein Dynamics from NMR. *Nature Structural Biology*. **1998**, pp 513–517.

(127) Palmer, A. Protein Dynamics from NMR Spectroscopy and MD Simulation. *Biophysical Journal*. **2013**, p 45a.

(128) Eichmüller, C.; Skrynnikov, N. R. A New Amide Proton R1ρ Experiment Permits Accurate Characterization of Microsecond Time-Scale Conformational Exchange. *Journal of Biomolecular NMR*. **2005**, pp 281–293.

(129) Rennella, E.; Huang, R.; Velyvis, A.; Kay, L. E. 13CHD2–CEST NMR Spectroscopy Provides an Avenue for Studies of Conformational Exchange in High Molecular Weight Proteins. *Journal of Biomolecular NMR*. **2015**, pp 187–199.

(130) Cressey, D.; Callaway, E. Cryo-Electron Microscopy Wins Chemistry Nobel. *Nature* **2017**, *550* (7675), 167.

(131) Bonomi, M.; Pellarin, R.; Vendruscolo, M. Simultaneous Determination of Protein Structure and Dynamics Using Cryo-Electron Microscopy. *Biophys. J.* **2018**, *114* (7), 1604–1613.

(132) Cheng, Y.; Grigorieff, N.; Penczek, P. A.; Walz, T. A Primer to Single-Particle Cryo-Electron Microscopy. *Cell* **2015**, *161* (3), 438–449.

(133) Neumann, P.; Dickmanns, A.; Ficner, R. Validating Resolution Revolution. *Structure* **2018**, *26* (12), 1678.

(134) Hannon, A. C. Neutron Diffraction, Instrumentation. *Encyclopedia of Spectroscopy and Spectrometry*. **1999**, pp 1479–1492.

(135) Piccoli, P. M. B.; Koetzle, T. F.; Schultz, A. J. SINGLE CRYSTAL NEUTRON DIFFRACTION FOR THE INORGANIC CHEMIST – A PRACTICAL GUIDE. *Comments on Inorganic Chemistry*. **2007**, pp 3–38.

(136) Takagi, T.; Amano, M.; Tomimoto, M. Novel Method for the Evaluation of 3D Conformation Generators. *J. Chem. Inf. Model.* **2009**, *49* (6), 1377–1388.

(137) Saunders, M.; Houk, K. N.; Wu, Y. D.; Clark Still, W.; Lipton, M.; Chang, G.; Guida, W. C. Conformations of Cycloheptadecane. A Comparison of Methods for Conformational Searching. *Journal of the American Chemical Society*. **1990**, pp 1419–1427.

(138) Rhodes, G. Crystallography Made Crystal Clear, 3rd Ed.; Academic Press: San Diego, **2006**.

(139) Ghose, A. K.; Jaeger, E. P.; Kowalczyk, P. J.; Peterson, M. L.; Treasurywala, A. M. Conformational Searching Methods for Small Molecules. I. Study of the Sybyl Search Method. *J. Comput. Chem.* **1993**, *14* (9), 1050–1065.

(140) Nissink, J. W. M.; Murray, C.; Hartshorn, M.; Verdonk, M. L.; Cole, J. C.; Taylor, R. A New Test Set for Validating Predictions of Protein-Ligand Interaction. *Proteins* **2002**, *49* (4), 457–471.

(141) Diller, D. J.; Merz, K. M., Jr. Can We Separate Active from Inactive Conformations? *J. Comput. Aided Mol. Des.* **2002**, *16* (2), 105–112.

(142) Boström, J.; Greenwood, J. R.; Gottfries, J. Assessing the Performance of OMEGA with Respect to Retrieving Bioactive Conformations. *J. Mol. Graph. Model.* **2003**, *21* (5), 449–462.

(143) Kirchmair, J.; Laggner, C.; Wolber, G.; Langer, T. Comparative Analysis of Protein-Bound Ligand Conformations with Respect to Catalyst's Conformational Space Subsampling Algorithms. *J. Chem. Inf. Model.* **2005**, *45* (2), 422–430.

(144) Izrailev, S.; Zhu, F.; Agrafiotis, D. K. A Distance Geometry Heuristic for Expanding the Range of Geometries Sampled during Conformational Search. *J. Comput. Chem.* **2006**, *27* (16), 1962–1969.

(145) Kirchmair, J.; Wolber, G.; Laggner, C.; Langer, T. Comparative Performance Assessment of the Conformational Model Generators Omega and Catalyst: A Large-Scale Survey on the Retrieval of Protein-Bound Ligand Conformations. *J. Chem. Inf. Model.* **2006**, *46* (4), 1848–1861.

(146) Agrafiotis, D. K.; Gibbs, A. C.; Zhu, F.; Izrailev, S.; Martin, E. Conformational Sampling of Bioactive Molecules: A Comparative Study. *J. Chem. Inf. Model.* **2007**, *47* (3), 1067–1086.

(147) Hartshorn, M. J.; Verdonk, M. L.; Chessari, G.; Brewerton, S. C.; Mooij, W. T. M.; Mortenson, P. N.; Murray, C. W. Diverse, High-Quality Test Set for the Validation of Protein-Ligand Docking Performance. *J. Med. Chem.* **2007**, *50* (4), 726–741.

(148) Chen, I.-J.; Foloppe, N. Conformational Sampling of Druglike Molecules with MOE and Catalyst: Implications for Pharmacophore Modeling and Virtual Screening. *J. Chem. Inf. Model.* **2008**, *48* (9), 1773–1791.

(149) Bonnet, P.; Agrafiotis, D. K.; Zhu, F.; Martin, E. Conformational Analysis of Macrocycles: Finding What Common Search Methods Miss. *J. Chem. Inf. Model.* **2009**, *49* (10), 2242–2259.

(150) Bai, F.; Liu, X.; Li, J.; Zhang, H.; Jiang, H.; Wang, X.; Li, H. Bioactive Conformational Generation of Small Molecules: A Comparative Analysis between Force-Field and Multiple Empirical Criteria Based Methods. *BMC Bioinformatics* **2010**, *11*, 545.

(151) Ebejer, J.-P.; Morris, G. M.; Deane, C. M. Freely Available Conformer Generation Methods: How Good Are They? *J. Chem. Inf. Model.* **2012**, *52* (5), 1146–1158.

(152) Warren, G. L.; Do, T. D.; Kelley, B. P.; Nicholls, A.; Warren, S. D. Essential Considerations for Using Protein–ligand Structures in Drug Discovery. *Drug Discovery Today*. **2012**, pp 1270–1281.

(153) Chen, I.-J.; Foloppe, N. Tackling the Conformational Sampling of Larger Flexible Compounds and Macrocycles in Pharmacology and Drug Discovery. *Bioorg. Med. Chem.* **2013**, *21* (24), 7898–7920.

(154) Watts, K. S.; Dalal, P.; Tebben, A. J.; Cheney, D. L.; Shelley, J. C. Macrocycle Conformational Sampling with MacroModel. *J. Chem. Inf. Model.* **2014**, *54* (10), 2680–2696.

(155) Sindhikara, D.; Spronk, S. A.; Day, T.; Borrelli, K.; Cheney, D. L.; Posy, S. L. Improving Accuracy, Diversity, and Speed with Prime Macrocycle Conformational Sampling. *J. Chem. Inf. Model.* **2017**, *57* (8), 1881–1894.

(156) The Biologically Interesting Molecule Reference Dictionary (BIRD). *RCSB Protein Data Bank*. **2019**.

(157) Clark, M.; Cramer, R. D.; Van Opdenbosch, N. Validation of the General Purpose Tripos 5.2 Force Field. *Journal of Computational Chemistry*. **1989**, pp 982–1012.

(158) Mayo, S. L.; Olafson, B. D.; Goddard, W. A. DREIDING: A Generic Force Field for Molecular Simulations. *The Journal of Physical Chemistry*. **1990**, pp 8897–8909.

(159) Judson, R. S.; Jaeger, E. P.; Treasurywala, A. M.; Peterson, M. L. Conformational Searching Methods for Small Molecules. II. Genetic Algorithm Approach. *J. Comput. Chem.* **1993**, *14* (11), 1407–1414.

(160) Jaeger, E. P.; Peterson, M. L.; Treasurywala, A. M. Conformational Energy Downward Driver (CEDD): Characterization and Calibration of the Method. *J. Comput. Aided Mol. Des.* **1995**, *9* (1), 55–64.

(161) Treasurywala, A. M.; Jaeger, E. P.; Peterson, M. L. Conformational Searching Methods for Small Molecules. III. Study of Stochastic Methods Available in SYBYL and MACROMODEL. *Journal of Computational Chemistry*. **1996**, pp 1171–1182.

(162) Confort, Version 3.9; Tripos Inc.: St Louis, MO. USA (http://www.tripos.com). McMartin C, Bohacek R. J. Comput-Aided Mol Des **1995**; 11:333–342.

(163) McMartin, C.; Bohacek, R. S. QXP: Powerful, Rapid Computer Algorithms for Structure-Based Drug Design. *J. Comput. Aided Mol. Des.* **1997**, *11* (4), 333–344.

(164) Abagyan, R.; Totrov, M. Biased Probability Monte Carlo Conformational Searches and Electrostatic Calculations for Peptides and Proteins. *J. Mol. Biol.* **1994**, *235* (3), 983–1002.

(165) Dominy, B. N.; Brooks, C. L. Development of a Generalized Born Model Parametrization for Proteins and Nucleic Acids. *J. Phys. Chem. B* **1999**, *103* (18), 3765–3773.

(166) Jones, T. A.; Zou, J. Y.; Cowan, S. W.; Kjeldgaard, M. Improved Methods for Building Protein Models in Electron Density Maps and the Location of Errors in These Models. *Acta Crystallographica Section A Foundations of Crystallography*. **1991**, pp 110–119.

(167) Murshudov, G. N.; Vagin, A. A.; Dodson, E. J. Refinement of Macromolecular Structures by the Maximum-Likelihood Method. *Acta Crystallogr. D Biol. Crystallogr.* **1997**, *53* (Pt 3), 240–255.

(168) Cruickshank, D. W. J. Remarks about Protein Structure Precision. *Acta Crystallographica Section D Biological Crystallography*. **1999**, pp 583–601.

(169) Goto, J.; Kataoka, R.; Hirayama, N. Ph4Dock: Pharmacophore-Based Protein-Ligand Docking. *J. Med. Chem.* **2004**, *47* (27), 6804–6811.

(170) Poongavanam, V.; Danelius, E.; Peintner, S.; Alcaraz, L.; Caron, G.; Cummings, M. D.; Wlodek, S.; Erdelyi, M.; Hawkins, P. C. D.; Ermondi, G.; et al. Conformational Sampling of Macrocyclic Drugs in Different Environments: Can We Find the Relevant Conformations? *ACS Omega* **2018**, *3* (9), 11742–11757.

(171) The RCSB PDB Web Service Interface. Http://www.pdb.org/ Pdb/software/rest.do (accessed Feb 12, 2016).

(172) Desaphy, J.; Bret, G.; Rognan, D.; Kellenberger, E. Sc-PDB: A 3D-Database of Ligandable Binding Sites--10 Years on. *Nucleic Acids Res.* **2015**, *43* (Database issue), D399–D404.

(173) Vainio, M. J. DPICalc, **2009**.

(174) Butina, D. Unsupervised Data Base Clustering Based on Daylight's Fingerprint and Tanimoto Similarity: A Fast and Automated Way To Cluster Small and Large Data Sets. *Journal of Chemical Information and Computer Sciences*. **1999**, pp 747–750.

(175) Kleywegt, G. J.; Harris, M. R.; Zou, J. Y.; Taylor, T. C.; Wählby, A.; Jones, T. A. The Uppsala Electron-Density Server. *Acta Crystallogr. D Biol. Crystallogr.* **2004**, *60* (Pt 12 Pt 1), 2240–2249.

(176) Meyder, A.; Nittinger, E.; Lange, G.; Klein, R.; Rarey, M. Estimating Electron Density Support for Individual Atoms and Molecular Fragments in X-Ray Structures. *J. Chem. Inf. Model.* **2017**, *57* (10), 2437–2447.

(177) Nittinger, E.; Flachsenberg, F.; Bietz, S.; Lange, G.; Klein, R.; Rarey, M. Placement of Water Molecules in Protein Structures: From Large-Scale Evaluations to Single-Case Examples. *J. Chem. Inf. Model.* **2018**, *58* (8), 1625–1637.

(178) Sommer, K.; Friedrich, N.-O.; Bietz, S.; Hilbig, M.; Inhester, T.; Rarey, M. UNICON: A Powerful and Easy-to-Use Compound Library Converter. *J. Chem. Inf. Model.* **2016**, *56* (6), 1105–1111.

(179) Wishart, D. S.; Feunang, Y. D.; Guo, A. C.; Lo, E. J.; Marcu, A.; Grant, J. R.; Sajed, T.; Johnson, D.; Li, C.; Sayeeda, Z.; et al. DrugBank 5.0: A Major Update to the DrugBank Database for 2018. *Nucleic Acids Research*. **2018**, pp D1074–D1082.

(180) Meyder, A.; Kampen, S.; Sieg, J.; Fährrolfes, R.; Friedrich, N.-O.; Flachsenberg, F.; Rarey, M. StructureProfiler: An All-in-One Tool for 3D Protein Structure Profiling. *Bioinformatics*. **2019**, pp 874–876.

(181) Zentrum für Bioinformatik: Universität Hamburg - AMD Software Server https://software.zbh.uni-hamburg.de (accessed Mar 21, 2020).

(182) Gutmanas, A.; Alhroub, Y.; Battle, G. M.; Berrisford, J. M.; Bochet, E.; Conroy, M. J.; Dana, J. M.; Fernandez Montecelo, M. A.; van Ginkel, G.; Gore, S. P.; et al. PDBe: Protein Data Bank in Europe. *Nucleic Acids Res.* **2014**, *42* (Database issue), D285–D291.

(183) Fährrolfes, R.; Bietz, S.; Flachsenberg, F.; Meyder, A.; Nittinger, E.; Otto, T.; Volkamer, A.; Rarey, M. ProteinsPlus: A Web Portal for Structure Analysis of Macromolecules. *Nucleic Acids Res.* **2017**, *45* (W1), W337–W343.

(184) Cole, J. C.; Korb, O.; McCabe, P.; Read, M. G.; Taylor, R. Knowledge-Based Conformer Generation Using the Cambridge Structural Database. *J. Chem. Inf. Model.* **2018**, *58* (3), 615–629.

(185) Brock, C. P.; Minton, R. P. Systematic Effects of Crystal-Packing Forces: Biphenyl Fragments with Hydrogen Atoms in All Four Ortho Positions. *Journal of the American Chemical Society*. **1989**, pp 4586–4593.

(186) Colbert, C. L.; Agar, N. Y. R.; Kumar, P.; Chakko, M. N.; Sinha, S. C.; Powlowski, J. B.; Eltis, L. D.; Bolin, J. T. Structural Characterization of Pandoraea Pnomenusa B-356 Biphenyl Dioxygenase Reveals Features of Potent Polychlorinated Biphenyl-Degrading Enzymes. *PLoS One* **2013**, *8* (1), e52550.

(187) Wiberg, K. B. Bent Bonds in Organic Compounds. *Acc. Chem. Res.* **1996**, *29* (5), 229–234.

(188) Chen, Y.; Stork, C.; Hirte, S.; Kirchmair, J. NP-Scout: Machine Learning Approach for the Quantification and Visualization of the Natural Product-Likeness of Small Molecules. *Biomolecules* **2019**, *9* (2).

(189) Bruno, I. J.; Cole, J. C.; Kessler, M.; Luo, J.; Motherwell, W. D. S.; Purkis, L. H.; Smith, B. R.; Taylor, R.; Cooper, R. I.; Harris, S. E.; et al. Retrieval of Crystallographically-Derived Molecular Geometry Information. *J. Chem. Inf. Comput. Sci.* **2004**, *44* (6), 2133–2144.

(190) Jain, A. N.; Cleves, A. E.; Gao, Q.; Wang, X.; Liu, Y.; Sherer, E. C.; Reibarkh, M. Y. Complex Macrocycle Exploration: Parallel, Heuristic, and Constraint-Based Conformer Generation Using ForceGen. *J. Comput. Aided Mol. Des.* **2019**, *33* (6), 531–558.

(191) Allinger, N. L. Conformational Analysis. 130. MM2. A Hydrocarbon Force Field Utilizing V1 and V2 Torsional Terms. *Journal of the American Chemical Society*. **1977**, pp 8127–8134.

(192) Sander, T.; Freyss, J.; von Korff, M.; Rufener, C. DataWarrior: An Open-Source Program for Chemistry Aware Data Visualization and Analysis. *J. Chem. Inf. Model.* **2015**, *55* (2), 460–473.

(193) Wahl, J.; Freyss, J.; von Korff, M.; Sander, T. Accuracy Evaluation and Addition of Improved Dihedral Parameters for the MMFF94s. *J. Cheminform.* **2019**, *11* (1), 53.

(194) Yoshikawa, N.; Hutchison, G. R. Fast, Efficient Fragment-Based Coordinate Generation for Open Babel. *J. Cheminform.* **2019**, *11* (1), 49.

(195) Gražulis, S.; Daškevič, A.; Merkys, A.; Chateigner, D.; Lutterotti, L.; Quirós, M.; Serebryanaya, N. R.; Moeck, P.; Downs, R. T.; Le Bail, A. Crystallography Open Database (COD): An Open-Access Collection of Crystal Structures and Platform for World-Wide Collaboration. *Nucleic Acids Res.* **2012**, *40* (Database issue), D420–D427.

(196) Feng, Z.; Chen, L.; Maddula, H.; Akcan, O.; Oughtred, R.; Berman, H. M.; Westbrook, J. Ligand Depot: A Data Warehouse for Ligands Bound to Macromolecules. *Bioinformatics* **2004**, *20* (13), 2153–2155.

(197) Chan, L.; Hutchison, G. R.; Morris, G. M. BOKEI: Bayesian Optimization Using Knowledge of Correlated Torsions and Expected Improvement for Conformer Generation. *Phys. Chem. Chem. Phys.* **2020**, *22* (9), 5211–5219.

(198) CHEMBL Database Release 25. **2019**.

(199) Wang, S.; Witek, J.; Landrum, G. A.; Riniker, S. Improving Conformer Generation for Small Rings and Macrocycles Based on Distance Geometry and Experimental Torsional-Angle Preferences. *J. Chem. Inf. Model.* **2020**.

(200) O'Boyle, N. M.; Guha, R.; Willighagen, E. L.; Adams, S. E.; Alvarsson, J.; Bradley, J.-C.; Filippov, I. V.; Hanson, R. M.; Hanwell, M. D.; Hutchison, G. R.; et al. Open Data, Open Source and Open Standards in Chemistry: The Blue Obelisk Five Years on. *J. Cheminform.* **2011**, *3* (1), 37.

(201) Parks, C. D.; Gaieb, Z.; Chiu, M.; Yang, H.; Shao, C.; Walters, W. P.; Jansen, J. M.; McGaughey, G.; Lewis, R. A.; Bembenek, S. D.; et al. D3R Grand Challenge 4: Blind Prediction of Protein-Ligand Poses, Affinity Rankings, and Relative Binding Free Energies. *J. Comput. Aided Mol. Des.* **2020**, *34* (2), 99–119.

(202) Urbaczek, S.; Kolodzik, A.; Fischer, J. R.; Lippert, T.; Heuser, S.; Groth, I.; Schulz-Gasch, T.; Rarey, M. NAOMI: On the Almost Trivial Task of Reading Molecules from Different File Formats. *J. Chem. Inf. Model.* **2011**, *51* (12), 3199–3207.

(203) Hudson, B. D.; Hyde, R. M.; Rahr, E.; Wood, J.; Osman, J. Parameter Based Methods for Compound Selection from Chemical Databases. *Quantitative Structure-Activity Relationships*. **1996**, pp 285–289.

(204) Kaufman, L.; Rousseeuw, P. Clustering by Means of Medoids. In *Statistical Data Analysis Based on the L1−Norm and Related Methods*; Birkhäuser: Basel, **1987**, pp 405−416.

(205) Jin, X.; Han, J. K-Medoids Clustering. *Encyclopedia of Machine Learning and Data Mining*. **2016**, pp 1–3.

(206) Kolodzik, A.; Urbaczek, S.; Rarey, M. Unique Ring Families: A Chemically Meaningful Description of Molecular Ring Topologies. *Journal of Chemical Information and Modeling*. **2012**, pp 2013–2021.

(207) Flachsenberg, F.; Andresen, N.; Rarey, M. RingDecomposerLib: An Open-Source Implementation of Unique Ring Families and Other Cycle Bases. *Journal of Chemical Information and Modeling*. **2017**, pp 122–126.

(208) Zhu, C.; Byrd, R. H.; Lu, P.; Nocedal, J. Algorithm 778: L-BFGS-B: Fortran Subroutines for Large-Scale Bound-Constrained Optimization. *ACM Transactions on Mathematical Software*. **1997**, pp 550–560.

(209) Morales, J. L.; Nocedal, J. Remark on "algorithm 778: L-BFGS-B: Fortran Subroutines for Large-Scale Bound Constrained Optimization." *ACM Transactions on Mathematical Software*. **2011**, pp 1–4.

(210) Dutta, S.; Dimitropoulos, D.; Feng, Z.; Persikova, I.; Sen, S.; Shao, C.; Westbrook, J.; Young, J.; Zhuravleva, M. A.; Kleywegt, G. J.; et al. Improving the Representation of Peptide-like Inhibitor and Antibiotic Molecules in the Protein Data Bank. *Biopolymers* **2014**, *101* (6), 659–668.

(211) Griewel, A.; Kayser, O.; Schlosser, J.; Rarey, M. Conformational Sampling for Large-Scale Virtual Screening: Accuracy versus Ensemble Size. *J. Chem. Inf. Model.* **2009**, *49* (10), 2303–2311.

(212) O'Boyle, N. M.; Vandermeersch, T.; Flynn, C. J.; Maguire, A. R.; Hutchison, G. R. Confab - Systematic Generation of Diverse Low-Energy Conformers. *J. Cheminform.* **2011**, *3*, 8.

(213) Sauton, N.; Lagorce, D.; Villoutreix, B. O.; Miteva, M. A. MS-DOCK: Accurate Multiple Conformation Generator and Rigid Docking Protocol for Multi-Step Virtual Ligand Screening. *BMC Bioinformatics* **2008**, *9* (1), 184.

(214) Vainio, M. J.; Johnson, M. S. Generating Conformer Ensembles Using a Multiobjective Genetic Algorithm. *J. Chem. Inf. Model.* **2007**, *47* (6), 2462–2474.

(215) Watts, K. S.; Dalal, P.; Murphy, R. B.; Sherman, W.; Friesner, R. A.; Shelley, J. C. ConfGen: A Conformational Search Method for Efficient Generation of Bioactive Conformers. *J. Chem. Inf. Model.* **2010**, *50* (4), 534–546.

(216) Cxcalc, Version 15.8.31.0, Part of the Discovery Toolkit; ChemAxon: Budapest, Hungary, **2015**.

(217) Kim, S.; Bolton, E. E.; Bryant, S. H. PubChem3D: Conformer Ensemble Accuracy. *J. Cheminform.* **2013**, *5* (1), 1.

(218) Hawkins, P. C. D.; Skillman, A. G.; Nicholls, A. Comparison of Shape-Matching and Docking as Virtual Screening Tools. *J. Med. Chem.* **2007**, *50* (1), 74–82.

(219) Baber, J. C.; Thompson, D. C.; Cross, J. B.; Humblet, C. GARD: A Generally Applicable Replacement for RMSD. *J. Chem. Inf. Model.* **2009**, *49* (8), 1889–1900.

(220) Andrews, P. R.; Craik, D. J.; Martin, J. L. Functional Group Contributions to Drug-Receptor Interactions. *J. Med. Chem.* **1984**, *27* (12), 1648–1657.

(221) Schulz-Gasch, T.; Schärfer, C.; Guba, W.; Rarey, M. TFD: Torsion Fingerprints as a New Measure to Compare Small Molecule Conformations. *J. Chem. Inf. Model.* **2012**, *52* (6), 1499–1512.

(222) Hawkins, P. C. D.; Nicholls, A. Conformer Generation with OMEGA: Learning from the Data Set and the Analysis of Failures. *J. Chem. Inf. Model.* **2012**, *52* (11), 2919–2936.

(223) Holm, S. A. Simple Sequentially Rejective Multiple Test Procedure Scand. J. Stat. **1979**, 6, 65–70.

(224) RDKit: Open-Source Cheminformatics, Version 2015.09.1, **2015**.

(225) Altschul, S. F.; Gish, W.; Miller, W.; Myers, E. W.; Lipman, D. J. Basic Local Alignment Search Tool. *J. Mol. Biol.* **1990**, *215* (3), 403–410.

(226) Camacho, C.; Coulouris, G.; Avagyan, V.; Ma, N.; Papadopoulos, J.; Bealer, K.; Madden, T. L. BLAST+: Architecture and Applications. *BMC Bioinformatics* **2009**, *10*, 421.

(227) BLAST, Version 2.2.31. Https://blast.ncbi.nlm.nih.gov (accessed Jan 14, 2018).

(228) R Core Team. R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria. **2017**. https://www.R-Project.org/.

(229) Baell, J. B.; Holloway, G. A. New Substructure Filters for Removal of Pan Assay Interference Compounds (PAINS) from Screening Libraries and for Their Exclusion in Bioassays. *Journal of Medicinal Chemistry*. **2010**, pp 2719–2740.

(230) Stork, C.; Wagner, J.; Friedrich, N.-O.; de Bruyn Kops, C.; Šícho, M.; Kirchmair, J. Hit Dexter: A Machine-Learning Model for the Prediction of Frequent Hitters. *ChemMedChem* **2018**, *13* (6), 564–571.

(231) Chenoweth, D. M.; Dervan, P. B. Allosteric Modulation of DNA by Small Molecules. *Proc. Natl. Acad. Sci. U. S. A.* **2009**, *106* (32), 13175–13179.

(232) Lambert, M.; Jambon, S.; Depauw, S.; David-Cordonnier, M.-H. Targeting Transcription Factors for Cancer Treatment. *Molecules* **2018**, *23* (6).

(233) Vathy-Fogarassy, A.; Kiss, A.; Abonyi, J. Improvement of Jarvis-Patrick Clustering Based on Fuzzy Similarity. *Applications of Fuzzy Sets Theory* **2007**. pp 195–202.

(234) Ishioka, T. Extended K-Means with an Efficient Estimation of the Number of Clusters. *Intelligent Data Engineering and Automated Learning — IDEAL 2000. Data Mining, Financial Engineering, and Intelligent Agents*. **2000**, pp 17–22.

(235) Bietz, S.; Inhester, T.; Lauck, F.; Sommer, K.; von Behren, M. M.; Fährrolfes, R.; Flachsenberg, F.; Meyder, A.; Nittinger, E.; Otto, T.; et al. From Cheminformatics to Structure-Based Design: Web Services and Desktop Applications Based on the NAOMI Library. *J. Biotechnol.* **2017**, *261*, 207–214.

(236) McCabe, P.; Korb, O.; Cole, J. Kernel Density Estimation Applied to Bond Length, Bond Angle, and Torsion Angle Distributions. *J. Chem. Inf. Model.* **2014**, *54* (5), 1284–1288.

Bibliography of this Dissertation's Publications

Publications Related to This Cumulative Thesis

[D1] **Friedrich, N.-O.**; Meyder, A.; de Bruyn Kops, C.; Sommer, K.; Flachsenberg, F.; Rarey, M.; Kirchmair, J. High-Quality Dataset of Protein-Bound Ligand Conformations and Its Application to Benchmarking Conformer Ensemble Generators. *J. Chem. Inf. Model.* **2017**, *57* (3), 529–539.

[D2] **Friedrich, N.-O.**; de Bruyn Kops, C.; Flachsenberg, F.; Sommer, K.; Rarey, M.; Kirchmair, J. Benchmarking Commercial Conformer Ensemble Generators. *J. Chem. Inf. Model.* **2017**, *57* (11), 2719–2728.

[D3] **Friedrich, N.-O.**; Simsir, M.; Kirchmair, J. How Diverse Are the Protein-Bound Conformations of Small-Molecule Drugs and Cofactors? *Front Chem* **2018**, *6*, 68.

[D4] **Friedrich, N.-O.**; Flachsenberg, F.; Meyder, A.; Sommer, K.; Kirchmair, J.; Rarey, M. Conformator: A Novel Method for the Generation of Conformer Ensembles. *J. Chem. Inf. Model.* **2019**, *59* (2), 731–742.

Further Authored Publications

(178) Sommer, K.; **Friedrich, N.-O.**; Bietz, S.; Hilbig, M.; Inhester, T.; Rarey, M. UNICON: A Powerful and Easy-to-Use Compound Library Converter. *J. Chem. Inf. Model.* **2016**, *56* (6), 1105–1111.

(237) de Bruyn Kops, Ch.; **Friedrich, N.-O.**; Kirchmair, J. Alignment-Based Prediction of Sites of Metabolism. *Journal of Chemical Information and Modeling*. **2017**, pp 1258–1264.

(230) Stork, C.; Wagner, J.; **Friedrich, N.-O.**; de Bruyn Kops, C.; Šícho, M.; Kirchmair, J. Hit Dexter: A Machine-Learning Model for the Prediction of Frequent Hitters. *ChemMedChem* **2018**, *13* (6), 564–571.

(109) Chen, Y.; Garcia de Lomana, M.; **Friedrich, N.-O.**; Kirchmair, J. Characterization of the Chemical Space of Known and Readily Obtainable Natural Products. *J. Chem. Inf. Model.* **2018**, *58* (8), 1518–1532.

(180) Meyder, A.; Kampen, S.; Sieg, J.; Fährrolfes, R.; **Friedrich, N.-O.**; Flachsenberg, F.; Rarey, M. StructureProfiler: An All-in-One Tool for 3D Protein Structure Profiling. *Bioinformatics*. **2019**, pp 874–876.

Abbreviations

BFGS-B	Broyden-Fletcher-Goldfarb-Shanno-B
BIRD	Biologically Interesting Molecule Reference Dictionary
BLAST	basic local alignment search tool
CCDC	Cambridge Crystallographic Data Center
cryo-EM	cryogenic electron microscopy
CSD	Cambridge Structural Database
CT	color-Tanimoto
DFT	density functional theory
DG	distance geometry
DPI	diffraction-component precision index
ECFP	extended-connectivity fingerprint
EDIA	Electron Density scores for Individual Atoms
EDIAm	Electron Density score for Multiple Atoms
EDS	Uppsala Electron Density Server
ETKDG	Experimental-Torsion basic Knowledge Distance Geometry
FWER	familywise error rate

GARD	Generally Applicable Replacement for rmsD
LBDD	ligand based drug design
MCOS	macrocyclic optimization score
MD	molecular dynamics
MM	molecular mechanics
MoA	mode of action
MOE	Molecular Operating Environment
NMR	nuclear magnetic resonance
OPLS	optimized potentials for liquid simulations
OWAB	occupancy-weighted B-factor
PAINS	pan-assay interference compounds
PCA	principal component analysis
PDB	Protein Data Bank
QC	quantum chemical
QM	quantum mechanics
QSAR	quantitative structure-activity relationship
RCSB	Research Collaboratory for Structural Bioinformatics
RMSD	root-mean-square deviation

RSCC	real-space correlation coefficient
RSR	real-space R-value
SBDD	structure based drug design
SLE	systemic lupus erythematosus
SOS	self-organizing superimposition
SPE	stochastic proximity embedding
ST	Shape-Tanimoto
TC	TanimotoCombo
TFD	Torsion Fingerprint Deviation
vdW	van der Waals

Appendix A

Publication and congress contributions

A.1 Contributions to Publications of the Cumulative Dissertation

The following overview summarizes the authors' contributions to the individual publications of this cumulative dissertation.

[D1] **Friedrich, N.-O.**; Meyder, A.; de Bruyn Kops, C.; Sommer, K.; Flachsenberg, F.; Rarey, M.; Kirchmair, J. High-Quality Dataset of Protein-Bound Ligand Conformations and Its Application to Benchmarking Conformer Ensemble Generators. *J. Chem. Inf. Model.* **2017**, *57* (3), 529–539.

This work reports on a novel fully automated cheminformatics pipeline for compiling high-quality datasets of protein-bound ligand conformations determined by X-ray crystallography and the application of the resulting datasets to benchmarking seven freely available conformer ensemble generators. The work represents a significant leap in the development and validation of conformer ensemble generators and related technologies, which the field has been working towards for more than two decades. N.-O. Friedrich developed the concept for the different filtering methods and the automated extraction of high-quality structures from the PDB by a fully automated, elaborate cheminformatics pipeline and implemented it. The cheminformatics pipeline evaluates the support of individual atom coordinates by the measured electron density with the EDIA$_m$, that was developed and provided as a command line tool by A. Meyder. With this cheminformatics pipeline N.-O. Friedrich compiled a complete set of high-quality structures of protein-bound ligand conformations from the PDB, the Sperrylite Dataset, and the subsets thereof, the Platinum Dataset and the Platinum Diverse Dataset. N.-O. Friedrich furthermore developed and implemented the validation tool, conducted the computational studies for the benchmarking of the different conformer ensemble generators and analyzed the results. This included examination of geometrical errors in the generated ensembles. C. de Bruyn Kops developed the concept for statistical analysis of the performance tests, verified the results and contributed to the manuscript. K. Sommer and F. Flachsenberg supported the implementation of the validation tool and contributed to the manuscript. The work was supervised by J. Kirchmair and M. Rarey.

[D2] **Friedrich, N.-O.**; de Bruyn Kops, C.; Flachsenberg, F.; Sommer, K.; Rarey, M.; Kirchmair, J. Benchmarking Commercial Conformer Ensemble Generators. *J. Chem. Inf. Model.* **2017**, *57* (11), 2719–2728.

This publication compares the performance of eight commercial conformer ensemble generators and the results from ref D1. It also reports on minor improvements to the data extraction pipeline and the compilation of an updated version of the dataset based on a more recent version of the PDB, as well as different application scenarios and parametrization of algorithms for best performance. N.-O. Friedrich wrote the manuscript, developed and implemented the improvements to the cheminformatics pipeline for the compilation of the latest version of the Sperrylite and Platinum datasets. N.-O. Friedrich also conducted the computational studies for the benchmarking of the different conformer ensemble generators and analyzed the results, including identification of geometrical errors and a statistical analysis. C. de Bruyn Kops verified the results of the statistical analysis of the performance tests and contributed to the manuscript. K. Sommer and F. Flachsenberg supported the implementation of the improvements to the cheminformatics pipeline and contributed to the manuscript. The work was supervised by J. Kirchmair.

[D3] **Friedrich, N.-O.**; Simsir, M.; Kirchmair, J. How Diverse Are the Protein-Bound Conformations of Small-Molecule Drugs and Cofactors? *Front Chem* **2018**, *6*, 68.

This publication focuses on the analysis of the bioactive conformational space of a representative set of 17 approved drugs and cofactors extracted from the Sperrylite Dataset and analyzes this part of the Sperrylite Dataset in great detail. It also provides a general overview and introduction to the topic of diversity of conformations of protein-bound ligands. J. Kirchmair and N.-O. Friedrich conceived this work. N.-O. Friedrich wrote the manuscript, developed and implemented the cheminformatics pipeline for the compilation of the Sperrylite Dataset described in ref D1 (and some improvements in D2). M. Simsir and N.-O. Friedrich conducted the computational studies. M. Simsir analyzed conformational ensembles generated by OMEGA and a prototype of Conformator (developed and implemented in NAOMI by N.-O. Friedrich) with respect to diversity, energy differences and completeness during her Master thesis, supervised by J. Kirchmair and N.-O. Friedrich. Based on this work N.-O. Friedrich compiled the dataset and the subsets and analyzed the bioactive conformational space of the approved drugs and cofactors. M. Simsir generated and investigated the conformer ensembles and their superpositions and contributed the score plots of the alignments with the minimum median RMSDs derived from principal component analysis. N.-O. Friedrich implemented the algorithm to compare the best superposition of each pair of conformers and select the minimum heavy-atom RMSD ensemble, based on the RMSD calculator in NAOMI. N.-O. Friedrich generated and investigated the alignments of bioactive ligand conformers, the alignments and all-against-all sequence identity of protein structures and individual pairs, as well as the interactions of proteins and ligands in the complexes. N.-O. Friedrich wrote the manuscript; J. Kirchmair and M. Simsir contributed to the interpretation of the data and the writing of the manuscript. The work was supervised by J. Kirchmair.

[D4] **Friedrich, N.-O.**; Flachsenberg, F.; Meyder, A.; Sommer, K.; Kirchmair, J.; Rarey, M. Conformator: A Novel Method for the Generation of Conformer Ensembles. *J. Chem. Inf. Model.* **2019**, *59* (2), 731–742.

In this publication the algorithm and evaluation of the novel conformer ensemble generation method Conformator are presented. N.-O. Friedrich wrote the manuscript, developed and implemented Conformator (in NAOMI), conducted the computational studies and analyzed the results. For Conformator N.-O. Friedrich developed, among other things, a new efficient clustering algorithm, an extended set of rules for sampling torsion angles and a novel approach to sampling the conformational space of macrocycles. N.-O. Friedrich and F. Flachsenberg developed the concept for conformer ensemble generation of macrocycles. F. Flachsenberg contributed to the implementation of Conformator, especially the integration into NAOMI and the macrocycle minimization. F. Flachensberg also developed the macrocyclic optimization score introduced in this work and contributed to the manuscript. A. Meyder supported the development of the checks for geometrical errors. K. Sommer helped in the development and implementation of Conformator, especially with the handling of different file formats, treatment of special cases in the conformer generation and by supplying a template for the user interface of the command line tool. The work was supervised by J. Kirchmair and M. Rarey.

A.2 Contributions to Further Publications

This overview describes the contributions of the author of this dissertation to further publications.

(178) Sommer, K.; **Friedrich, N.-O.**; Bietz, S.; Hilbig, M.; Inhester, T.; Rarey, M. UNICON: A Powerful and Easy-to-Use Compound Library Converter. *J. Chem. Inf. Model.* **2016**, *56* (6), 1105–1111.

UNICON is a command-line tool for file conversion between standard formats commonly used in cheminformatics. It allows conversion between SDF, MOL2, SMILES, and PDB files via the generation of 2D structure coordinates and generation of 3D structures. It also facilitates the enumeration of tautomeric forms, protonation states and conformer ensembles. The conformer ensemble generation in UNICON is performed by a prototype of the conformer ensemble generator Conformator developed and implemented by N.-O. Friedrich during his master's thesis. The enhancements in the conformation ensemble generation process over the previously introduced CONFECT algorithm are shown in Table S2 of the supporting information of ref 178. N.-O. Friedrich intensively tested the command line tool during different stages of development on thousands of molecules and was involved in evaluating the results. K. Sommer developed the concept of UNICON, implemented it into NAOMI and wrote the manuscript. N.-O. Friedrich, S. Bietz, M. Hilbig and T. Inhester contributed to the manuscript. M. Rarey supervised this work.

(237) de Bruyn Kops, Ch.; **Friedrich, N.-O.**; Kirchmair, J. Alignment-Based Prediction of Sites of Metabolism. *Journal of Chemical Information and Modeling*. **2017**, pp 1258–1264.

This work presents a detailed analysis of the breadth of applicability of alignment-based site of metabolism prediction and discusses the transfer of the approach from a structure- to ligand-based method and an extension of the applicability domain. It also analyzes the effect of molecular similarity of the query and reference molecules on the prediction capability of the approach. The work combines the alignment-based method with a leading chemical reactivity model.
C. de Bruyn Kops wrote the manuscript, developed and conducted all the experiments. N.-O. Friedrich contributed to the conformer generation, the concept development and to the design of the experiments. The work was supervised by J. Kirchmair.

(230) Stork, C.; Wagner, J.; **Friedrich, N.-O.**; de Bruyn Kops, C.; Šícho, M.; Kirchmair, J. Hit Dexter: A Machine-Learning Model for the Prediction of Frequent Hitters. *ChemMedChem* **2018**, *13* (6), 564–571.

Hit Dexter is a machine learning approach that predicts frequent hitters, to allow filtering of potential PAINS and aggregators, as well as compounds with undesirable fragments. C. Stork wrote the manuscript, developed and conducted the experiments. N.-O. Friedrich contributed to concept development and to the design of the experiments. J. Wagner contributed to the development of a prototype of the machine-learning model throughout his master thesis, supervised by J. Kirchmair and N.-O. Friedrich. C. de Bruyn Kops, M. Šícho, N.-O. Friedrich and J. Kirchmair contributed to the manuscript. The work was supervised by J. Kirchmair.

(109) Chen, Y.; Garcia de Lomana, M.; **Friedrich, N.-O.**; Kirchmair, J. Characterization of the Chemical Space of Known and Readily Obtainable Natural Products. *J. Chem. Inf. Model.* **2018**, *58* (8), 1518–1532.

This publication assesses the readily available product space for natural products and devises a rule-based approach for the automated classification of natural products. For this work comprehensive data sets of known and readily obtainable natural products were compiled from 18 virtual databases (including the Dictionary of Natural Products), nine physical libraries, and the Protein Data Bank (PDB). The algorithm SugarBuster was deployed for the removal of sugars and sugar-like moieties, which are generally not of interest for drug discovery, from the natural products. The first executable version of SugarBuster was developed and implemented (in NAOMI) by N.-O. Friedrich, based on work by K. Sommer. SugarBuster was later refined, completed, and applied by M. Garcia de Lomana during her Master thesis. Y. Chen wrote the manuscript, developed and conducted the experiments. N.-O. Friedrich contributed to concept development and to the design and execution of the experiments. The work was supervised by J. Kirchmair.

(180) Meyder, A.; Kampen, S.; Sieg, J.; Fährrolfes, R.; **Friedrich, N.-O.**; Flachsenberg, F.; Rarey, M. StructureProfiler: An All-in-One Tool for 3D Protein Structure Profiling. *Bioinformatics*. **2019**, pp 874–876.

StructureProfiler is a command line tool for automated profiling of X-ray protein structures based on customizable criteria catalogues. Its “Platinum set of criteria” include most of the steps in the workflow (developed by N.-O. Friedrich and described in ref D1) for generating the Sperrylite and Platinum datasets. A. Meyder wrote the manuscript, developed and implemented StructureProfiler (in NAOMI). N.-O. Friedrich intensively tested the command line tool during different stages of development on hundreds of thousands of molecules and analyzed exceptions, oddities and special cases. This included EDIA_{m} violations, as well as inter- and intramolecular clashes detected in structures of the Platinum Dataset. M. Rarey supervised this work.

A.3 Conference Contributions

This section lists the author’s oral presentations and posters presented at national and international conferences.

Poster: **N.-O. Friedrich**, S. Sommer, M. Rarey, J. Kirchmair. A new benchmarking dataset for conformer ensemble generators. *11th German Conference on Chemoinformatic*s *(GCC)*, **2015**, Fulda, Germany

Talk: J. Kirchmair, **N.-O. Friedrich**, A. Meyder, K. Sommer, M. Rarey. Method for the automated compilation of datasets of accurate protein-bound ligand structures. *12th German Conference on Chemoinformatic*s *(GCC)*, **2016**, Fulda, Germany

Poster: N.-O. Friedrich, J. Wagner, J. Kirchmair. Prediction of compound promiscuity using machine learning algorithms. 12th German Conference on Chemoinformatics (GCC), 2016, Fulda, Germany

Poster: K. Sommer, N.-O. Friedrich, S. Bietz, M. Hilbig, T. Inhester, M. Rarey. An easy-to-use software tool for compound library conversion and isomeric enumeration, Seventh Joint Sheffield Conference on Chemoinformatics, 2016, Sheffield, GBA

Poster: N.-O. Friedrich, A. Meyder, C. de Bruyn Kops, M. Rarey, J. Kirchmair. Benchmarking Commercial Conformer Ensemble Generators. Vienna Summer School on Drug Design, 2017, Vienna, Austria

Poster: C. de Bruyn Kops, N.-O. Friedrich, Kirchmair. Prediction of Xenobiotic Sites of Metabolism: Exploring an Alignment-Based Approach. Vienna Summer School on Drug Design, 2017, Vienna, Austria

Poster: N.-O. Friedrich, A. Meyder, C. de Bruyn Kops, M. Rarey, J. Kirchmair. Benchmarking Commercial Conformer Ensemble Generators. 31st Molecular Modelling Workshop (MMWS), 2017, Erlangen, Germany

Talk: N.-O. Friedrich, M. Simsir, A. Meyder, K. Sommer, C. de Bruyn Kops, M. Rarey, J. Kirchmair. Assessment of the diversity of protein-bound ligand conformations and their representation with conformer ensembles. 32nd Molecular Modelling Workshop (MMWS), 2018, Erlangen, Germany

Appendix B

Software Architecture and Application

In this chapter, the software architecture, and the application of the conformer ensemble generator Conformator are presented. The program was developed for the NAOMI software library and implemented in C ++. Figure B1 shows the most relevant parts of the implementation of Conformator and existing basic libraries of the NAOMI platform as well as newly developed classes and functions. The complete workflow of Conformator when generating an ensemble for a molecule with a macrocyclic system is depicted in Figure S2 in the supporting information of ref D4. The figure roughly follows the workflow of the actual program and can be used as a reference point for clarity.

B.1 Libraries and Functions of Conformator

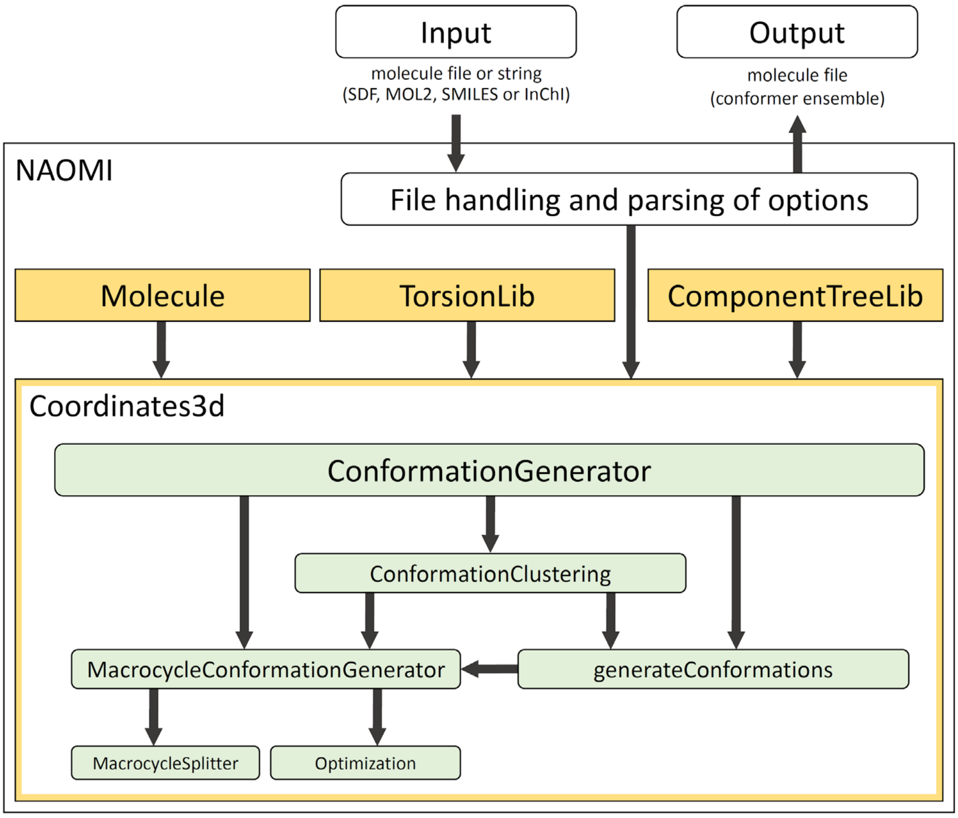


Figure B1: Simplified overview of the implementation of the conformer ensemble generator Conformator and existing basic libraries (yellow) in the NAOMI platform. The newly developed classes and functions (green) are part of the Coordinates3d library.

The main class of the Conformator algorithm ConformationGenerator is part of the NAOMI library that deals with 3D coordinates (Coordinates3d); it includes methods for general conformer generation (generateConformations), macrocycle conformer generation (MacrocycleConformerGenerator) and clustering of conformations (ConformationClustering). Precomputed, force field-optimized templates of small rings are used for sampling ring conformations (RingConformationGenerator). Conformers with geometrical errors are detected (GeometryCheckUtils) and discarded. (RingConformationGenerator and GeometryCheckUtils are also part of the Coordinates3d library and not depicted in Figure B1).

generateConformations: Generates conformations for a given molecule. The method uses the torsion library to generate torsion data (TorsionLib) for the component tree of the molecule (ComponentTreeLib). The result of this method can be influenced by the quality level and the maxim number of conformations to be returned. The quality level does influence the clustering of the conformations (e.g. maximum number of conformations to generate before initial clustering, RMSD starting threshold and enlargement per round in Å, cf. supporting information ref D4). If standard parameters are used, it resets the initial coordinates of the molecule. (New coordinates are generated with CoordinateGenerator.) The function also performs Platinum geometry and planarity checks for ring systems (including macrocycles).

ConformationClustering: Clusters conformations by RMSD to form the ensemble. The method calls functions from the Clustering library with the corresponding template arguments. The clustering method is a special case of sphere exclusion clustering and strongly depends on the order of input conformations. The cluster threshold is increased between iterations. Conformations that are no cluster centers are deleted. The cluster algorithm is described in detail and with a visual representation in Figure S1 in the supporting information of ref D4.

MacrocycleConformationGenerator: Generates conformations for macrocycles (rings formed by 10 or more atoms; definition is adjustable). Utilizes a special procedure to cut macrocycles (splitter) and a local optimization algorithm (optimizer) or the rebuilding of the macrocycles after conformer generation. The splitter function slices macrocycles by cutting bonds until no macrocycles are left. Conformations are then generated for these structures (without macrocycles) with the standard conformer generation procedure. The resulting conformations serve as starting points for the optimization. The tailored optimizer utilizes simplified force field terms for bond distortion, angle bending, and torsion energy to evaluate the deviations of molecular geometries from the ideal values and to assess steric clashes. The macrocycle conformer generation algorithm of Conformator is described in detail in ref D4 and a visual representation can be found in Figure S2 in the supporting information.

B.2 Conformator User Guide

This section explains the usage and basic program options of Conformator, the conformer ensemble generator developed for this thesis. Conformator is available as a standalone command-line tool within the NAOMI ChemBio Suite (from https://uhh.de/naomi). Conformator is a straightforward command line tool, with no setup required and can be easily implemented into a cheminformatics pipeline.

Conformator can be called in the following manner implicitly using the default values for conformer generation:

```
conformator.exe -i inputfile.smi -o outputfile.sdf
```

or with explicit configuration parameters, e.g. with quality level “Fast” (instead of the standard “Best”) and a maximum ensemble size of 50 (instead of the standard 250):

```
conformator.exe -i inputfile.smi -o outputfile.sdf -q 1 -n 50
```

From a file with multiple molecules a range of molecules to be processed can be specified, e.g. to only process molecules 3 to 8:

```
conformator.exe -i inputfile.smi -o outputfile.sdf -f 3 -t 8
```

It is also possible to include hydrogen atoms in the clash calculations (--hydrogens) and to keep initial coordinates as starting point for conformer generation (--keep3d).

Command line option	**Description**
-h [--help]	Show command line options
-v [--verbosity] arg (=3)	Set verbosity level (0 = Quiet, 1 = One-line-summary, 2 = Errors, 3 = Warnings, 4 = Info)[a]
-i [--input] arg	Input file (sdf, mol2, smi, inchi), suffix is required.
-o [--output] arg (=temp.sdf)	Output file, suffix is required.
-n [--nOfConfs] arg (=250)	Set maximum number of conformations to be generated.
-q [--quality] arg (=2)	Set quality level (1 = Fast, 2 = Best)
-f [--from] arg (=1)	Position of first entry in the calculation [start:1].
-t [--to] arg (=4294967295)	Position of last entry in the calculation.
--hydrogens	Consider hydrogen clashes during conformation generation.

--keep3d	Keep initial 3D coordinates for molecule as starting point for conformation generation.
--macrocycle_size arg (=10)	Define minimum size of macrocycles (<= 10)
--rmsd_input	Calculate the minimum RMSD of the closest ensemble member to the input structure.[b]
--rmsd_ensemble	Calculate the minimum RMSD of the closest ensemble members to each other.[b]

[a] One-line-summary prints a single line per molecule to standard out that reports on:

<infile id> <name> <nof conf> <stereo> [Error Message | ok]

Where 'stereo' can be: Absolute stereochemistry detected and preserved (INPUT), Ambiguous stereochemistry detected; single stereochemistry assigned (PURE) and Ambiguous stereochemistry detected; stereo centers enumerated in macrocycle (RACEMATE).
Info additionally includes debug output and the list of unambiguous SMILES ("USMILES") corresponding to the conformers generated.

[b] Due to symmetry correction, calculating the minimum pairwise RMSD between a generated conformer and the input conformer (--rmsd_input), or even the minimum pairwise RMSD between any generated conformers (--rmsd_ensemble), may lead to substantially longer runtimes.

B.3 Conformer Generation with RDKit

The following contains the simple python code that was used for conformer generation with RDKit. It is directly based on the recommendations and examples in the RDKit Cookbook (http://www.rdkit.org/docs/Cookbook.html).

```
#!/usr/bin/python
import os
import subprocess
import sys
import getopt

#Created on: June 25, 2015
#Author: Nils-Ole Friedrich
#RDKit conformer ensemble generation

def main(argv):
   inputfile = ''
   outputfile = ''
   nofConfs = 0
   try:
      opts, args =
            getopt.getopt(argv,"hi:o:q:n:",
            ["ifile=","ofile=","nconfs="])
   except getopt.GetoptError:
      print 'rdkit_generate_conformers.py -i <inputfile> -o
            <outputfile> -n <nofConfs>'
      sys.exit(2)
   for opt, arg in opts:
      if opt == '-h':
      print 'rdkit_generate_conformers.py -i <inputfile> -o
            <outputfile> -n <nofConfs>'
      sys.exit()
      elif opt in ("-i", "--ifile"):
      inputfile = arg
      elif opt in ("-o", "--ofile"):
      outputfile = arg
      elif opt in ("-n", "--nconfs"):
      nofConfs = int(arg)
   print 'Input file: ', inputfile
   print 'Output file: ', outputfile
   print 'nofConfs: ', nofConfs

   molname=os.path.basename(inputfile)
   os.path.splitext(molname)
   molname = os.path.splitext(molname)[0]
   print 'Input molecule: ', molname

   from rdkit import Chem
   from rdkit.Chem import AllChem
   m = Chem.MolFromMolFile(inputfile)
   m2 = Chem.AddHs(m)
   print 'generating conformers, no pruning, but optimization
     with universal force field'

#no clustering (pruning) but optimization with universal force
#field
   conformers =
     AllChem.EmbedMultipleConfs(m2, numConfs=nofConfs)

#with clustering (pruning) and optimization with universal
#force field
#conformers =
     #AllChem.EmbedMultipleConfs(m2, numConfs=nofConfs,
     #pruneRmsThresh=1.0)
#print 'generating conformers, with pruning (1 A) and
#optimization with universal force field'

   print 'optimization and writing to file'
   with open(outputfile, "a") as myoutfile:
     w = Chem.SDWriter(myoutfile)
     for id in conformers: AllChem.UFFOptimizeMolecule(m2,
          confId=id)
     for id in conformers: w.write(m2, confId=id)
     w.flush()
   print 'generation done'

if __name__ == "__main__":
  main(sys.argv[1:])
```

B.4 Conformer Generation with MOE

The following contains the SVL code used for conformer generation with MOE. The different conformer generation algorithms in MOE (Stochastic, LowModeMD and Systematic) are implemented as different “quality levels”. The code was run in batch mode, as described in the usage section of the comments.

```
//
//moe_confsearch.svl computes conformations for an SD file
//
//    created: 20.05.2014
//    last update: 22.03.2016
//
//    author: Nils-Ole Friedrich
//
//   Description:
//    Given an SD file token the function imports
//    the SD file, computes conformations and exports the
//    results back to another SD file
//
//   Usage:
//    (1) Load thus function;
//    (2) At the SVL command line, enter:
//
//   svl> moe_confsearch
//         ['input_sd_filename','output_sd_filename',
//         'pure_filename', 'number_of_conformations',
//         'algorithm']
//
//    Alternately, this file can be run from batch mode
//
//    moebatch
//          -exec "run['moe_confsearch',
//                     ['input_sd_filename',
//                     'output_sd_filename',
//                     'pure_filename',
//                     'number_of_conformations',
//                     'algorithm']]" -exit
//


#set main 'cs'

const SD_OPTIONS = [
     append:          1,    //1 to append all new records
```

```
      add_hydrogens:  1,    //1 to add hydrogens to molecules
      start_entry:    1,    //range of entries to import
      end_entry:     [],    //if null then import all entries
      file_field:     0,    //if true write field w. file name
    no_fields:        0     //import mol field only
];

function   db_ImportSD, ConfSearch, db_ExportSD,
           db_ComputeCompound;

function cs [sdfile, outfile, purefilename, nofconf, algo]

      local options = SD_OPTIONS;

      if not length outfile then
           outfile = tok_cat [fbase sdfile, '_out.sdf'];
      endif

      //algo to qualitylevel
      local qualitylevel = '1';

      write ['{}', tok_cat ['qualitylevel: ',qualitylevel,
             '\n']];
      write ['{}', tok_cat ['algo: ',algo, '\n']];

      if(algo == 'LowModeMD') then
           qualitylevel = '2';
      endif
      if(algo == 'Systematic')then
           qualitylevel = '3';
      endif

// create temp MOE database
      local dbfile = tok_cat ['/output/moe_',
                qualitylevel,'_',nofconf,'/',
                purefilename, 'temp.mdb'];
      local mdb = db_Open [dbfile, 'create'];
      db_Close mdb;

// import SD file
      write ['{}', tok_cat['Importing ', sdfile, ';...\n']];
      db_ImportSD [dbfile, sdfile, 'mol', [], [], [],
                  options];

      local ensembleoutfile = tok_cat ['/output/moe_',
                     qualitylevel,'_',nofconf,'/',
                      purefilename, 'csearch_ensemble.mdb'];


//ConfSearch call

      write ['{}', tok_cat ['Computing Conformations on ',
            sdfile, ';...\n']];

      ConfSearch [
      infile          : dbfile   // '' for current system
      ,   infile_data : 1        // copy source data fields?
      ,   infile_esel : 0        // selected entries only?
      ,   outfile     : ensembleoutfile     // output database
      ,   dbview      : 0        // open output mdb in viewer?
      ,   dbappend    : 0        // append to existing mdb?
      ,   method      : algo     // the algorithm to use
                                 // 'LowModeMD' | 'Stochastic'
                                 // | 'Systematic'
      ,   cutoff      : 7.0      // the strain energy
                                 // cutoff, default 7.0
      ,   cutoff_chi  : 1        // strain within stereo class
      ,   maxconf     : nofconf  // max number of conf
                                 // originally 10000
      ,   maxfail     : 100      // stochastic failure limit,
                                 // orig. 100
      ,   maxit       : 1000     // iteration limit,
                                 //orig. 10000
      ,   gtest       : 0.005    // energy gradient test,
                                 //default 0.005
      ,   mm_maxit    : 500      // minimization iteratio
                                 //limit, originally 500
      ,   rmsd        : 0.25     // rmsd tolerance for
                                 //duplicates, orig. 0.25
      ,   rmsd_H      : 0        // include H/LP in rmsd calc?
      ,   free_shape  : 0        // unfixed atoms only in
                                 // shape? default 0
      ,   pot_charge  : 0        // re-calculate partial
                                 // charges? default 1
      ,   invert_sp3  : 0        // invert sp3 stereo centers?
                                 // default 0
      ,   rot_amide   : 0        // rotate amide bonds?
      ,   rot_double  : 0        // rotate double bonds?
      ,   chair_only  : 1        // chair conformations only?
      ,   verbose     : 1        // write to SVL Commands
                                 // window?
      ];
```

```
// compute mol names, does not work in more recent versions
      //db_EnsureField [dbfile, 'name', 'char'];
   // db_ComputeCompound [dbfile,'mol', 'name'];

      local ensemblefile = ensembleoutfile;
      local ensemblemdb = db_Open [ensemblefile];
      db_Close ensemblemdb;

// export ensemble to output SD file
      write ['{}', tok_cat ['Exporting ', ensemblefile, ' to
           ', outfile,'...\n']];
      db_ExportSD[
           ensemblefile,
           outfile,
           first db_Fields ensemblefile,
           db_Entries ensemblefile
      ];

// delete temp MOE mdb files
      fdelete dbfile;
      fdelete ensemblefile;

      write ['{}', tok_cat ['Finished conformational search on
           ',sdfile, '\n']];
endfunction
```

Appendix C

C.1 Published Journal Articles

High-Quality Dataset of Protein-Bound Ligand Conformations and Its Application to Benchmarking Conformer Ensemble Generators

Reprinted with permission from
Friedrich, N.-O.; Meyder, A.; de Bruyn Kops, C.; Sommer, K.; Flachsenberg, F.; Rarey, M.; Kirchmair, J. High-Quality Dataset of Protein-Bound Ligand Conformations and Its Application to Benchmarking Conformer Ensemble Generators. *J. Chem. Inf. Model.* **2017**, 57 (3), 529–539.

DOI: 10.1021/acs.jcim.6b00613

Journal of
CHEMICAL INFORMATION
AND MODELING

Article

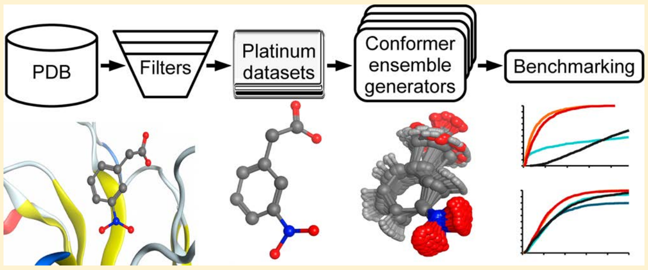
pubs.acs.org/jcim

High-Quality Dataset of Protein-Bound Ligand Conformations and Its Application to Benchmarking Conformer Ensemble Generators

Nils-Ole Friedrich, Agnes Meyder, Christina de Bruyn Kops, Kai Sommer, Florian Flachsenberg, Matthias Rarey, and Johannes Kirchmair*

University of Hamburg, ZBH − Center for Bioinformatics, Bundesstraße 43, Hamburg 20146, Germany

Ⓢ *Supporting Information*



ABSTRACT: We developed a cheminformatics pipeline for the fully automated selection and extraction of high-quality protein-bound ligand conformations from X-ray structural data. The pipeline evaluates the validity and accuracy of the 3D structures of small molecules according to multiple criteria, including their fit to the electron density and their physicochemical and structural properties. Using this approach, we compiled two high-quality datasets from the Protein Data Bank (PDB): a comprehensive dataset and a diversified subset of 4626 and 2912 structures, respectively. The datasets were applied to benchmarking seven freely available conformer ensemble generators: Balloon (two different algorithms), the RDKit standard conformer ensemble generator, the Experimental-Torsion basic Knowledge Distance Geometry (ETKDG) algorithm, Confab, Frog2 and Multiconf-DOCK. Substantial differences in the performance of the individual algorithms were observed, with RDKit and ETKDG generally achieving a favorable balance of accuracy, ensemble size and runtime. The Platinum datasets are available for download from http://www.zbh.uni-hamburg.de/platinum_dataset.

■ INTRODUCTION

Three-dimensional approaches in computer-aided molecular design such as ligand docking, pharmacophore modeling and 3D QSAR rely on the accurate representation of the protein-bound conformations of small molecules. A common approach to covering the conformational space is to compute ensembles of representative conformers. Given its importance to the field, conformer ensemble generation is a well-studied problem[1] that continues to be a current, relevant and challenging topic in cheminformatics.

Algorithms for conformer ensemble generation can be divided into two main broad categories: systematic and stochastic approaches. A large variety of methods are available today, including simulations, evolutionary algorithms, geometric distance- and knowledge-based approaches, as well as random and systematic searches. Furthermore, combined approaches exist, e.g., it is common practice to check or evaluate randomly generated conformations with force field simulations.[2,3] Computational efficacy varies dramatically among the various algorithms available to date.[4]

The geometric deviation between the experimentally determined bioactive ligand conformation(s) and the best matching, computed conformer is generally expected to correlate with ensemble size and computational cost. Naturally, large conformer ensembles are more likely to include accurate representations of protein-bound ligand conformations. However, they come at increased computational cost during generation and, more importantly, in downstream applications (e.g., 3D virtual screening). The ultimate goal is a computationally efficient algorithm, capable of producing conformational ensembles of small size that accurately represent protein-bound ligand conformations.

A wide variety of algorithms for conformer ensemble generation are available today, among them freely available tools such as Balloon,[5] RDKit,[6] Confab,[3] Frog2[7] and Multiconf-DOCK.[8] Several benchmarking studies of conformer ensemble generators have been published in recent years. Most are not comparable to each other because of the different datasets (Table 1) and protocols used. One of the earliest works compared Catalyst,[9,10] Confort,[11] Flo99,[12] MacroModel[13] and OMEGA[14] for their ability to reproduce bioactive conformations.[15] The dataset consisted of 32 ligands, and a cutoff for resolution of 2.0 Å was applied as the primary quality criterion for selection. However, resolution is not a measure of the quality of a model but the quantity of the underlying data[16] and therefore not appropriate as an exclusive quality criterion. A follow-up study examined the performance of OMEGA on a set of 36 ligands.[17]

Kirchmair et al.[18,19] published the first studies that employed datasets of several hundred structures of protein-bound ligand conformations extracted from the PDB[20] to test conformer ensemble generators, in this case two algorithms implemented in the pharmacophore modeling tool Catalyst and OMEGA. Again, the resolution of the X-ray structures served as the primary quality criterion for data selection. A related set of 918

Received: October 10, 2016
Published: February 16, 2017

DOI: 10.1021/acs.jcim.6b00613

Table 1. Overview of Datasets Used for Benchmarking Conformer Ensemble Generators[a]

Dataset name	CSD	PDB	Others	Year	Reference
Boström		32		2001	15
Original GOLD Validation (Nissink et al.)		134		2002	35
Boström et al.		36		2003	17
Perola and Charifson		100	50	2004	36
Kirchmair et al.		510		2005	18
Kirchmair et al.		778		2006	19
Astex diverse (Hartshorn et al.)		85		2007	23
Li et al.		918		2007	21
Vernalis		130		2008	24
Bai et al.		742		2010	22
Hawkins et al.	480	197		2010	14
Ebejer et al.	469	239		2012	4
Iridium-HT (Warren et al.)		121		2012	32
Riniker and Landrum	1290	238		2015	31

[a]Most of the datasets are derived from PDB structures; a few also include structures from the Cambridge Structural Database (CSD) or publically unavailable structures ("others").[34]

molecules was extracted from the PDB for the validation of CAESAR,[21] and a subset thereof was later used in a comparative analysis of methods based on molecular force fields and multiple empirical criteria.[22]

Hartshorn et al.[23] defined a high-quality dataset of 85 structures of proteins cocrystallized with drug-like compounds, which they designed for benchmarking ligand docking algorithms. Later, this dataset was used for comparing the performance of Frog and OMEGA.[7] There are also several studies in which different datasets were merged in order to improve coverage of the (mostly drug-like) chemical space. For example, Chen and Foloppe[24] combined three sets of compounds to come up with a diverse set of 130 ligand structures to compare the conformer ensemble generators implemented in MOE[25] and Catalyst. Recently, this dataset was also used for comparing the performance of Confgen, MOE, OMEGA and RDKit to BCL::Conf.[26] A significant push toward larger-sized, high-quality datasets of protein-bound ligand conformations came from Hawkins et al.[14] They systematically analyzed the insufficiencies of current approaches and defined a panel of stringent quality criteria to come up with a benchmarking dataset of 197 ligand structures extracted from the PDB. These criteria included the real-space correlation coefficient (RSCC),[27] real-space *R*-value (RSR)[28] and the diffraction-component precision index (DPI).[29] The RSR is used to detect parts of the structures where the calculated and observed electron density maps disagree. The RSCC is the correlation coefficient between those two electron density maps.[27] Both RSR and RSCC are based on matching electron density shapes. Thus, well-shaped but weak electron density can result in a misleading positive score.[30] The DPI is a global precision estimate of structure model and data quality. It takes R_{free} into account for estimating the uncertainty of atomic coordinates obtained by structural refinement of protein diffraction data. This benchmarking dataset was used to analyze the performance of OMEGA[14] and served as a basis for further benchmarking datasets and comparative studies.[4,31] The highest-grade dataset published so far is Iridium-HT.[32] It comprises 121 high-quality structures, manually selected according to a set of stringent criteria. In addition to established criteria (e.g., *R*-factor), many additional parameters were taken into account, including the completeness of the electron density in the active site. A significant effort was made in manually assigning the correct ligand topology, stereochemistry, ionization and tautomeric states, etc. The dataset was used, e.g., to evaluate the performance of the conformer ensemble generator CONFECT.[33]

Because of the significant effort involved in detecting high-quality structures of protein-bound ligands, the currently available datasets are small and therefore limit statistical analysis. In this study, we report on the development of a new cheminformatics pipeline for the fully automated compilation of high-quality datasets from the PDB. This approach not only allows for the effective compilation of large, customized datasets but also enables frequent updating. The dataset was applied to the assessment of the performance of seven current, freely available conformer ensemble generators: Balloon (two different algorithms), RDKit (two different algorithms), Confab, Frog2 and Multiconf-DOCK.

■ METHODS

Dataset Compilation. In the first step (Figure 1, step 1), the PDB web service[37] was queried for any entries matching the following criteria: (i) the electron density map is available from the Uppsala Electron Density server (EDS),[38,39] (ii) a free (i.e., noncovalently bound) ligand is present, (iii) R_{work} is lower than 0.4 and (iv) R_{free} is lower than 0.45. The definition of these criteria is in line with those of the Iridium-HT dataset.[32] All further steps were fully automated using shell and Python scripts. RDKit[40] was used for computing canonical SMILES, the number of rotatable bonds, number of heavy atoms and structural similarity, and for clustering.

Common buffer compounds, crystallization agents and metal−organic compounds were discarded (step 2) based on the lists of "unwanted ligands" and "organo-metallic complexes" obtained from the sc-PDB[41] and identified by their HET codes. Ligands lighter than 130 u and heavier than 600 u (step 3), structures originating from crystal structures with resolution of less than 2.5 Å (step 4) and covalently bound ligands were discarded (step 5) based on information available from the PDB ligand summary. Only molecules consisting of H, C, N, O, F, Si, P, S, Cl, Br and I were retained (step 6). Twenty more entries were removed; four of them lacked the required information from the PDB and 16 contained uncommon chemical bonds. Next, ligands with a minimum of 10 heavy atoms (step 7) and 1 to 16 rotatable bonds were selected (step 8).

All 24 550 PDB structures matching the above criteria were downloaded from the PDB website and the DPI was calculated with DPICalc[42] according to the definition by Goto.[43] Following the work of Hawkins et al.,[14] a maximum DPI of 0.42 Å was allowed (step 9), which leads to a maximum average positional uncertainty of 0.6 Å in the remaining structures.[29] To eliminate ligands with alternative conformations, only structures with occupancy equal to 1 for all atoms were selected (step 10).

The electron density of all ligands was examined with the Electron Density score for Multiple Atoms ($\mathrm{EDIA_m}$),[44,45] which results from the combination of the respective scores for the individual atoms (Electron Density scores for Individual Atoms, EDIA). Only ligands with $\mathrm{EDIA_m}$ greater than 0.8 were incorporated into the dataset (step 11). Electron density maps required for computing $\mathrm{EDIA_m}$ values were downloaded from

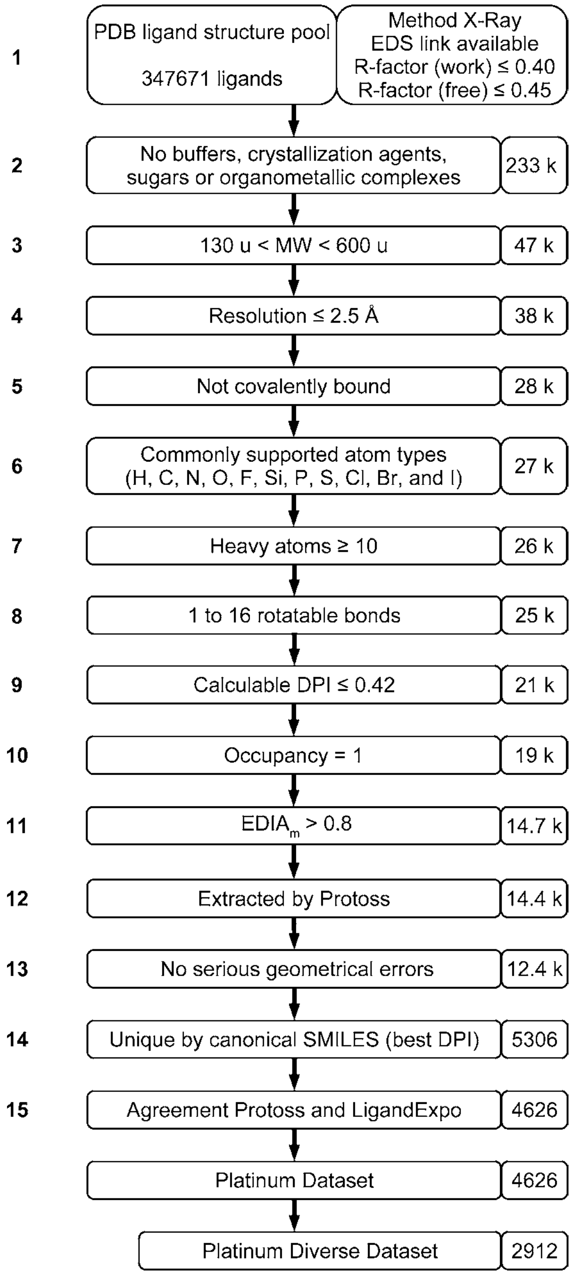


Figure 1. Overview of the cheminformatics pipeline for selecting high-quality X-ray structures of protein-bound ligand conformations from the PDB. The numbers of ligands that passed each filtering step are reported on the right. The sequence of the individual steps (indicated on the left) was optimized for short runtimes.

the EDS. The most likely protonation states and hydrogen coordinates of the ligands were determined in the protein binding pocket with Protoss[46] (step 12). In total, 148 nucleic acids or other molecules were discarded during that process. RDKit was used to generate canonical SMILES of the remaining molecules. For 182 structures, this was not possible.

The dataset was checked for molecules with geometric errors using the software library NAOMI.[47] Any molecules having at least one bond that differs by more than 0.2 Å from its ideal value[48] were removed from the dataset (step 13). Also, molecules having at least one atom angle differing by more than 12° from the VSEPR angle (16° in the case of oxygen, sulfur and phosphorus) were removed from the dataset. Corrections were made for cases like the smaller C−N−O angles in oximes that are mostly between 110° and 114°.[49]

Duplicates among the remaining 12 409 ligands were removed based on canonical SMILES (considering heavy atoms only), preserving the structure with the best DPI value (step 14). The remaining 5306 structures were checked for topological correctness by comparing the chemical structure computed with Protoss with those deposited in Ligand Expo:[50] The chemical structure of a ligand was deemed correct if the canonical SMILES representations from both sources were identical after canonization with RDKit (step 15). This was the case for 4626 ligands, and these constitute the Platinum benchmarking dataset. A diverse subset of 2912 structures (the Platinum Diverse Dataset) was selected by Butina clustering[51] with ECFP6-like Morgan fingerprints and a Tanimoto similarity cutoff of 0.5 (computed with RDKit).

Conformer Ensemble Generation. Conformer ensembles were generated with Balloon,[52] RDKit,[42] Confab[53] and Multiconf-DOCK.[54] The conformer ensemble generators were fed with a standard 3D conformation computed for each molecule with NAOMI from its SMILES notation. Ensembles comprising a maximum of 10, 50, 250 or 500 conformers were generated with default parameters. For the Balloon Genetic Algorithm (GA), the interconformer RMSD limit was set to 0.0 Å, whereas for Confab an RMSD cutoff of 0.0 Å and for Multiconf-DOCK an RMSD window of 0.0 Å were used to produce ensembles that are as close to the maximum allowed ensemble size as possible. For Frog2, force field minimization for each conformer was enabled for the generation of ensembles with a maximum of 250 or 500 conformers (the default algorithm minimizes high energy conformers only). RDKit supports the minimization of conformers with the MMFF and UFF.[6] In this work, the latter was used for all ensemble sizes.

Statistical Analysis. Reported RMSDs were calculated with NAOMI as the minimum heavy atom RMSDs measured between the reference structure and any of the computed conformers of an ensemble, considering symmetry. The significance of differences in the performance of conformer generators was determined by pairwise Mann−Whitney U tests. The significance level for each Mann−Whitney U test was adjusted according to the Bonferroni procedure for controlling the family wise error rate (FWER). Repeated runtime tests showed deviations of less than 5%.

Hardware Setup. All calculations were performed on Linux workstations running openSUSE 13.1 and equipped with Intel Xeon processors (2.2 to 2.7 GHz) and 126 GB of main memory.

RESULTS

Platinum Benchmarking Datasets. Two high-quality datasets for benchmarking conformer ensemble generators were compiled using a fully automated cheminformatics pipeline: the Platinum and Platinum Diverse datasets, consisting of 4626 and 2912 high-quality structures, respectively. The cheminformatics pipeline for dataset compilation consists of a cascade of filtering steps (Figure 1) that remove unwanted molecules (e.g., crystallization agents or organometallic complexes) and molecules not relevant to drug discovery (e.g., very small or large molecules, highly flexible molecules, molecules with uncommon atom types), structures with topological or geometrical errors and structures of low quality (e.g., with low resolution, high DPI or low EDIA$_m$). The

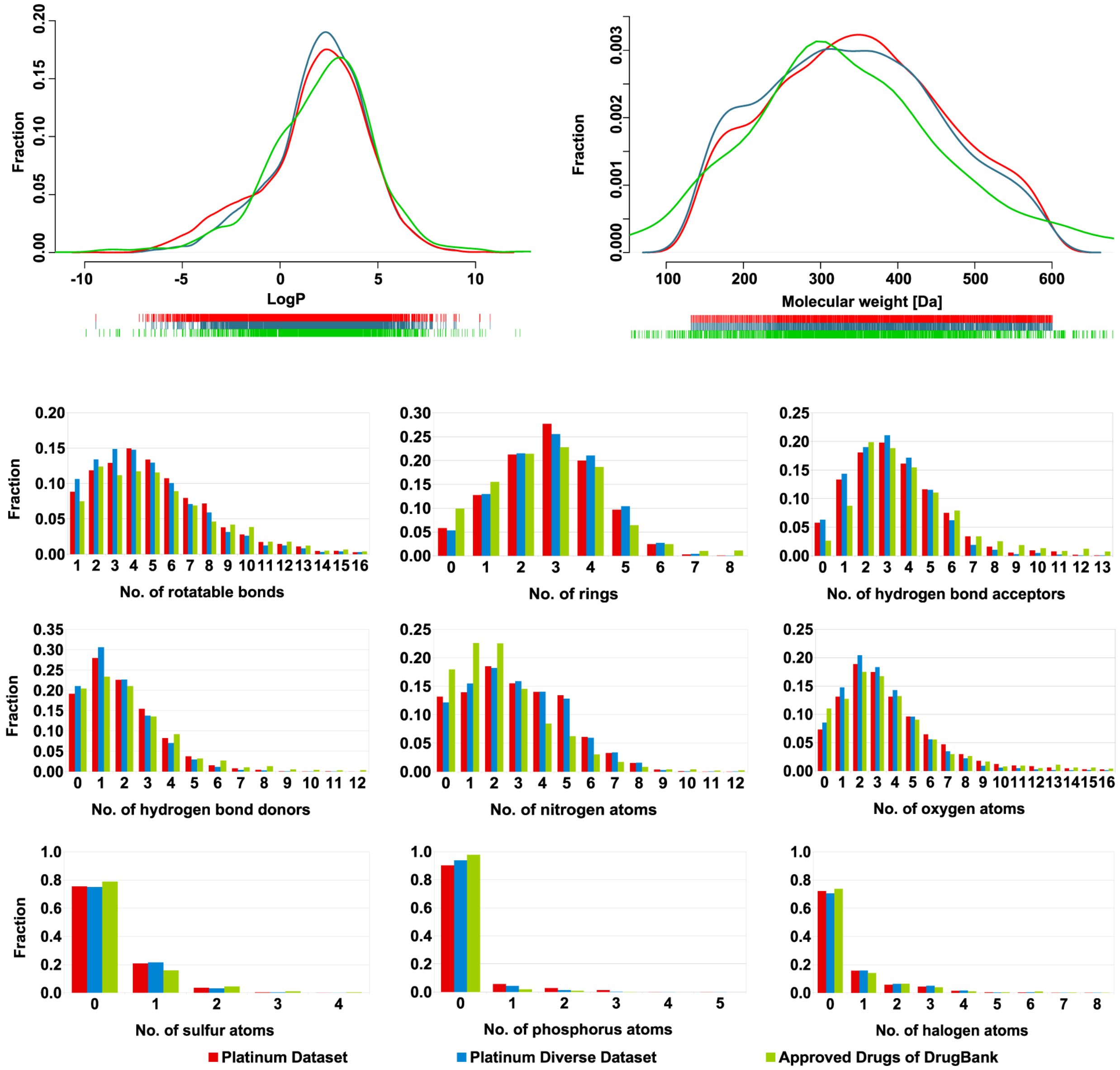


Figure 2. Distributions of molecular properties calculated for the Platinum datasets and the Approved Drugs subset of DrugBank.

EDIA is a fully automated scoring approach that evaluates the support of an atom by the electron density. It is based on the work of Nittinger et al.[55] and freely available via the Proteins*Plus* Server.[45] Previously, the compilation of datasets (e.g., the Astex, Iridium and PDBbind core set 2013[56] datasets) required the manual inspection of electron density maps due to the limitations of the RSCC and other density correlation scores.[55,57] The EDIA considers inappropriate electron density contours and electron density sphere clashes of noncovalently bound atoms by analyzing the *2fo-fc* electron density map. By combining a shape and an intensity match to compute the EDIA, the program reliably marks outliers in contrast to, e.g., the RSCC. For the quality assessment of a set of atoms, such as those comprising a ligand, the EDIA scores are combined with the help of the power mean to compute the EDIA_m. The scoring range of EDIA from 0 to 0.4 marks a structure as badly supported, 0.4 to 0.8 as mediocre supported and 0.8 to 1.2 as well supported by the experimental data.

Physicochemical Properties of the Platinum Datasets. The distributions of computed physicochemical properties among compounds of the Platinum datasets were compared with those of the compounds present in the Approved Drugs subset of DrugBank.[58] These distributions are very similar among the three datasets (Figure 2) and also the averages are within small margins (Table 2). Thus, the Platinum datasets are representative of drug-like molecules.

Performance of Freely Available Conformer Ensemble Generators. The Platinum datasets were used for benchmarking seven freely available algorithms for conformer ensemble generation: Balloon (the distance geometry and genetic algorithm), the standard conformer ensemble generator implemented in RDKit and the Experimental-Torsion basic Knowledge Distance Geometry (ETKDG) recently introduced to RDKit,[31] as well as Confab, Frog2 and Multiconf-DOCK.

Balloon generates one initial conformer by distance geometry and uses a multiobjective genetic algorithm (Balloon GA) to generate the ensemble. Torsion angles, stereochemistry of double bonds, tetrahedral chiral centers and ring conformations are modified. A postprocessing step with an MMFF94-like force field releases strain and removes duplicates and strained

DOI: 10.1021/acs.jcim.6b00613
J. Chem. Inf. Model. 2017, 57, 529−539

Table 2. Arithmetic Mean and Standard Deviation Computed for Physicochemical Properties for the Platinum Datasets and the Approved Drugs Subset of DrugBank[a]

	Platinum	Platinum Diverse	Approved Drugs of DrugBank
MW [Da]	351 ± 113	343 ± 114	386 ± 290
log *P*	1.8 ± 2.7	2.1 ± 2.4	2.1 ± 3.3
N_HBAs	3.4 ± 2.2	3.2 ± 2.0	4.7 ± 6.1
N_HBDs	1.9 ± 1.6	1.7 ± 1.5	2.7 ± 4.6
N_rot_bonds	5.1 ± 3.0	4.7 ± 2.9	5.8 ± 7.5
N_rings	2.8 ± 1.4	2.9 ± 1.5	2.8 ± 2.0

[a]Molecular weight (MW), log *P*, number of hydrogen bond acceptors (N_HBAs), hydrogen bond donors (N_HBDs), rotational bonds (N_rot_bonds) and rings (N_rings) computed with MOE.[25]

structures. Optionally, the ensemble can be generated by distance geometry only (Balloon DG).[5]

The standard conformer ensemble generator implemented in RDKit follows a distance geometry approach.[59] Based on a set of rules and the connection table of the molecule, the algorithm computes a distance bounds matrix. Random distance matrices that satisfy these bounds are used to produce atom coordinates. The resulting conformers are optionally minimized with a force field.

ETKDG is a recently developed conformer ensemble generator that was implemented in RDKit. It is based on a distance geometry approach that uses torsional-angle preferences obtained from small-molecule crystallographic data. Importantly, it also uses a chemical knowledge component that replaces force field minimization.[31]

Confab is a knowledge-based tool for conformer ensemble generation. It generates all conformers described by a set of torsion rules during a systematic search by changing torsion angles. Confab requires molecules to have at least one rotatable bond. The algorithm normally aims at building a maximum of one million conformers for a given molecule.

Frog2 is a graph-based approach in which the nodes are rings, interconnecting linkers or appendices of rings. It uses DG-AMMOS[60] to generate rings missing from the ring library. It does not address ring flexibility but includes an option for minimizing the energy of the generated conformers using AMMOS.[61] Frog2 depends on the Open Babel software package[62,63] for minimization and data conversion.[7]

Multiconf-DOCK[8] is based on an implementation of a systematic search for ligand flexibility in the program DOCK 5.[64,65] It extends multiple possible anchor segments incrementally and generates conformations by rotating all single, nonterminal, acyclic bonds in specified increments. The user can define an RMSD cutoff as well as an energy threshold relative to the initial single conformation. Multiconf-DOCK uses the Amber force field[66] in DOCK 5 to select low-energy conformers.

The performance of the individual algorithms was evaluated with respect to three key parameters: accuracy, ensemble size and computing time. An overview of the benchmarking workflow is provided in Figure 3. Ensembles consisting of a maximum of 10, 50, 250 and 500 conformers were generated. The small ensemble sizes represent use cases where speed is of essence, whereas the larger ensemble sizes represent scenarios where the accurate representation of protein-bound ligand conformations is the overriding priority.

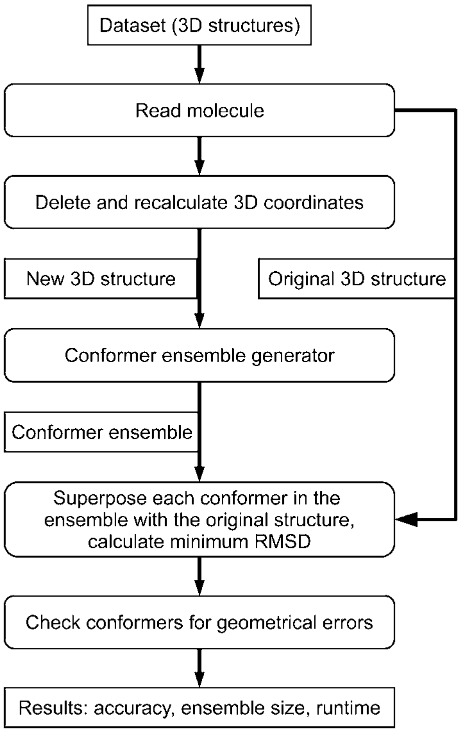

Dataset (3D structures)
Read molecule
Delete and recalculate 3D coordinates
New 3D structure
Original 3D structure
Conformer ensemble generator
Conformer ensemble
Superpose each conformer in the ensemble with the original structure, calculate minimum RMSD
Check conformers for geometrical errors
Results: accuracy, ensemble size, runtime

Figure 3. Workflow for benchmarking conformer ensemble generators.

Accuracy and Success Rates. Conformer ensemble generators are commonly evaluated with respect to their capability of reproducing the experimentally determined conformations of small molecules, primarily protein-bound ligand conformations as observed in crystal structures. The term "accuracy of a conformational ensemble" usually refers to the RMSD [Å], calculated for the best-fitting conformer of an ensemble compared to the experimentally observed conformer. The RMSD depends on the size of the molecule and is not normalized. Despite these limitations, the RMSD remains the *de facto* standard in benchmarking conformer ensemble generators. It is regarded as an objective, universal and intuitive function. RMSDs were calculated with NAOMI, which determines the minimum RMSD of each molecular pair by superposing them and enumerating all automorphisms (corresponding to chemical symmetries).

The seven algorithms were tested on the Platinum datasets. We found that the results obtained with both datasets were very similar. Hence, for the sake of clarity, we decided to report only the results obtained for the Platinum Diverse Dataset as the most representative data source. For the same reasons we focus our discussion on ensembles with a maximum of 250 conformers. The results for all investigated ensemble sizes are reported in Table 3 and Figure 4 for the Platinum Diverse Dataset.

Significant differences in the accuracy of the tested algorithms were observed (Table S1). For example, ensembles with a maximum of 250 conformers achieved mean RMSDs between 0.63 and 0.92 Å (median between 0.52 and 0.77 Å). The algorithms cluster in three groups with respect to accuracy. The top-performing group consists of Balloon GA, RDKit and ETKDG. These algorithms reproduced more than 80% of all protein-bound ligand conformations with an RMSD of less than 1 Å (Table 4). These top performers are followed by

Table 3. Arithmetic Mean and Median RMSD in Å Obtained for the Platinum Diverse Dataset[a]

	10		50		250		500	
Maximum ensemble size	Mean	Median	Mean	Median	Mean	Median	Mean	Median
Balloon DG	1.10	0.97	1.00	0.86	0.92	0.77	0.89	0.74
Balloon GA	1.22	1.10	0.90	0.80	0.72	0.63	0.67	0.58
RDKit	1.00	0.89	0.77	0.64	0.63	0.52	0.59	0.48
ETKDG	0.98	0.87	0.77	0.66	0.63	0.54	0.59	0.51
Confab	0.81	0.70	0.72	0.61	0.65	0.53	0.64	0.52
Frog2	1.18	1.19	0.93	0.85	0.75	0.65	0.77	0.67
Multiconf-DOCK	0.99	0.89	0.84	0.72	0.80	0.69	0.80	0.69

[a] Note that Confab did not produce ensembles for a large number of molecules (Table 7); therefore, its performance should not be directly compared to that of any of the other tools. This is also true for Frog2 with a maximum ensemble size of 10. Interquartile ranges are provided in Table S2.

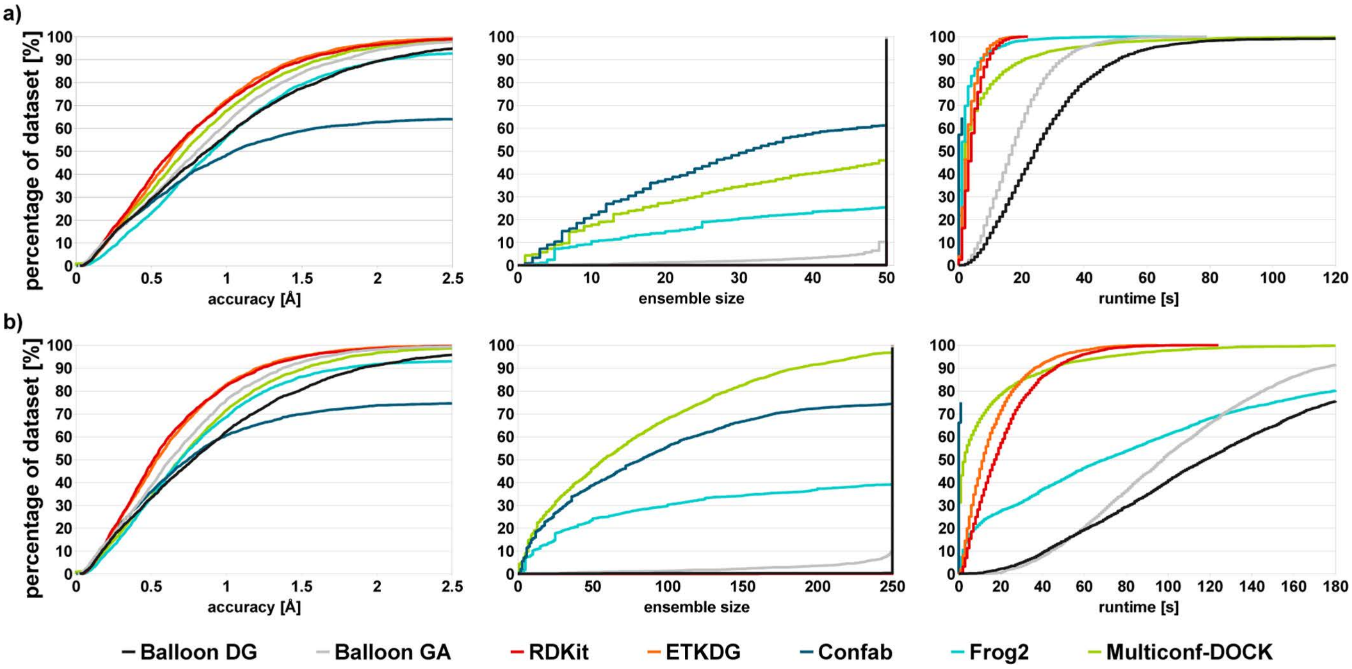


Figure 4. Percentage of protein-bound ligand conformations of the Platinum Diverse Dataset that are reproduced by the different conformer ensemble generators vs accuracy, ensemble size and runtime. Maximum ensemble size (a) 50 and (b) 250 conformations. Steeper curves indicate better performance with respect to all three criteria. The graphs reporting the ensemble sizes for Balloon DG, RDKit and ETKD overlap because all these algorithms fully exploit the maximum allowed ensemble size.

Table 4. Fraction of Structures of the Platinum Diverse Dataset Successfully Reproduced within a Specified RMSD Threshold

Maximum ensemble size	50				250			
Minimum accuracy [Å]	0.5	1.0	1.5	2.0	0.5	1.0	1.5	2.0
Balloon DG	0.29	0.57	0.77	0.92	0.33	0.62	0.81	0.92
Balloon GA	0.30	0.72	0.90	0.97	0.43	0.84	0.96	0.99
RDKit	0.39	0.71	0.89	0.96	0.48	0.82	0.95	0.98
ETKDG	0.36	0.72	0.91	0.97	0.45	0.83	0.95	0.99
Confab	0.28	0.48	0.59	0.63	0.36	0.61	0.70	0.74
Frog2	0.23	0.56	0.79	0.89	0.33	0.68	0.86	0.92
Multiconf-DOCK	0.32	0.68	0.87	0.96	0.34	0.71	0.89	0.97

Multiconf-DOCK and Frog2, of which more than 63% of ensembles satisfied the RMSD criterion. The third group is composed of Balloon DG and Confab, which produced ensembles for 56% and 57% of the ligands in this test, respectively. Confab did not produce any ensembles for a significant number of molecules.

Accuracy is a function of maximum ensemble size. For very small ensembles (maximum 10 conformers per ensemble), only 17 to 53% of the protein-bound ligand conformations were reproduced with RMSDs of less than 1 Å. With these very small ensembles, Frog2 performed substantially worse than for larger ensemble sizes (see Figures S1−S3). Ensembles consisting of a maximum of 50 conformers showed success rates of 45 to 67% with this criterion and hence nearly achieve the performance observed with ensembles of a maximum of 250 conformers (i.e., 56 to 78%). Going beyond a maximum ensemble size of 250

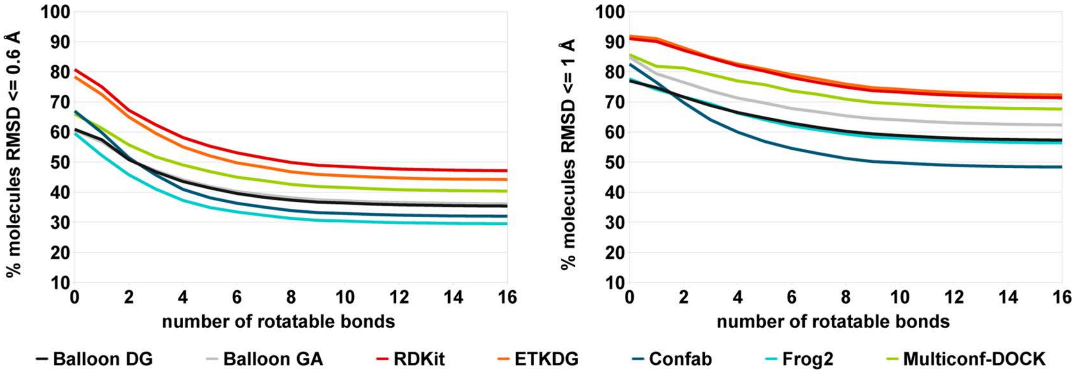


Figure 5. Percentage of molecules of the Platinum Diverse Dataset that were reproduced by the tested tools with RMSD smaller than 0.6 Å (left) and smaller than 1 Å (right) as a function of the number of rotatable bonds. The maximum ensemble size was set to 50.

Table 5. Arithmetic Mean and Median Ensemble Sizes Measured for the Platinum Diverse Dataset[a]

	10		50		250		500	
Maximum ensemble size	mean	median	mean	median	mean	median	mean	median
Balloon DG	10	10	50	50	249	250	498	500
Balloon GA	9	10	49	50	244	250	487	500
RDKit	10	10	50	50	250	250	500	500
ETKDG	10	10	50	50	250	250	500	500
Confab	6	6	20	17	65	48	109	70
Frog2	9	10	42	50	176	250	300	381
Multiconf-DOCK	9	10	36	50	78	57	80	57

[a] Note that Confab did not produce ensembles for a large number of molecules (Table 7); therefore, its performance should not be directly compared to that of any of the other tools. This is also true for Frog2 with a maximum ensemble size of 10. Interquartile ranges are provided in Table S3.

Table 6. Arithmetic Mean and Median Runtimes in Seconds Measured for the Platinum Diverse Dataset[a]

	10		50		250		500	
Maximum ensemble size	mean	median	mean	median	mean	median	mean	median
Balloon DG	6	5	27	24	132	117	260	230
Balloon GA	4	3	19	17	105	98	256	234
RDKit	1	1	5	4	22	18	42	34
ETKDG	1	1	4	3	16	12	32	23
Confab	<1	<1	<1	<1	<1	<1	<1	<1
Frog2	3	2	3	1	128	67	501	295
Multiconf-DOCK	5	1	8	2	15	3	15	3

[a] Note that Confab did not produce ensembles for a large number of molecules (Table 7); therefore, its performance should not be directly compared to that of any of the other tools. This is also true for Frog2 with a maximum ensemble size of 10. Interquartile ranges are provided in Table S4.

conformers does not yield substantial improvements for ligands with less than 12 rotatable bonds (see Figure S3).

Accuracy is also a function of molecular flexibility, which is in part reflected by the number of rotatable bonds in a molecule. Success rates drop substantially with increasing number of rotatable bonds (Figure 5). For example, at an RMSD cutoff of 1 Å, success rates of RDKit decreased by eight percentage points between molecules with two and eight rotatable bonds. The drop in performance is 16 percentage points for an RMSD cutoff of 0.6 Å, which is the guaranteed maximum positional uncertainty of any molecules in the Platinum datasets because of the DPI filter (see the Methods section). For highly flexible molecules (number of rotatable bonds greater than 8), no further drop in the success rates of any of the algorithms is observed.

Conformer Ensemble Size. Success rates and accuracy depend on ensemble size: The larger the ensembles, the higher the chance of one of the conformations fitting closely to the experimentally determined ligand conformation. For example, for RDKit the mean RMSDs improve from 1.0 to 0.59 Å (median 0.89 to 0.48 Å) by increasing the maximum number of conformers per ensemble from 10 to 500 (as measured for the Platinum Diverse Dataset; Table 3).

Although Balloon DG, Balloon GA, RDKit and ETKDG largely fill up the ensembles to the defined threshold, the size of the ensembles generated with Frog2, Confab and Multiconf-DOCK strongly depends on the input molecule. Which of the second group of algorithms produces the smallest ensembles depends on the maximum ensemble size: Confab with maximum ensemble sizes of 10, 50 and 250 (6, 20 and 66 on average, respectively) and Multiconf-DOCK with a maximum ensemble size of 500 (80 on average). Frog2, on the contrary, generates large ensembles (176 conformers on average with a maximum ensemble size of 250; Table 5). For some conformer ensemble generators (e.g., Balloon), the maximum ensemble size defined by the input parameters is not a hard limit. For the sake of comparability, we removed all supernumerary conformers that were generated after the maximum ensemble size was reached.

Runtime. The largest difference among the tested algorithms was observed for their runtimes (Table 6). Confab was measured as the fastest algorithm by far, with mean runtimes well below 1 s per conformer ensemble. Note that this algorithm, however, did not produce ensembles for a substantial subset of the test molecules. Apart from Confab, ETKDG and RDKit were fastest in producing ensembles with a maximum of 10 (ETKDG 1 s, RDKit 1 s) and 50 (ETKDG 4 s, RDKit 5 s) conformers, and Multiconf-DOCK was the fastest algorithm for generating large ensembles with a maximum of 250 and 500 conformers (15 s per ensemble).

Balloon DG was the slowest algorithm in most tests, with a mean runtime of 132 s for ensembles with a maximum of 250 conformers. Frog2 was in general very fast, with a mean runtime of 3 s per molecule for ensembles of up to 50 conformers. However, with force field minimization enabled (in this study, this was enabled for ensembles with a maximum of 250 and 500 conformers), computing times increased substantially to an average of 128 s for ensembles with a maximum of 250 conformers. In some extreme cases, computing time increased to nearly 2500 s for a single ensemble with a maximum of 250 conformers.

Processing Failures. With the exception of Frog2 and Confab, all tested algorithms produced ensembles for more than 99% of all molecules in the Platinum Diverse Dataset (Table 7). Frog2 uses the Open Babel software package. In our tests, Frog2 produced best results in combination with Open Babel version 2.3.0, which was used in this study. This algorithm produced ensembles for up to 93% of all molecules (with maximum ensemble sizes between 50 and 500). Depending on the maximum ensemble size used, Confab generated ensembles for 53 to 79% of all molecules of the Platinum Diverse Dataset. Molecules with processing failures were not included in the calculation of mean and median accuracy, ensemble size and runtime.

Table 7. Percentage of Successfully Processed Molecules for the Platinum Diverse Dataset

Maximum ensemble size	10	50	250	500
Balloon DG	99	99	99	99
Balloon GA	100	100	100	100
RDKit	99	100	100	100
ETKDG	100	100	100	100
Confab	53	65	75	79
Frog2	89	93	93	92
Multiconf-DOCK	100	100	100	100

Overall Performance. Ensembles generated with RDKit and ETKDG were of comparable quality on average, but differences in RMSDs of up to 2 Å (at maximum ensemble size 250) were observed for individual molecules (see Figure S4). Because ETKDG is up to 25% faster than the original algorithm implemented in RDKit, it is expected that ETKDG will become the default ensemble generator in this cheminformatics toolkit.

Only weak correlations were observed between accuracy and runtime: Frog2 showed the lowest correlation with an R^2 of 0.20 and RDKit the strongest with an R^2 of 0.42 (as measured for the Platinum Diverse Dataset and ensembles with a maximum of 250 conformers; see Figure S5). No correlation was observed for ensemble size and accuracy for any of the investigated algorithms (Figure S6).

Multiconf-DOCK benefits less from larger maximum ensemble sizes than the other algorithms: We measured a mean RMSD of 1.0 and 0.8 Å for ensembles with a maximum of 10 and 500 conformers, respectively (for the Platinum Diverse Dataset). One of the reasons for this difference is that Multiconf-DOCK does not exploit the maximum allowed number of conformers per ensemble. Confab produces small ensembles of high quality, faster than all other algorithms, but fails to process a substantial number of molecules of the Platinum Diverse Dataset.

We performed pairwise Mann–Whitney U tests on the complete set of RMSD values to compare the performance of the conformer generators. The significance level for each Mann–Whitney U test was adjusted to α/N according to the Bonferroni procedure for controlling the FWER, where N is the number of tests. This adjustment ensures that the FWER is less than α. Because 21 pairwise comparisons of the conformer generators were carried out, the significance level of each individual test was 0.0024 and 0.00048 for α = 0.05 and α = 0.01, respectively (Table S1). Controlling the FWER is important because the chance of making at least one Type 1 error (i.e., false positive prediction) increases with the number of statistical tests performed on the data. With 21 tests, as in this study, there is a 65.9% chance of making at least one Type 1 error at α = 0.05 unless the significance level for the individual tests is adjusted.

As hypothesized based on the accumulation curves presented in Figure 4, RDKit performs significantly better than Balloon GA, Balloon DG, Frog2 and Multiconf-DOCK according to the Mann–Whitney U test ($p < 0.00048$; Table S1), because the null hypothesis was rejected and RDKit tends to have lower RMSD values. In addition, nearly all differences in accuracy are significant at α = 0.01, with the exception of the following pairs: RDKit vs ETKDG, RDKit vs Confab, ETKDG vs Confab, and Frog2 vs Multiconf-DOCK. Any statistically significant results for Confab, however, should be considered in the context of its low success rates (Table 7).

We observed some geometrical errors regarding bond length, bond angles and planarity of aromatic rings in individual conformers of the ensembles from each tool, except Multiconf-DOCK and Confab. For example, the planarity of aromatic rings was severely disturbed in conformers of PDB Ligand ID JD5 by Balloon DG, FHC by Balloon GA and XMM by RDKit (Figure 6). A few conformers generated by RDKit for XMM also contained carbon–carbon single bonds of 1.15 instead of 1.52 Å. Large strain in the angle of an sp^3-hybridized carbon (i.e., 125.9° instead of 109.5°) was observed for a conformer of ZBF generated by ETKDG. Frog2 significantly changed the bond lengths of carbon–nitrogen triple bonds (i.e., 1.46 Å instead of 1.14 Å) for ligand 264 from 2RBN.

■ CONCLUSIONS

In this work, we report on a new cheminformatics pipeline for compiling high-quality datasets of protein-bound ligand conformations determined by X-ray crystallography. To the best of our knowledge, the two Platinum datasets derived with this pipeline are the largest publically available datasets of such high quality. Dataset size is of key importance to assuring statistical significance in detecting more subtle differences

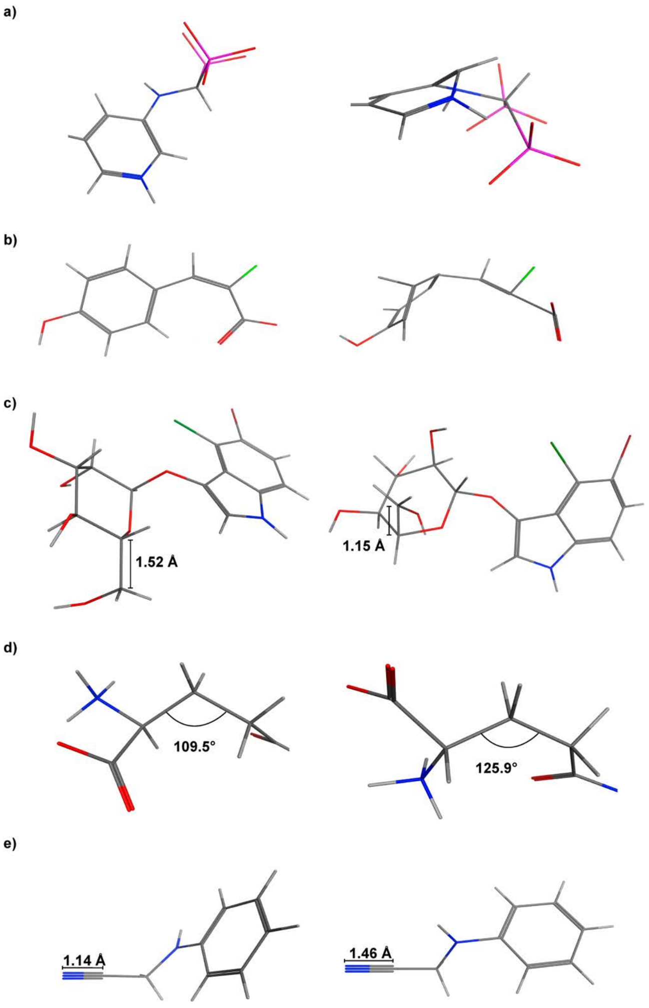


Figure 6. Examples of geometrical errors introduced by conformer ensemble generators. Input conformations are depicted on the left and computed conformers on the right. (a) Disturbed aromatic ring geometries observed with Balloon DG (JD5 from 4NFK) and (b) Balloon GA (FHC from 2OPA), (c) disturbed aromatic ring and a bond of wrong length observed with RDKit (XMM from 2JE7), (d) large strain in the angle of an sp3-hybridized carbon observed with ETKDG (ZBF from 4OPN) and (e) bond of wrong length observed with Frog2 (264 from 2RBN).

among the various algorithms, a fact which has recently been highlighted by Riniker et al. in a comparative study.[31] The Iridium-HT (121 molecules) and Hawkins datasets (197 molecules) are comparable in quality to the Platinum datasets. Our analysis of the sufficient sample size for a 99% confidence level and 5% margin of error showed that the minimum required sample size is 663 molecules. This minimum sample size is calculated for very large or unknown population size based on the standard deviation, confidence level and margin of error. Lowering the confidence level to 95% results in a necessary sample size of 384 structures. Hence, neither dataset nor any combination of the two could include enough structures for a direct comparison, let alone the same level of confidence that can be reached with the Platinum datasets.

The automated compilation procedure is based entirely on objective measures with no expert biases involved, which is not the case for any of the previously available datasets. The vast majority of compounds in the Platinum datasets are drug-like, which makes them particularly useful for testing methods employed in drug discovery. In addition, the automated procedure allows us to provide regular updates based on the latest data from the PDB.

The two datasets were presented and applied to benchmarking seven freely available conformer ensemble generators. We found significant differences in the performance of the tested algorithms. Overall, RDKit and ETKDG emerged as preferred algorithms under most tested scenarios. Some other algorithms were found to perform particularly well in more specific use cases (e.g., scenarios where very large conformer ensembles are needed or where speed is of essence). Confab and Frog2 did not produce ensembles for a substantial number of molecules. Also, we detected errors in the assignment of bond lengths, bond angles and the planarity of rings for Balloon DG, Balloon GA, RDKit, ETKDG and Frog2.

We hope that the Platinum datasets will find widespread application in the scientific community and help advance the development of conformer ensemble generators and related technologies.

■ ASSOCIATED CONTENT

Ⓢ Supporting Information

The Supporting Information is available free of charge on the ACS Publications website at DOI: 10.1021/acs.jcim.6b00613.

Additional figures and tables: percentage of protein-bound ligand conformations reproduced by the different conformer ensemble generators with different settings; percentage of molecules reproduced with RMSDs less or equal to 0.6 and 1 Å with different ensemble sizes; comparison of the accuracy of RDKit and ETKDG for individual molecules; correlation between runtime and accuracy; correlation between ensemble size and accuracy; results of the Mann−Whitney U tests; tables reporting the arithmetic mean, median and interquartile range of RMSD values, ensemble sizes and runtimes (PDF)

Text file with information on the DPI and EDIA$_m$ values for all ligands of the Platinum Dataset (TXT)

Text file with information on the DPI and EDIA$_m$ values for all ligands of the Platinum Diverse Dataset (TXT)

■ AUTHOR INFORMATION

Corresponding Author

*J. Kirchmair. E-mail: kirchmair@zbh.uni-hamburg.de. Tel.: +49 (0)40 42838 7303.

ORCID

Christina de Bruyn Kops: 0000-0001-8890-2137
Florian Flachsenberg: 0000-0001-7051-8719
Matthias Rarey: 0000-0002-9553-6531
Johannes Kirchmair: 0000-0003-2667-5877

Notes

The authors declare no competing financial interest.
The Platinum datasets are available for download from http://www.zbh.uni-hamburg.de/platinum_dataset.

■ ACKNOWLEDGMENTS

The authors thank Gerard J. Kleywegt and his team in Uppsala for providing the electron density maps for all PDB structures analyzed in the context of this work.

REFERENCES

(1) Schwab, C. H. Conformations and 3D Pharmacophore Searching. *Drug Discovery Today: Technol.* **2010**, *7*, e245−e253.

(2) Leite, T. B.; Gomes, D.; Miteva, M. A.; Chomilier, J.; Villoutreix, B. O.; Tuffery, P. Frog: A Free Online Drug 3D Conformation Generator. *Nucleic Acids Res.* **2007**, *35*, W568−W572.

(3) O'Boyle, N. M.; Vandermeersch, T.; Flynn, C. J.; Maguire, A. R.; Hutchison, G. R. Confab - Systematic Generation of Diverse Low-Energy Conformers. *J. Cheminf.* **2011**, *3*, 8.

(4) Ebejer, J.-P.; Morris, G. M.; Deane, C. M. Freely Available Conformer Generation Methods: How Good Are They? *J. Chem. Inf. Model.* **2012**, *52*, 1146−1158.

(5) Vainio, M. J.; Johnson, M. S. Generating Conformer Ensembles Using a Multiobjective Genetic Algorithm. *J. Chem. Inf. Model.* **2007**, *47*, 2462−2474.

(6) Tosco, P.; Stiefl, N.; Landrum, G. Bringing the MMFF Force Field to the RDKit: Implementation and Validation. *J. Cheminf.* **2014**, *6*, 37.

(7) Miteva, M. A.; Guyon, F.; Tuffery, P. Frog2: Efficient 3D Conformation Ensemble Generator for Small Compounds. *Nucleic Acids Res.* **2010**, *38*, W622−W627.

(8) Sauton, N.; Lagorce, D.; Villoutreix, B. O.; Miteva, M. A. MS-DOCK: Accurate Multiple Conformation Generator and Rigid Docking Protocol for Multi-Step Virtual Ligand Screening. *BMC Bioinf.* **2008**, *9*, 184.

(9) Smellie, A.; Teig, S. L.; Towbin, P. Poling: Promoting Conformational Variation. *J. Comput. Chem.* **1995**, *16*, 171−187.

(10) Smellie, A.; Kahn, S. D.; Teig, S. L. Analysis of Conformational Coverage. 1. Validation and Estimation of Coverage. *J. Chem. Inf. Model.* **1995**, *35*, 285−294.

(11) *Confort*, version 3.9; Tripos Inc.: St Louis, MO.

(12) McMartin, C.; Bohacek, R. S. QXP: Powerful, Rapid Computer Algorithms for Structure-Based Drug Design. *J. Comput.-Aided Mol. Des.* **1997**, *11*, 333−344.

(13) *Macromodel*, version 2016-3; Schrödinger, LLC: New York, NY, 2016.

(14) Hawkins, P. C. D.; Skillman, A. G.; Warren, G. L.; Ellingson, B. A.; Stahl, M. T. Conformer Generation with OMEGA: Algorithm and Validation Using High Quality Structures from the Protein Databank and Cambridge Structural Database. *J. Chem. Inf. Model.* **2010**, *50*, 572−584.

(15) Boström, J. Reproducing the Conformations of Protein-Bound Ligands: A Critical Evaluation of Several Popular Conformational Searching Tools. *J. Comput.-Aided Mol. Des.* **2001**, *15*, 1137−1152.

(16) Rhodes, G. *Crystallography Made Crystal Clear*, 3rd ed.; Academic Press: San Diego, 2006.

(17) Boström, J.; Greenwood, J. R.; Gottfries, J. Assessing the Performance of OMEGA with Respect to Retrieving Bioactive Conformations. *J. Mol. Graphics Modell.* **2003**, *21*, 449−462.

(18) Kirchmair, J.; Laggner, C.; Wolber, G.; Langer, T. Comparative Analysis of Protein-Bound Ligand Conformations with Respect to Catalyst's Conformational Space Subsampling Algorithms. *J. Chem. Inf. Model.* **2005**, *45*, 422−430.

(19) Kirchmair, J.; Wolber, G.; Laggner, C.; Langer, T. Comparative Performance Assessment of the Conformational Model Generators OMEGA and Catalyst: A Large-Scale Survey on the Retrieval of Protein-Bound Ligand Conformations. *J. Chem. Inf. Model.* **2006**, *46*, 1848−1861.

(20) Berman, H. M. The Protein Data Bank. *Nucleic Acids Res.* **2000**, *28*, 235−242.

(21) Li, J.; Ehlers, T.; Sutter, J.; Varma-O'Brien, S.; Kirchmair, J. CAESAR: A New Conformer Generation Algorithm Based on Recursive Buildup and Local Rotational Symmetry Consideration. *J. Chem. Inf. Model.* **2007**, *47*, 1923−1932.

(22) Bai, F.; Liu, X.; Li, J.; Zhang, H.; Jiang, H.; Wang, X.; Li, H. Bioactive Conformational Generation of Small Molecules: A Comparative Analysis between Force-Field and Multiple Empirical Criteria Based Methods. *BMC Bioinf.* **2010**, *11*, 545.

(23) Hartshorn, M. J.; Verdonk, M. L.; Chessari, G.; Brewerton, S. C.; Mooij, W. T. M.; Mortenson, P. N.; Murray, C. W. Diverse, High-Quality Test Set for the Validation of Protein−Ligand Docking Performance. *J. Med. Chem.* **2007**, *50*, 726−741.

(24) Chen, I. J.; Foloppe, N. Conformational Sampling of Druglike Molecules with MOE and Catalyst: Implications for Pharmacophore Modeling and Virtual Screening. *J. Chem. Inf. Model.* **2008**, *48*, 1773−1791.

(25) *Molecular Operating Environment (MOE)*, version 2013.08; Chemical Computing Group Inc.: Montreal, QC, 2013.

(26) Kothiwale, S.; Mendenhall, J. L.; Meiler, J. Bcl::Conf: Small Molecule Conformational Sampling Using a Knowledge Based Rotamer Library. *J. Cheminf.* **2015**, *7*, DOI: 10.1186/s13321-015-0095-1.

(27) Murshudov, G. N.; Vagin, A. A.; Dodson, E. J. Refinement of Macromolecular Structures by the Maximum-Likelihood Method. *Acta Crystallogr., Sect. D: Biol. Crystallogr.* **1997**, *53*, 240−255.

(28) Jones, T. A.; Zou, J. Y.; Cowan, S. W.; Kjeldgaard, M. Improved Methods for Building Protein Models in Electron Density Maps and the Location of Errors in These Models. *Acta Crystallogr., Sect. A: Found. Crystallogr.* **1991**, *47*, 110−119.

(29) Cruickshank, D. W. J. Remarks About Protein Structure Precision. *Acta Crystallogr., Sect. D: Biol. Crystallogr.* **1999**, *55*, 583−601.

(30) Read, R. J.; Adams, P. D.; Arendall, W. B.; Brunger, A. T.; Emsley, P.; Joosten, R. P.; Kleywegt, G. J.; Krissinel, E. B.; Lütteke, T.; Otwinowski, Z.; Perrakis, A.; Richardson, J. S.; Sheffler, W. H.; Smith, J. L.; Tickle, I. J.; Vriend, G.; Zwart, P. H. A New Generation of Crystallographic Validation Tools for the Protein Data Bank. *Structure* **2011**, *19*, 1395−1412.

(31) Riniker, S.; Landrum, G. A. Better Informed Distance Geometry: Using What We Know to Improve Conformation Generation. *J. Chem. Inf. Model.* **2015**, *55*, 2562−2574.

(32) Warren, G. L.; Do, T. D.; Kelley, B. P.; Nicholls, A.; Warren, S. D. Essential Considerations for Using Protein−Ligand Structures in Drug Discovery. *Drug Discovery Today* **2012**, *17*, 1270−1281.

(33) Schärfer, C.; Schulz-Gasch, T.; Hert, J.; Heinzerling, L.; Schulz, B.; Inhester, T.; Stahl, M.; Rarey, M. Confect: Conformations from an Expert Collection of Torsion Patterns. *ChemMedChem* **2013**, *8*, 1690−1700.

(34) Groom, C. R.; Bruno, I. J.; Lightfoot, M. P.; Ward, S. C. The Cambridge Structural Database. *Acta Crystallogr., Sect. B: Struct. Sci., Cryst. Eng. Mater.* **2016**, *72*, 171−179.

(35) Nissink, J. W. M.; Murray, C.; Hartshorn, M.; Verdonk, M. L.; Cole, J. C.; Taylor, R. A New Test Set for Validating Predictions of Protein-Ligand Interaction. *Proteins: Struct., Funct., Genet.* **2002**, *49*, 457−471.

(36) Perola, E.; Charifson, P. S. Conformational Analysis of Drug-Like Molecules Bound to Proteins: An Extensive Study of Ligand Reorganization Upon Binding. *J. Med. Chem.* **2004**, *47*, 2499−2510.

(37) The RCSB PDB Web Service Interface. http://www.pdb.org/pdb/software/rest.do (accessed February 12, 2016).

(38) Kleywegt, G. J.; Harris, M. R.; Zou, J.-y.; Taylor, T. C.; Wählby, A.; Jones, T. A. The Uppsala Electron-Density Server. *Acta Crystallogr., Sect. D: Biol. Crystallogr.* **2004**, *60*, 2240−2249.

(39) The Electron Density Server. http://eds.bmc.uu.se/eds (accessed February 16, 2016).

(40) *RDKit: Open-Source Cheminformatics*, version 2015.09.1, 2015.

(41) Desaphy, J.; Bret, G.; Rognan, D.; Kellenberger, E. sc-PDB: A 3D-Database of Ligandable Binding Sites - 10 Years On. *Nucleic Acids Res.* **2015**, *43*, D399−D404.

(42) Vainio, M. J. *DPICalc*, 2009.

(43) Goto, J.; Kataoka, R.; Hirayama, N. Ph4Dock: Pharmacophore-Based Protein−Ligand Docking. *J. Med. Chem.* **2004**, *47*, 6804−6811.

(44) Meyder, A.; Nittinger, E.; Lange, G.; Klein, R.; Rarey, M. EDIA: Estimating Electron Density Support for Individual Atoms in X-ray Structures. *Book of Abstracts*, 7th Joint Sheffield Conference on Chemoinformatics, Sheffield, United Kingdom, July 4−6, 2016; Molecular Graphics and Modelling Society and Chemical Structure

Association Trust: Oxford, United Kingdom and Radnor, PA, 2016; p 19.

(45) The Proteins Plus Server. http://proteinsplus.zbh.uni-hamburg.de (accessed February 1, 2017).

(46) Bietz, S.; Urbaczek, S.; Schulz, B.; Rarey, M. Protoss: A Holistic Approach to Predict Tautomers and Protonation States in Protein-Ligand Complexes. *J. Cheminf.* **2014**, *6*, 12.

(47) Urbaczek, S.; Kolodzik, A.; Fischer, J. R.; Lippert, T.; Heuser, S.; Groth, I.; Schulz-Gasch, T.; Rarey, M. NAOMI: On the Almost Trivial Task of Reading Molecules from Different File Formats. *J. Chem. Inf. Model.* **2011**, *51*, 3199−3207.

(48) Tables of Interatomic Distances and Configuration in Molecules and Ions. In *Interatomic Distances Supplement. Special Publication No. 18*; Sutton, L. E., Ed.; The Chemical Society: London, United Kingdom, 1965; pp S1s−S23s.

(49) Politzer, P.; Murray, J. S. Structural Analysis of Hydroxylamines, Oximes and Hydroxamic Acids: Trends and Patterns. In *The Chemistry of Hydroxylamines, Oximes, and Hydroxamic Acids*; Rappoport, Z.; Liebman, J. F., Eds.; Wiley: Chichester, United Kingdom, 2009; pp 29−51.

(50) Feng, Z.; Chen, L.; Maddula, H.; Akcan, O.; Oughtred, R.; Berman, H. M.; Westbrook, J. Ligand Depot: A Data Warehouse for Ligands Bound to Macromolecules. *Bioinformatics* **2004**, *20*, 2153−2155.

(51) Butina, D. Unsupervised Data Base Clustering Based on Daylight's Fingerprint and Tanimoto Similarity: A Fast and Automated Way to Cluster Small and Large Data Sets. *J. Chem. Inf. Comput. Sci.* **1999**, *39*, 747−750.

(52) Vainio, M. J.; Puranen, J. S. *Balloon*, version 1.5.0.1143, 2015.

(53) O'Boyle, N. M. *Confab*, version 1.0.1, 2011.

(54) Miteva, M.; Villoutreix, B. *Multiconf-DOCK*, version 08-04-20, 2008.

(55) Nittinger, E.; Schneider, N.; Lange, G.; Rarey, M. Evidence of Water Molecules - A Statistical Evaluation of Water Molecules Based on Electron Density. *J. Chem. Inf. Model.* **2015**, *55*, 771−783.

(56) Li, Y.; Liu, Z.; Li, J.; Han, L.; Liu, J.; Zhao, Z.; Wang, R. Comparative Assessment of Scoring Functions on an Updated Benchmark: 1. Compilation of the Test Set. *J. Chem. Inf. Model.* **2014**, *54*, 1700−1716.

(57) Hawkins, P. C. D.; Kelley, B. P.; Warren, G. L. The Application of Statistical Methods to Cognate Docking: A Path Forward? *J. Chem. Inf. Model.* **2014**, *54*, 1339−1355.

(58) Wishart, D. S. Drugbank: A Comprehensive Resource for in Silico Drug Discovery and Exploration. *Nucleic Acids Res.* **2006**, *34*, D668−D672.

(59) Blaney, J. M., Dixon, J. S. Distance Geometry in Molecular Modeling. In *Reviews in Computational Chemistry*; Lipkowitz, K. B.; Boyd, D. B., Eds.; John Wiley & Sons, Inc.: Hoboken, NJ, 2007; Vol. *5*, pp 299−335.

(60) Lagorce, D.; Pencheva, T.; Villoutreix, B. O.; Miteva, M. A. DG-AMMOS: A New Tool to Generate 3D Conformation of Small Molecules Using Distance Geometry and Automated Molecular Mechanics Optimization for in Silico Screening. *BMC Chem. Biol.* **2009**, *9*, 6.

(61) Pencheva, T.; Lagorce, D.; Pajeva, I.; Villoutreix, B. O.; Miteva, M. A. AMMOS: Automated Molecular Mechanics Optimization Tool for in Silico Screening. *BMC Bioinf.* **2008**, *9*, 438.

(62) O'Boyle, N. M.; Banck, M.; James, C. A.; Morley, C.; Vandermeersch, T.; Hutchison, G. R. Open Babel: An Open Chemical Toolbox. *J. Cheminf.* **2011**, *3*, 33.

(63) *The Open Babel Package*, version 2.3.0, 2015.

(64) Kuntz, I. D.; Blaney, J. M.; Oatley, S. J.; Langridge, R.; Ferrin, T. E. A Geometric Approach to Macromolecule-Ligand Interactions. *J. Mol. Biol.* **1982**, *161*, 269−288.

(65) Moustakas, D. T.; Lang, P. T.; Pegg, S.; Pettersen, E.; Kuntz, I. D.; Brooijmans, N.; Rizzo, R. C. Development and Validation of a Modular, Extensible Docking Program: DOCK 5. *J. Comput.-Aided Mol. Des.* **2006**, *20*, 601−619.

(66) Cornell, W. D.; Cieplak, P.; Bayly, C. I.; Gould, I. R.; Merz, K. M.; Ferguson, D. M.; Spellmeyer, D. C.; Fox, T.; Caldwell, J. W.; Kollman, P. A. A Second Generation Force Field for the Simulation of Proteins, Nucleic Acids, and Organic Molecules. *J. Am. Chem. Soc.* **1995**, *117*, 5179−5197.

Benchmarking Commercial Conformer Ensemble Generators

Reprinted with permission from
Friedrich, N.-O.; de Bruyn Kops, C.; Flachsenberg, F.; Sommer, K.; Rarey, M.; Kirchmair, J. Benchmarking Commercial Conformer Ensemble Generators. *J. Chem. Inf. Model.* **2017**, 57 (11), 2719–2728.

DOI: 10.1021/acs.jcim.7b00505

Journal of
CHEMICAL INFORMATION
AND MODELING

Article

Cite This: J. Chem. Inf. Model. 2017, 57, 2719-2728

pubs.acs.org/jcim

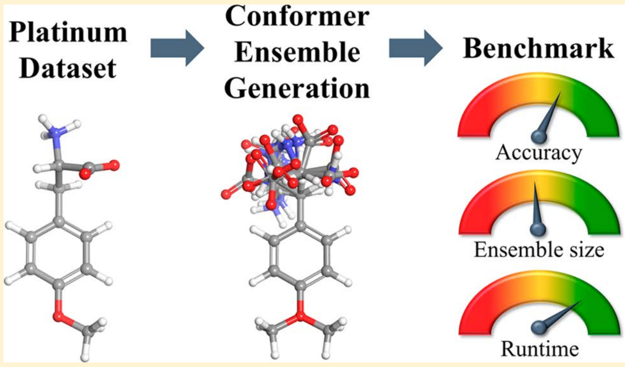
Benchmarking Commercial Conformer Ensemble Generators

Nils-Ole Friedrich, Christina de Bruyn Kops, Florian Flachsenberg, Kai Sommer, Matthias Rarey, and Johannes Kirchmair*

Center for Bioinformatics, Universität Hamburg, Bundesstr. 43, Hamburg 20146, Germany

Supporting Information

ABSTRACT: We assess and compare the performance of eight commercial conformer ensemble generators (ConfGen, ConfGenX, cxcalc, iCon, MOE LowModeMD, MOE Stochastic, MOE Conformation Import, and OMEGA) and one leading free algorithm, the distance geometry algorithm implemented in RDKit. The comparative study is based on a new version of the Platinum Diverse Dataset, a high-quality benchmarking dataset of 2859 protein-bound ligand conformations extracted from the PDB. Differences in the performance of commercial algorithms are much smaller than those observed for free algorithms in our previous study (*J. Chem. Inf. Model.* **2017**, *57*, 529−539). For commercial algorithms, the median minimum root-mean-square deviations measured between protein-bound ligand conformations and ensembles of a maximum of 250 conformers are between 0.46 and 0.61 Å. Commercial conformer ensemble generators are characterized by their high robustness, with at least 99% of all input molecules successfully processed and few or even no substantial geometrical errors detectable in their output conformations. The RDKit distance geometry algorithm (with minimization enabled) appears to be a good free alternative since its performance is comparable to that of the midranked commercial algorithms. Based on a statistical analysis, we elaborate on which algorithms to use and how to parametrize them for best performance in different application scenarios.



INTRODUCTION

Knowledge of protein-bound ligand conformations is an essential precondition for the application of 3D computational approaches such as docking and pharmacophore modeling. In most cases, experimental data on the bioactive conformations of small molecules bound to a biomacromolecule of interest are not available and must therefore be predicted. The most common approach to computing protein-bound ligand conformations is to sample their low-energy conformational space and generate representative conformer ensembles. Conformer ensemble generation is a well-studied problem, and modern algorithms are generally able to represent the bioactive conformations of drug-like molecules with adequate accuracy for virtual screening and further applications in cheminformatics.[1−7] One major challenge in the development of algorithms for conformer ensemble generation is the conflict of objectives between accuracy (generally measured as the minimum root-mean-square deviation (RMSD) in Å between the experimentally determined bioactive conformation and any computed conformers of an ensemble), ensemble size, and computing time. Depending on the specific application scenario, varying emphasis may be put on each of these parameters. For example, if accuracy is the primary objective, one would preferably choose a sophisticated and possibly computationally expensive algorithm that generates large ensembles of high-quality conformers. However, if large numbers of molecules are to be screened repeatedly, smaller ensembles may be preferred to reduce the number of molecular representations that need to be screened. If time is of essence, computationally efficient approaches would be preferred even though this choice usually coincides with cutbacks on quality. The ultimate goal is of course to have a computationally efficient algorithm accurately reproducing protein-bound ligand conformations with small ensembles.

During the last 20 years, not only the performance of the algorithms but also the quality of the benchmarking datasets and statistical analysis has improved substantially.[8,9] Benchmarking datasets, formerly compiled by taking into account the quality of the X-ray structures based almost exclusively on resolution, are now selected based on multiple criteria such as the real-space correlation coefficient (RSCC),[10] the real-space R-value (RSR),[11] and the diffraction-component precision index[12] (DPI).[9] Recently, we introduced the Platinum Dataset as the largest collection of high-quality protein-bound ligand conformations to date.[8] This dataset contains 4626 structures extracted from a total of over 347k structures of co-crystallized ligands stored in the PDB. The fully automated compilation pipeline included, apart from established filters and checks (e.g., removal of "unwanted ligands" such as crystallization agents and removal of structures of low resolution or high DPI), several new components, such as a method for the accurate

Received: August 22, 2017
Published: October 2, 2017

derivation and cross-checking of ligand topology and a novel approach for scoring the support of the atom positions in a molecule based on the electron density maps (EDIA).[13]

A representative subset of the Platinum Dataset (the Platinum Diverse Dataset),[8] which resulted from a clustering procedure to reduce any bias by the accumulation of certain molecular scaffolds in the dataset, was used in a previous study for benchmarking seven freely available conformer ensemble generators: Balloon (two different algorithms),[14] the standard distance geometry (DG)[15] and the experimental-torsion basic knowledge distance geometry[5] (ETKDG) algorithms implemented in RDKit, Confab,[16] Frog2,[17] and Multiconf-DOCK.[18] Significant differences in the performance of the individual algorithms became apparent, with the RDKit DG and ETKDG algorithms emerging as the best performing methods under most tested scenarios. For example, both algorithms were able to reproduce the bioactive conformations of more than 80% of all tested molecules with RMSDs below 1 Å. On the other hand, it also became apparent that some free algorithms did not produce ensembles for a large number of molecules and that some calculated conformers had geometric errors (e.g., wrong bond lengths, wrong bond angles, or planarity errors in rings).

In this follow-up work, we extend our benchmarking studies to eight commercial conformer ensemble generators and compare their performance with that of the standard DG algorithm implemented in RDKit. We also analyze the influence of force fields and conformer clustering procedures within some of the conformer ensemble generators on the quality of conformer ensembles.

■ RESULTS

Benchmarking Dataset. The Platinum benchmarking dataset presented in our previous study[8] was updated based on a more recent version of the PDB[19] (Table 1). Minor improvements to the data extraction pipeline were also made, such as the implementation of a revised version of EDIA[13] and additional quality checks for structural data (see Methods for details). The revised Platinum dataset consists of 4548 structures. Benchmarking was performed on a diversified subset of 2859 structures ("Platinum Diverse Dataset"). Accuracy tests with the RDKit DG algorithm yielded identical mean and median RMSD values with both versions of the Platinum Diverse Dataset, indicating that the results obtained with both datasets can be directly compared.

Table 1. Comparison of Platinum Dataset Versions

	Platinum Dataset 2016_01 (used in ref 8)	Platinum Dataset 2017_01 (this work)
Data extracted from the PDB on	February 12, 2016	February 16, 2017
Total no. of co-crystallized ligands in the PDB[a]	347671	350454
No. of compounds - Platinum Dataset	4626	4548
No. of compounds - Platinum Diverse Dataset	2912	2859
Compounds present in both versions of the Platinum Diverse Dataset	2763	
Compounds removed from the 2016_01 Platinum Dataset	170[b]	
Compounds added to the 2017_01 Platinum Dataset	92	

[a] With PDB advanced search query (i) "has external EDS link", (ii) "has free ligands", (iii) "experimental method is X-ray" and "has experimental data", (iv) has R_{work} below 0.4, and (v) R_{free} below 0.45. [b] Most of these structures were removed by the revised version of EDIA.

Commercial Conformer Ensemble Generators. The eight commercial conformer ensemble generators evaluated in this work are briefly introduced here. ConfGenX[20] (Schrödinger) is a further development of ConfGen,[21,22] a knowledge-based method that combines empirically derived heuristics and physics-based force field calculations. Cxcalc[23] (ChemAxon) is based on a fragment fusion method that uses the Dreiding force field[24] for calculation and optimization of conformers. The conformer ensemble generator iCon[25] is a systematic, knowledge-based search algorithm implemented in LigandScout (Inte:Ligand)[26] that uses a modified version of the MMFF94s[27] force field for refinement. The stochastic conformational search algorithm implemented in MOE (Molecular Operating Environment by Chemical Computing Group),[28] "MOE Stochastic", randomly rotates all bonds and performs an all-atom energy minimization to generate conformations. In contrast to the other methods included in this comparison, MOE Stochastic automatically generates enantiomers by also randomly inverting tetrahedral centers. The MOE LowModeMD method uses a fast, implicit vibrational analysis method in combination with a short molecular dynamics simulation to produce conformer ensembles.[29] A third algorithm implemented in MOE, the Conformation Import ("MOE Import"), breaks molecules into fragments and uses conformations from a standard library to assign torsional angles to common fragments. New fragments are generated using the MOE Stochastic approach described above. The program OMEGA (OpenEye)[2,30] also makes extensive use of fragment templates. The OMEGA algorithm attempts to generate energetically accessible combinations of the template conformations and uses a modified version of MMFF94s for scoring.

Accuracy with Default Parameter Sets Supplied by the Developers. The performance of conformer ensemble generators was measured for ensembles with a maximum of 50 and 250 conformers, and their accuracy was defined as the minimum RMSD in Å measured between the experimentally determined protein-bound conformation and any conformer of the computed ensemble. Unless stated otherwise, all RMSD values mentioned hereafter refer to the minimum RMSD calculated for a maximum ensemble size of 250. A complete overview of results (both median and mean RMSD values) is provided in Table 2. The results of the Mann−Whitney U test that was used to test for statistical significance are provided in the Supporting Information.

In a first experiment, we tested the performance of all algorithms with the default parameter sets supplied by the developers of the individual conformer ensemble generators (Figure 1). For ensembles with a maximum of 250 conformers, the eight commercial conformer ensemble generators were able to represent the bioactive conformations with median RMSDs between 0.46 and 0.61 Å over all ligands in the dataset.

OMEGA obtained the highest accuracy for ensembles with a maximum of 250 conformers (median RMSD 0.46 Å). However, at a maximum ensemble size of 250 OMEGA's ensembles were not significantly more accurate than those obtained with ConfGenX (median RMSD 0.49 Å), iCon (median RMSD 0.47 Å), and MOE LowModeMD (median RMSD 0.50 Å). The highest RMSDs were measured for cxcalc

Table 2. Arithmetic Mean and Median RMSD in Å Obtained for the Platinum Diverse Dataset[a]

Algorithm	Mode[b]	Clustering[c]	Force field	Maximum ensemble size 50		Maximum ensemble size 250	
				RMSD			
				mean	median	mean	median
ConfGen (default)	comprehensive	n/a	none	0.77	0.65	0.70	0.60
ConfGen	fast	n/a	none	0.90	0.83	0.90	0.83
ConfGenX (default)	n/a	n/a	none	0.69	0.59	0.58	0.49
ConfGenX	n/a	n/a	OPLS 2005	0.64	0.54	0.55	0.46
ConfGenX	n/a	n/a	OPLS3	**0.63**	0.52	**0.54**	0.44
cxcalc (default)	n/a	enabled	Dreiding	0.87	0.77	0.73	0.61
iCon (default)	fast	enabled	MMFF94s[d]	0.72	0.53	0.60	0.47
iCon	best	enabled	MMFF94s[d]	0.72	0.53	0.59	0.47
iCon	fast	disabled	MMFF94s[d]	0.80	0.59	0.64	0.46
MOE Stochastic (default)	n/a	enabled	MMFF94x[d]	0.75	0.55	0.64	0.52
MOE Stochastic	n/a	disabled	MMFF94x[d]	0.99	0.85	0.72	0.63
MOE LowModeMD (default)	n/a	enabled	MMFF94x[d]	0.75	0.54	0.62	0.50
MOE LowModeMD	n/a	disabled	MMFF94x[d]	1.10	0.93	0.75	0.66
MOE Import (default)	n/a	enabled	MMFF94x[d]	0.90	0.79	0.65	0.56
OMEGA (default)	n/a	enabled	mmff94s_NoEstat[e]	0.67	**0.51**	0.57	0.46
OMEGA	n/a	disabled	mmff94s_NoEstat[e]	0.74	0.56	0.59	**0.43**
RDKit DG (default)	n/a	disabled	UFF	0.77	0.64	0.63	0.52
RDKit DG	n/a	disabled	none	0.99	0.86	0.85	0.71
RDKit DG	n/a	disabled	MMFF94	0.79	0.69	0.66	0.56
RDKit DG	n/a	enabled	UFF	0.82	0.64	0.64	0.52
RDKit DG	n/a	enabled	none	1.01	0.85	0.88	0.72
RDKit DG	n/a	enabled	MMFF94	0.79	0.69	0.66	0.56

[a] The values of the best-performing algorithms per column are marked in bold. Algorithm performance is indicated by a color gradient, ranging from dark red (worst performance among all algorithms) via white to dark green (best performance among all algorithms). [b] Parameter sets and search modes offered by the various conformer ensemble generators. [c] Clustering of conformers by RMSD. [d] MMFF94 variants with altered out of plane bending parameters for conjugated nitrogens. [e] MMFF94 variant that includes all MMFF94s terms except Coulomb interactions.

(median RMSD 0.61 Å). There was also no significant difference observed between MOE LowModeMD (median RMSD 0.50 Å) and MOE Stochastic (median RMSD 0.52 Å). OMEGA additionally obtained the highest accuracy for small ensembles with a maximum of 50 conformers and was significantly more accurate at this ensemble size than all algorithms except for ConfGenX.

In comparison to the accuracy achieved with the RDKit DG algorithm (default configuration, i.e., with UFF enabled and clustering disabled), the commercial algorithms could be divided into three contrasting groups. OMEGA and ConfGenX obtained significantly higher accuracy than RDKit. MOE Stochastic, MOE LowModeMD, MOE Import, and iCon did not perform significantly better or worse than RDKit in terms of accuracy. ConfGen and cxcalc were significantly less accurate than RDKit.

For ConfGen and iCon we also tested alternative ensemble generation modes. ConfGen obtained a significantly lower accuracy in the "fast" mode (median RMSD 0.83 Å) as compared to the default (i.e., "comprehensive") mode (median RMSD 0.60 Å). Ensembles generated with the iCon "best" mode (median RMSD 0.47 Å) showed no significant difference in accuracy compared to the default (i.e., "fast") mode (median RMSD 0.47 Å).

Another way to measure the accuracy of conformer ensemble generators is their success rates in representing protein-bound ligand conformations below a certain RMSD threshold. Commonly used RMSD thresholds are 0.5, 1.0, 1.5, and 2.0 Å, and success rates for all tested algorithms and parameter sets

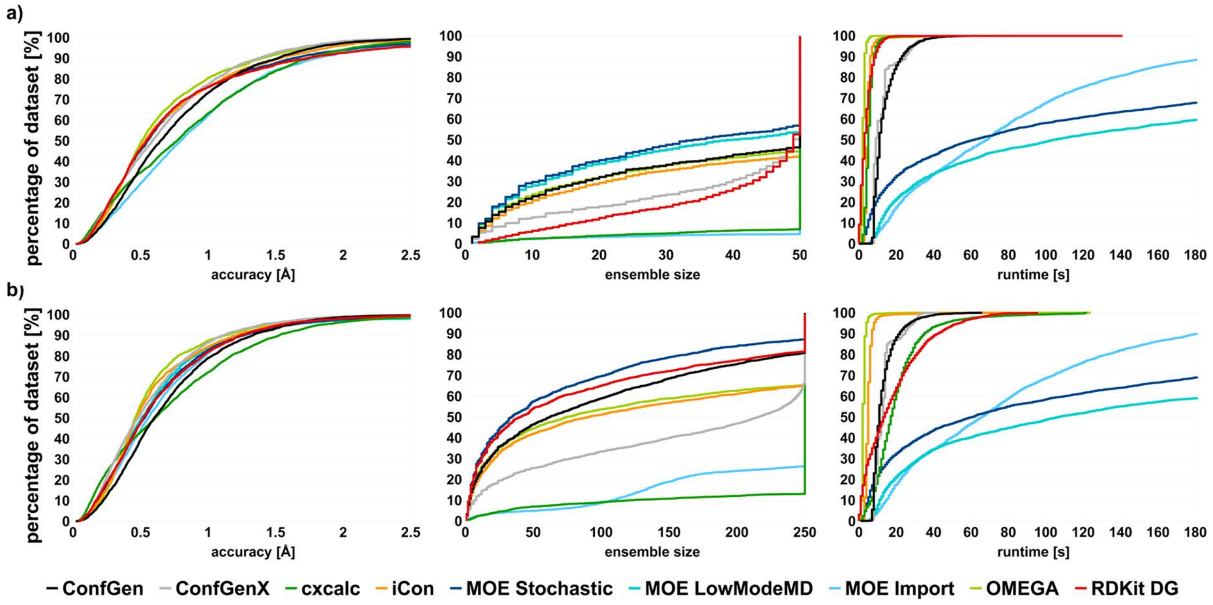


Figure 1. Percentage of protein-bound ligand conformations of the Platinum Diverse Dataset reproduced by the different algorithms within a certain accuracy (left), ensemble size (middle), and runtime per molecule (right) at maximum ensemble sizes (a) 50 and (b) 250 conformers. All commercial algorithms were run with default parameters.

for these thresholds are reported in Table 3. Our analysis focuses on two relevant RMSD thresholds: 0.6 Å, which is the maximum positional uncertainty for atoms in the Platinum Dataset, and 1.0 Å, below which docking poses are generally considered to adequately represent ligand binding modes.

At an RMSD threshold of 0.6 Å, for ligands with up to three rotatable bonds, OMEGA had the highest success rate of 96% (Figure 2a), closely followed by iCon (91%) and ConfGenX (90%). The lowest success rates for ligands with up to three rotatable bonds were measured for MOE Import (79%). The success rates sank drastically for ligands with up to 16 rotatable bonds. The smallest impact was found for iCon (66%) and OMEGA (64%), while cxcalc had the lowest success rate of 49%.

With an RMSD cutoff of 1.0 Å applied, all eight commercial ensemble generators had success rates above 70% for the tested molecules, even for those molecules with up to 16 rotatable bonds (Figure 2b; Table 3). For molecules with up to three rotatable bonds, the success rates of all conformer ensemble generators were >96%, with the exception of MOE LowModeMD (default; 86%) and Conformation Import (93%).

The smallest impact of the number of rotatable bonds on accuracy was observed for MOE LowModeMD (default), whose success rate dropped only to 84% at 16 rotatable bonds. The largest effect of the number of rotatable bonds was measured for cxcalc, whose success rate for molecules with three rotatable bond decreased from 95% to 72% at 16 rotatable bonds.

Accuracy with Conformer Clustering Procedures Disabled. Most conformer ensemble generators include a clustering procedure for the assembly of small, diverse ensembles. Although small ensembles are generally favorable in practice, when comparing their accuracy directly, these algorithms may be at a disadvantage over algorithms that fully exploit the maximum allowed ensemble size. For this reason, in the second experiment, and in analogy to our previous benchmarking study,[8] we disabled the clustering procedures in an attempt to maximize the accuracy of ensembles by forcing the algorithms to produce ensembles that exploit the allowed ensemble size. For iCon and OMEGA, deactivation of clustering procedures did not result in significant changes in performance, with median RMSD values of 0.46 and 0.43 Å (mean RMSDs 0.64 and 0.59 Å), respectively, as compared to 0.47 and 0.46 Å (mean RMSDs 0.60 and 0.57 Å), respectively, with clustering enabled (Table 2 and Figure S1). For the MOE Stochastic and LowModeMD algorithms, the median RMSDs increased significantly from 0.52 to 0.63 Å and from 0.50 to 0.66 Å, respectively, when clustering was disabled. These results also indicate that commercial generators are generally more accurate than free ones, which in our previous study obtained median RMSDs between 0.52 and 0.77 Å (mean RMSDs 0.63 to 0.92 Å) with clustering disabled.

Overall, the clustering procedures implemented in commercial conformer ensemble generators appear to be capable of producing small and, at the same time, accurate ensembles. Hence the recommendation based on these findings is to enable built-in clustering when using any of these conformer ensemble generators. The further discussion therefore focuses on results generated with conformer ensemble generators run with the default parameter sets supplied by the developers.

Accuracy of Conformer Ensemble Generators When Used in Combination with Different Force Fields. Force fields are used in conformer generation to improve initial coordinates, minimize generated conformers, or cluster them based on energy difference. Recently, the latest version of the OPLS force field, OPLS3, was released,[31] which includes nearly an order of magnitude more stretch, bend, and torsion parameters compared to MMFF[32] and OPLS_2005. OPLS3 was reported to improve protein−ligand binding predictions by 30% over earlier variants of the OPLS force field, without empirical fitting to protein−ligand binding data.[31] This prompted us to assess the influence of the optional geometry optimization with OPLS_2005 and OPLS3 on the performance of ConfGenX. However, the performance with OPLS3 (median RMSD 0.44 Å) was not significantly higher than with OPLS_2005 (median RMSD 0.46 Å). ConfGenX with the additional minimization by

DOI: 10.1021/acs.jcim.7b00505
J. Chem. Inf. Model. 2017, 57, 2719−2728

Table 3. Percentage of Structures of the Platinum Diverse Dataset Successfully Reproduced within a Specified RMSD Threshold[a]

Algorithm	Mode[b]	Clustering[c]	Force field	Maximum ensemble size 50				Maximum ensemble size 250			
				RMSD threshold							
				0.5	1.0	1.5	2.0	0.5	1.0	1.5	2.0
ConfGen (default)	comprehensive	n/a	none	37	73	90	97	40	79	93	**99**
ConfGen	fast	n/a	none	22	65	87	96	22	65	87	96
ConfGenX (default)	n/a	n/a	none	43	77	**93**	**98**	52	86	**96**	**99**
ConfGenX	n/a	n/a	OPLS 2005	46	80	**93**	**98**	54	**87**	**96**	**99**
ConfGenX	n/a	n/a	OPLS3	48	**81**	**93**	**98**	**56**	**87**	**96**	**99**
cxcalc (default)	n/a	enabled	Dreiding	34	63	84	94	43	72	89	96
iCon (default)	fast	enabled	MMFF94s[d]	47	76	89	96	54	84	95	**99**
iCon	best	enabled	MMFF94s[d]	47	76	90	96	55	85	95	**99**
iCon	fast	disabled	MMFF94s[d]	44	69	85	94	53	79	92	98
MOE Stochastic (default)	n/a	enabled	MMFF94x[d]	45	76	88	94	48	83	94	98
MOE Stochastic	n/a	disabled	MMFF94x[d]	27	58	80	91	38	75	93	97
MOE LowModeMD (default)	n/a	enabled	MMFF94x[d]	46	76	87	93	49	84	94	97
MOE LowModeMD	n/a	disabled	MMFF94x[d]	24	54	77	89	35	73	92	98
MOE Import (default)	n/a	enabled	MMFF94x[d]	29	61	85	94	44	81	95	98
OMEGA (default)	n/a	enabled	mmff94s_NoEstat[e]	**49**	80	92	97	**56**	**87**	**96**	**99**
OMEGA	n/a	disabled	mmff94s_NoEstat[e]	46	73	87	95	**56**	83	94	98
RDKit DG (default)	n/a	disabled	UFF	39	71	89	96	48	82	95	98
RDKit DG	n/a	disabled	none	22	58	80	91	32	69	87	95
RDKit DG	n/a	disabled	MMFF94	34	71	90	97	44	81	95	**99**
RDKit DG	n/a	enabled	UFF	38	71	89	96	47	82	95	98
RDKit DG	n/a	enabled	none	18	57	78	88	26	68	87	95
RDKit DG	n/a	enabled	MMFF94	34	71	89	96	43	80	94	97

[a] The values of the best-performing algorithms per column are marked in bold. Algorithm performance is indicated by a color gradient, ranging from dark red (worst performance among all algorithms) via white to dark green (best performance among all algorithms). [b] Parameter sets and search modes offered by the various conformer ensemble generators. [c] Clustering of conformers by RMSD. [d] MMFF94 variants with altered out of plane bending parameters for conjugated nitrogens. [e] MMFF94 variant that includes all MMFF94s terms except Coulomb interactions.

the OPLS3 force field reached the same conformation reproducibility as OMEGA with default parameters for the thresholds of 0.5 and 1.0 Å (56 and 87%, respectively; Table 3). Note that for individual test molecules, the minimum RMSDs for ensembles generated with ConfGenX (OPLS3 force field) and OMEGA differed by up to 2.5 Å (Figure S2).

Conformer Ensemble Size. The size of conformer ensembles is an important factor for accuracy, data size, and speed of downstream processes. When used with default settings, the size of the ensembles generated with the tested algorithms varies substantially (Table 4). In this section, all median values reported denote the median number of conformers per ensemble. MOE Stochastic produced the smallest ensembles (median 34), followed by MOE LowModeMD (median 39), ConfGen (median 63), OMEGA (median 74), ConfGenX with OPLS_2005 or OPLS3 (median 82 and 77, respectively; Figure 3), and iCon (median 90). MOE Import, cxcalc, and the RDKit DG algorithm produced the largest ensembles (median 250).

Runtime. In this section, all median values reported denote the median calculation time of conformer ensembles in seconds. The range of runtimes observed for both commercial and freely available conformer ensemble generators is between 1 and 110 s. OMEGA was the fastest of all tested algorithms, with a median runtime of 2 s per molecule (Table 5), and was also the only ensemble generator faster than the DG algorithm

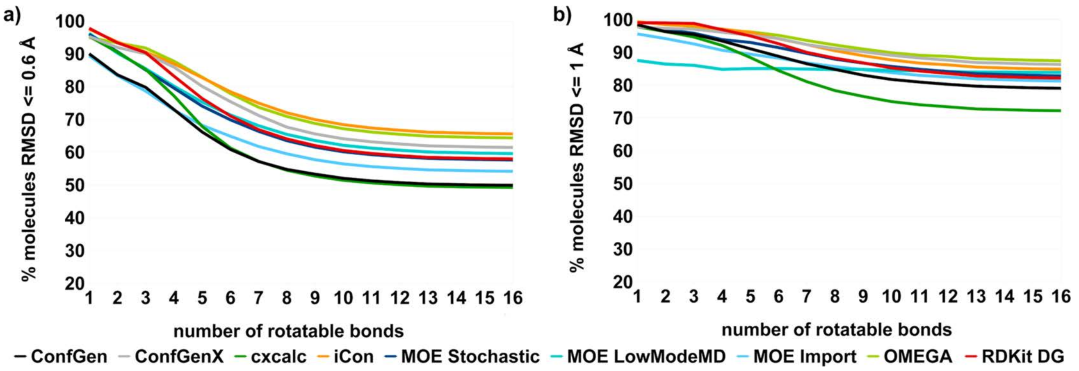


Figure 2. Percentage of molecules of the Platinum Diverse Dataset reproduced by the tested algorithms (default settings) with RMSD (a) ≤0.6 Å and (b) ≤1 Å as a function of the number of rotatable bonds. The maximum ensemble size was set to 250.

Table 4. Arithmetic Mean and Median Ensemble Sizes Obtained for the Platinum Diverse Dataset

				Maximum ensemble size 50 [conformers per ensemble]		Maximum ensemble size 250 [conformers per ensemble]	
Algorithm	Mode[a]	Clustering[b]	Force field	mean	median	mean	median
ConfGen (default)	comprehensive	n/a	none	34	50	100	63
ConfGen	fast	n/a	none	13	10	12	8
ConfGenX (default)	n/a	n/a	none	39	49	160	214
ConfGenX	n/a	n/a	OPLS 2005	30	34	102	82
ConfGenX	n/a	n/a	OPLS3	30	34	100	77
cxcalc (default)	n/a	enabled	Dreiding	48	50	227	250
iCon (default)	fast	enabled	MMFF94s[d]	35	50	123	90
iCon	best	enabled	MMFF94s[d]	36	50	131	117
iCon	fast	disabled	MMFF94s[d]	42	50	174	250
MOE Stochastic (default)	n/a	enabled	MMFF94x[d]	30	34	77	34
MOE Stochastic	n/a	disabled	MMFF94x[d]	50	50	237	250
MOE LowModeMD (default)	n/a	enabled	MMFF94x[d]	31	39	88	39
MOE LowModeMD	n/a	disabled	MMFF94x[d]	50	50	247	250
MOE Import (default)	n/a	enabled	MMFF94x[d]	48	50	215	250
OMEGA (default)	n/a	enabled	mmff94s_NoEstat[e]	34	50	118	74
OMEGA	n/a	disabled	mmff94s_NoEstat[e]	42	50	172	250
RDKit DG (default)	n/a	disabled	UFF	50	50	250	250
RDKit DG	n/a	disabled	none	50	50	250	250
RDKit DG	n/a	disabled	MMFF94	50	50	250	250
RDKit DG	n/a	enabled	UFF	42	49	180	229
RDKit DG	n/a	enabled	none	42	49	180	229
RDKit DG	n/a	enabled	MMFF94	42	49	180	229

[a] Parameter sets and search modes offered by the various conformer ensemble generators. [b] Clustering of conformers by RMSD. [d] MMFF94 variants with altered out of plane bending parameters for conjugated nitrogens. [e] MMFF94 variant that includes all MMFF94s terms except Coulomb interactions.

implemented in RDKit without force field minimization. OMEGA was followed by iCon with a median runtime of 5 s. ConfGen was slightly faster than ConfGenX, with a median runtime of 11 s compared to the median runtime of 13 s of ConfGenX. MOE Import (median 66 s) was of similar speed than MOE Stochastic (median 61 s). Both MOE Import and MOE Stochastic were much faster than MOE LowModeMD (median 106 s) but still substantially slower than all other tested algorithms. Note that MOE LowModeMD is not intended to be a high-throughput method.[29]

The ConfGen modes "fast" and "comprehensive" (default) had very similar runtimes, with median values of 10 and 11 s, respectively. The same was observed for the iCon modes "best" (median 5 s) and "fast" (default; median 5 s). ConfGenX was slower with force fields OPLS_2005 (median 35 s) and OPLS3 (median 21 s) than without them (default; median 13 s). All Schrödinger algorithms except ConfGen (default) show a characteristic staircase pattern in runtime plots. A reason for this behavior could not be determined.

For all of the algorithms with a clustering option, the effect of disabling clustering on the runtime was not substantial enough to justify the loss in accuracy or increase in ensemble size.

Processing Failures. All conformer ensemble generators were able to process 99−100% of the molecules of the Platinum Diverse Dataset. These rates are mostly higher than those of the previously tested free conformer ensemble generators, for which success rates between 75 and 100% for a maximum ensemble size of 250 were observed (RDKit DG algorithm with minimization enabled: 100%).[8]

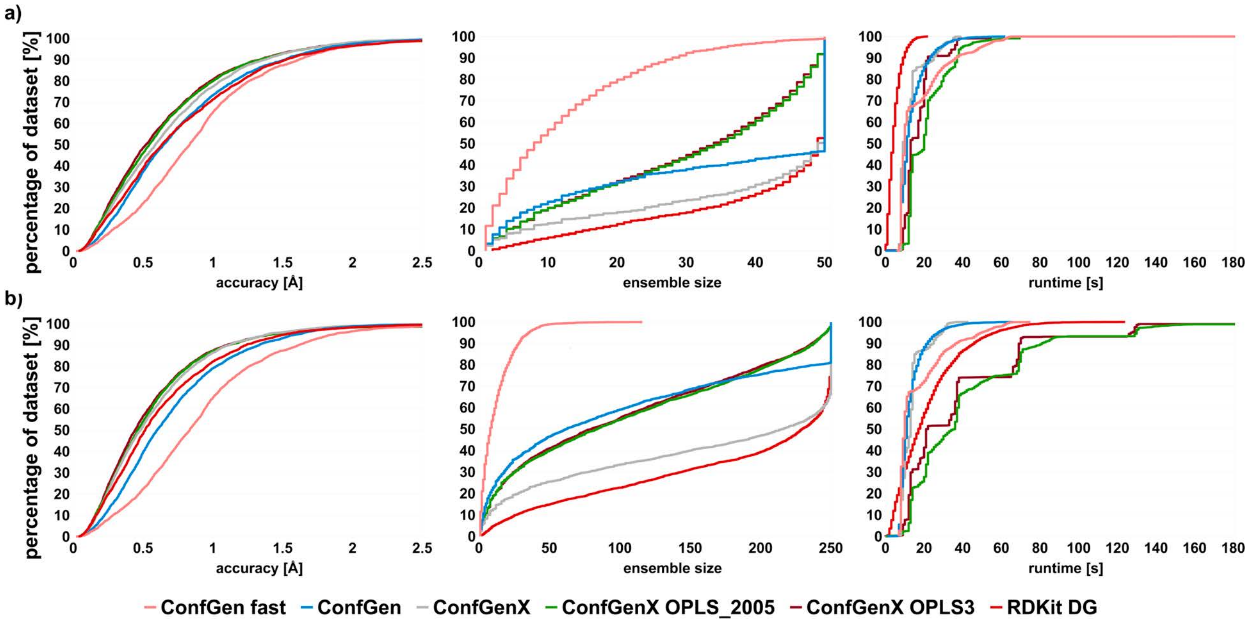


Figure 3. Percentage of protein-bound ligand conformations of the Platinum Diverse Dataset reproduced by ConfGen and ConfGenX within a certain accuracy (left), ensemble size (middle), and runtime (right) at maximum ensemble sizes (a) 50 and (b) 250. Ensembles generated with ConfGen in default (i.e., "comprehensive") and fast mode, ConfGenX in default mode, and with the force fields OPLS_2005 and OPLS3.

Anomalous Geometries Found in Output Conformers. In our previous study, we observed geometrical errors (such as wrong bond lengths, wrong bond angles, or out-of-plane errors in aromatic systems) in conformers generated with several of the free conformer ensemble generators (including the RDKit DG algorithm with minimization enabled).[8] In contrast, very few problems in the geometry of conformers generated with commercial algorithms were identified. Most errors observed for ConfGen, ConfGenX, cxcalc, and the three MOE algorithms were related to bond lengths (see Figure 4 for examples). No geometrical errors were identified for conformers generated with iCon and OMEGA.

■ CONCLUSIONS

The race for the best-performing conformer ensemble generator is much closer for commercial than for free algorithms. Commercial ensemble generators are characterized by their high robustness. All of the algorithms tested herein successfully processed at least 99% of all input molecules. No geometrical errors could be identified in conformations generated with iCon and OMEGA, and only a few anomalous geometries were identified for the other tested commercial algorithms. This robustness of the commercial ensemble generators is an important factor for their usability, in particular in industry.

OMEGA emerged as the algorithm with top accuracy and speed, while keeping ensemble sizes comparatively small. ConfGenX reached the same level of accuracy at the cost of larger ensembles and longer runtimes. These algorithms were closely followed by iCon, which presented itself as an accurate and fast alternative to the two algorithms. If ensemble size is of the essence, the MOE algorithms could be first choice. Nonetheless, the best free algorithm, the DG algorithm implemented in RDKit (with minimization enabled), was competitive with the commercial algorithms, landing in the middle ranks in terms of accuracy, ensemble size, and runtime. Note that most knowledge-based algorithms may have an inherent advantage in this kind of benchmark as they use PDB-derived libraries for biasing torsion angles.

The use of clustering procedures and force fields is recommended for the commercial conformer ensemble generators. Both generally lead to smaller ensembles and better accuracy, at low additional computational cost. Most conformer ensemble generators offer a wide range of customizable functions and parameters that can lead to a further improvement of their performance.

Overall, the benchmarking studies based on the Platinum Dataset showed that there are several free and commercial algorithms available today that allow the representation of protein-bound ligand conformations with adequate accuracy for most applications in computational drug discovery. In particular, the robustness of the commercial algorithms, and also of the RDKit DG algorithm, proved to be high.

■ METHODS

Data Set Compilation. The Platinum Dataset version 2017_01 was compiled according to the method described in ref 8, with the following improvements:

(1) The dataset was compiled more recently from the PDB, accessed February 16, 2017.

(2) A refined version of EDIA,[13] which is accessible online.[33]

(3) A method (based on NAOMI)[34] to filter ligands that are wrongly annotated as "free" ligands in the PDB while actually being covalently bound.

(4) Additional checks for out-of-plane errors of aromatic rings and ring systems with six or fewer atoms per relevant cycle were implemented in NAOMI: All triplets of atoms in aromatic ring systems connected by bonds were enumerated, and the out of plane angles for all adjacent bonds within the same aromatic ring system were subsequently calculated. Molecules with any out-of-plane angles >20° were removed from the dataset.

Table 5. Arithmetic Mean and Median Runtimes in Seconds Measured for the Platinum Diverse Dataset[a]

Algorithm	Mode[b]	Clustering[c]	Force field	Runtimes for maximum ensemble size 50 [s]		Runtimes for maximum ensemble size 250 [s]	
				mean	median	mean	median
ConfGen (default)	comprehensive	n/a	none	14	11	13	11
ConfGen	fast	n/a	none	17	10	17	10
ConfGenX (default)	n/a	n/a	none	13	9	14	13
ConfGenX	n/a	n/a	OPLS 2005	21	20	41	35
ConfGenX	n/a	n/a	OPLS3	16	13	37	21
cxcalc (default)	n/a	enabled	Dreiding	6	5	21	17
iCon (default)	fast	enabled	MMFF94s[d]	5	5	5	5
iCon	best	enabled	MMFF94s[d]	5	5	5	5
iCon	fast	disabled	MMFF94s[d]	8	8	8	8
MOE Stochastic (default)	n/a	enabled	MMFF94x[d]	158	62	153	61
MOE Stochastic	n/a	disabled	MMFF94x[d]	164	123	166	122
MOE LowModeMD (default)	n/a	enabled	MMFF94x[d]	232	103	475	106
MOE LowModeMD	n/a	disabled	MMFF94x[d]	282	237	278	226
MOE Import (default)	n/a	enabled	MMFF94x[d]	85	67	85	66
OMEGA (default)	n/a	enabled	mmff94s_NoEstat[e]	2	2	3	**2**
OMEGA	n/a	disabled	mmff94s_NoEstat[e]	2	2	**2**	**2**
RDKit DG (default)	n/a	disabled	UFF	5	4	22	17
RDKit DG	n/a	disabled	none	2	**1**	6	4
RDKit DG	n/a	disabled	MMFF94	6	5	34	28
RDKit DG	n/a	enabled	UFF	4	3	18	14
RDKit DG	n/a	enabled	none	**1**	**1**	5	4
RDKit DG	n/a	enabled	MMFF94	6	5	30	24

[a] The values of the fastest algorithms per column are marked in bold. Algorithm performance is indicated by a color gradient, ranging from dark red (longest runtime among all algorithms) via white to dark green (shortest runtime among all algorithms). [b] Parameter sets and search modes offered by the various conformer ensemble generators. [c] Clustering of conformers by RMSD. [d] MMFF94 variants with altered out of plane bending parameters for conjugated nitrogens. [e] MMFF94 variant that includes all MMFF94s terms except Coulomb interactions.

A representative subset, the Platinum Diverse Dataset version 2017_01, was derived by following the clustering protocol described in ref 8. Both datasets are available for download.[35]

Conformer Ensemble Generation. Standard 3D conformations generated from SMILES with NAOMI served as input for conformer ensemble generation. Conformer ensembles were calculated with the parameters described in the Results section. Clustering was disabled for individual experiments, by setting the RMSD cluster threshold for iCon, MOE Stochastic, MOE LowModeMD, and OMEGA to 0.0 Å, and the pruneRmsThresh for the RDKit DG algorithm to −1.0.

RMSD Calculations, Geometry Checks, and Runtime Measurements. All RMSD values were calculated with NAOMI, which selects the minimum heavy-atom RMSD for the best superposition of each pair of conformers, taking into account molecular symmetry via complete automorphism enumeration.

Deviation from known optimal values of atom angles and bond lengths as well as divergence from planarity of aromatic rings and ring systems (up to 6 bonds per relevant cycle) were measured with NAOMI. Runtimes were measured for SD files containing single molecules and rounded to full seconds. The deviations observed between repeated runtime experiments were <5%.

Statistical Analysis. The Mann−Whitney U test was used to test for statistical significance at $\alpha = 0.05$ and $\alpha = 0.01$, with the Holm−Bonferroni method[36] applied to control the family wise error rate. The raw p-values and the p-values adjusted with

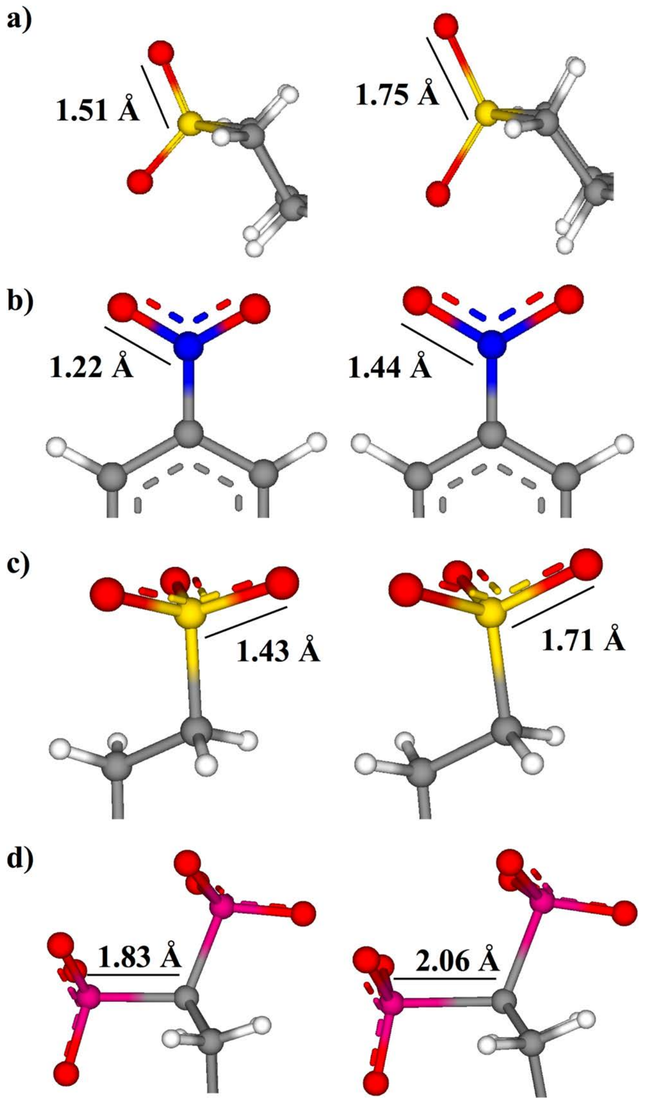


Figure 4. Examples of anomalous bond lengths found in conformers generated by commercial conformer ensemble generators. Input conformations are depicted on the left and computed conformers on the right. Excessively long bond lengths of (a) the double bond between sulfur and oxygen generated by ConfGen (AQ2 from 4AWI), (b) bonds in a nitro group (DCB from 4A3H) computed with ConfGenX, (c) the double bond between sulfur and oxygen in a sulfonate group (1PS from 1R4P) generated by cxcalc, and (d) the carbon phosphorus single bond (UNV from 3S4J) generated by MOE (all three modes).

the Holm−Bonferroni method are reported for all 253 pairwise comparisons of the conformer ensemble generators (and their different force fields and clustering algorithms) in the Supporting Information.

Hardware Setup. All calculations were performed on Linux workstations running openSUSE 13.1 and equipped with Intel Xeon processors (2.2−2.7 GHz) and 126 GB of main memory.

■ ASSOCIATED CONTENT

Ⓢ Supporting Information

The Supporting Information is available free of charge on the ACS Publications website at DOI: 10.1021/acs.jcim.7b00505.

Additional figures and tables: percentage of protein-bound ligand conformations reproduced by the different tools with different settings; comparison of the accuracy of ConfGenX and OMEGA for individual molecules (PDF)

Results of the Mann−Whitney U tests and *p*-values adjusted with the Holm−Bonferroni method for ensembles with a maximum of 50 conformers (TXT)

Results of the Mann−Whitney U tests and *p*-values adjusted with the Holm−Bonferroni method for ensembles with a maximum of 250 conformers (TXT)

■ AUTHOR INFORMATION

Corresponding Author

*E-mail: kirchmair@zbh.uni-hamburg.de. Tel.: +49 (0)40 42838 7303.

ORCID

Nils-Ole Friedrich: 0000-0002-8983-388X
Christina de Bruyn Kops: 0000-0001-8890-2137
Florian Flachsenberg: 0000-0001-7051-8719
Kai Sommer: 0000-0003-1866-8247
Matthias Rarey: 0000-0002-9553-6531
Johannes Kirchmair: 0000-0003-2667-5877

Notes

The authors declare no competing financial interest.

■ ACKNOWLEDGMENTS

The authors thank ChemAxon, Inte:Ligand, and OpenEye for providing free licences for their conformer ensemble generators, and the Chemical Computing Group and Schrödinger for providing academic licences for their conformer ensemble generators.

■ REFERENCES

(1) Chen, I.-J.; Foloppe, N. Conformational Sampling of Druglike Molecules with MOE and Catalyst: Implications for Pharmacophore Modeling and Virtual Screening. *J. Chem. Inf. Model.* **2008**, *48*, 1773−1791.

(2) Hawkins, P. C. D.; Skillman, A. G.; Warren, G. L.; Ellingson, B. A.; Stahl, M. T. Conformer Generation with OMEGA: Algorithm and Validation Using High Quality Structures from the Protein Databank and Cambridge Structural Database. *J. Chem. Inf. Model.* **2010**, *50*, 572−584.

(3) Ebejer, J.-P.; Morris, G. M.; Deane, C. M. Freely Available Conformer Generation Methods: How Good Are They? *J. Chem. Inf. Model.* **2012**, *52*, 1146−1158.

(4) Schärfer, C.; Schulz-Gasch, T.; Hert, J.; Heinzerling, L.; Schulz, B.; Inhester, T.; Stahl, M.; Rarey, M. CONFECT: Conformations from an Expert Collection of Torsion Patterns. *ChemMedChem* **2013**, *8*, 1690−1700.

(5) Riniker, S.; Landrum, G. A. Better Informed Distance Geometry: Using What We Know To Improve Conformation Generation. *J. Chem. Inf. Model.* **2015**, *55*, 2562−2574.

(6) Kothiwale, S.; Mendenhall, J. L.; Meiler, J. BCL::Conf: Small Molecule Conformational Sampling Using a Knowledge Based Rotamer Library. *J. Cheminf.* **2015**, *7*, 47.

(7) Hawkins, P. C. D. Conformation Generation: The State of the Art. *J. Chem. Inf. Model.* **2017**, *57*, 1747−1756.

(8) Friedrich, N.-O.; Meyder, A.; de Bruyn Kops, C.; Sommer, K.; Flachsenberg, F.; Rarey, M.; Kirchmair, J. High-Quality Dataset of Protein-Bound Ligand Conformations and Its Application to Benchmarking Conformer Ensemble Generators. *J. Chem. Inf. Model.* **2017**, *57*, 529−539.

(9) Warren, G. L.; Do, T. D.; Kelley, B. P.; Nicholls, A.; Warren, S. D. Essential Considerations for Using Protein-Ligand Structures in Drug Discovery. *Drug Discovery Today* **2012**, *17*, 1270−1281.

(10) Murshudov, G. N.; Vagin, A. A.; Dodson, E. J. Refinement of Macromolecular Structures by the Maximum-Likelihood Method. *Acta Crystallogr., Sect. D: Biol. Crystallogr.* **1997**, *53*, 240−255.

(11) Jones, T. A.; Zou, J. Y.; Cowan, S. W.; Kjeldgaard, M. Improved Methods for Building Protein Models in Electron Density Maps and the Location of Errors in These Models. *Acta Crystallogr., Sect. A: Found. Crystallogr.* **1991**, *47*, 110−119.

(12) Cruickshank, D. W. J. Remarks about Protein Structure Precision. *Acta Crystallogr., Sect. D: Biol. Crystallogr.* **1999**, *55*, 583−601.

(13) Meyder, A.; Nittinger, E.; Lange, G.; Klein, R.; Rarey, M. Estimating Electron Density Support for Individual Atoms and Molecular Fragments in X-Ray Structures. *J. Chem. Inf. Model.* **2017** 10.1021/acs.jcim.7b00391.

(14) Vainio, M. J.; Johnson, M. S. Generating Conformer Ensembles Using a Multiobjective Genetic Algorithm. *J. Chem. Inf. Model.* **2007**, *47*, 2462−2474.

(15) Tosco, P.; Stiefl, N.; Landrum, G. Bringing the MMFF Force Field to the RDKit: Implementation and Validation. *J. Cheminf.* **2014**, *6*, 37.

(16) O'Boyle, N. M.; Vandermeersch, T.; Flynn, C. J.; Maguire, A. R.; Hutchison, G. R. Confab - Systematic Generation of Diverse Low-Energy Conformers. *J. Cheminf.* **2011**, 3, 8.

(17) Miteva, M. A.; Guyon, F.; Tufféry, P. Frog2: Efficient 3D Conformation Ensemble Generator for Small Compounds. *Nucleic Acids Res.* **2010**, *38*, W622−W627.

(18) Sauton, N.; Lagorce, D.; Villoutreix, B. O.; Miteva, M. A. MS-DOCK: Accurate Multiple Conformation Generator and Rigid Docking Protocol for Multi-Step Virtual Ligand Screening. *BMC Bioinf.* **2008**, *9*, 184.

(19) Berman, H. M.; Westbrook, J.; Feng, Z.; Gilliland, G.; Bhat, T. N.; Weissig, H.; Shindyalov, I. N.; Bourne, P. E. The Protein Data Bank. *Nucleic Acids Res.* **2000**, *28*, 235−242.

(20) *ConfGenX*, Version 2016−2, Part of the Schrödinger Small-Molecule Drug Discovery Suite; Schrödinger: New York, NY, 2016.

(21) Watts, K. S.; Dalal, P.; Murphy, R. B.; Sherman, W.; Friesner, R. A.; Shelley, J. C. ConfGen: A Conformational Search Method for Efficient Generation of Bioactive Conformers. *J. Chem. Inf. Model.* **2010**, *50*, 534−546.

(22) *ConfGen*, Version 2016−2, Part of the Schrödinger Small-Molecule Drug Discovery Suite; Schrödinger: New York, NY, 2016.

(23) *cxcalc*, Version 15.8.31.0, Part of the Discovery Toolkit; ChemAxon: Budapest, Hungary, 2015.

(24) Mayo, S. L.; Olafson, B. D.; Goddard, W. A. DREIDING: A Generic Force Field for Molecular Simulations. *J. Phys. Chem.* **1990**, *94*, 8897−8909.

(25) *iCon*, Part of LigandScout Version 4.1; Inte:Ligand: Vienna, Austria, 2017.

(26) Wolber, G.; Langer, T. LigandScout: 3-D Pharmacophores Derived from Protein-Bound Ligands and Their Use as Virtual Screening Filters. *J. Chem. Inf. Model.* **2005**, *45*, 160−169.

(27) Halgren, T. A. MMFF VI. MMFF94s Option for Energy Minimization Studies. *J. Comput. Chem.* **1999**, *20*, 720−729.

(28) *Molecular Operating Environment (MOE)*, Version 2016.08; Chemical Computing Group: Montreal, QC, 2017.

(29) Labute, P. LowModeMD - Implicit Low-Mode Velocity Filtering Applied to Conformational Search of Macrocycles and Protein Loops. *J. Chem. Inf. Model.* **2010**, *50*, 792−800.

(30) *OMEGA*, Version 2.5.1.4; OpenEye Scientific Software: Santa Fe, NM, 2017.

(31) Harder, E.; Damm, W.; Maple, J.; Wu, C.; Reboul, M.; Xiang, J. Y.; Wang, L.; Lupyan, D.; Dahlgren, M. K.; Knight, J. L.; Kaus, J. W.; Cerutti, D. S.; Krilov, G.; Jorgensen, W. L.; Abel, R.; Friesner, R. A. OPLS3: A Force Field Providing Broad Coverage of Drug-like Small Molecules and Proteins. *J. Chem. Theory Comput.* **2016**, *12*, 281−296.

(32) Halgren, T. A. Merck Molecular Force Field. I. Basis, Form, Scope, Parameterization, and Performance of MMFF94. *J. Comput. Chem.* **1996**, *17*, 490−519.

(33) *The Proteins Plus Server*. http://proteinsplus.zbh.uni-hamburg.de (accessed June 1, 2017).

(34) Urbaczek, S.; Kolodzik, A.; Fischer, J. R.; Lippert, T.; Heuser, S.; Groth, I.; Schulz-Gasch, T.; Rarey, M. NAOMI: On the Almost Trivial Task of Reading Molecules from Different File Formats. *J. Chem. Inf. Model.* **2011**, *51*, 3199−3207.

(35) *The Platinum Datasets*. http://www.zbh.uni-hamburg.de/platinum_dataset (accessed June 27, 2017).

(36) Holm, S. A. Simple Sequentially Rejective Multiple Test Procedure. *Scand. J. Stat.* **1979**, *6*, 65−70.

How Diverse Are the Protein-Bound Conformations of Small-Molecule Drugs and Cofactors?

Reprinted with permission from
Friedrich, N.-O.; Simsir, M.; Kirchmair, J. How Diverse Are the Protein-Bound Conformations of Small-Molecule Drugs and Cofactors? *Front Chem* **2018**, 6, 68. Copyright 2018 Frontiers in Chemistry.

DOI: 10.3389/fchem.2018.00068

frontiers in Chemistry

ORIGINAL RESEARCH
published: 27 March 2018
doi: 10.3389/fchem.2018.00068

How Diverse Are the Protein-Bound Conformations of Small-Molecule Drugs and Cofactors?

Nils-Ole Friedrich[1], Méliné Simsir[1,2] and Johannes Kirchmair[1]*

[1] Department of Informatics, Center for Bioinformatics, Universität Hamburg, Hamburg, Germany, [2] Molécules Thérapeutiques In Silico, Université Paris Diderot, Sorbonne Paris Cité, Paris, France

Knowledge of the bioactive conformations of small molecules or the ability to predict them with theoretical methods is of key importance to the design of bioactive compounds such as drugs, agrochemicals, and cosmetics. Using an elaborate cheminformatics pipeline, which also evaluates the support of individual atom coordinates by the measured electron density, we compiled a complete set ("Sperrylite Dataset") of high-quality structures of protein-bound ligand conformations from the PDB. The Sperrylite Dataset consists of a total of 10,936 high-quality structures of 4,548 unique ligands. Based on this dataset, we assessed the variability of the bioactive conformations of 91 small molecules—each represented by a minimum of ten structures—and found it to be largely independent of the number of rotatable bonds. Sixty-nine molecules had at least two distinct conformations (defined by an RMSD greater than 1 Å). For a representative subset of 17 approved drugs and cofactors we observed a clear trend for the formation of few clusters of highly similar conformers. Even for proteins that share a very low sequence identity, ligands were regularly found to adopt similar conformations. For cofactors, a clear trend for extended conformations was measured, although in few cases also coiled conformers were observed. The Sperrylite Dataset is available for download from http://www.zbh.uni-hamburg.de/sperrylite_dataset.

Keywords: bioactive conformational space, protein-bound ligand conformation, conformational variability, PDB, protein-ligand interaction, binding site, small-molecule drug, cofactor

OPEN ACCESS

Edited by:
Daniela Schuster,
Paracelsus Private Medical University of Salzburg, Austria

Reviewed by:
Esther Kellenberger,
Université de Strasbourg, France
Francesco Ortuso,
Magna Græcia University, Italy
Sereina Riniker,
ETH Zürich, Switzerland

****Correspondence:***
Johannes Kirchmair
kirchmair@zbh.uni-hamburg.de

Specialty section:
This article was submitted to Medicinal and Pharmaceutical Chemistry, a section of the journal Frontiers in Chemistry

Received: *16 January 2018*
Accepted: *05 March 2018*
Published: *27 March 2018*

Citation:
Friedrich N-O, Simsir M and Kirchmair J (2018) How Diverse Are the Protein-Bound Conformations of Small-Molecule Drugs and Cofactors? Front. Chem. 6:68. doi: 10.3389/fchem.2018.00068

INTRODUCTION

The protein-bound ("bioactive") conformations of ligands can differ substantially from those observed in solution, the gas phase and small-molecule crystal structures (Boström, 2001; Perola and Charifson, 2004; Seeliger and de Groot, 2010). Bioactive conformations can be distributed over large regions of the ligand's conformational space and can have considerable strain energy (Nicklaus et al., 1995; Boström et al., 1998; Boström, 2001; Perola and Charifson, 2004; Günther et al., 2006). For the application of 3D computational approaches such as docking or *de novo* design methods in drug discovery, the protein-bound conformations of small molecules need to be known or at least determinable (Brameld et al., 2008).

The Protein Data Bank (PDB) is the most comprehensive resource of experimental structural data on biomacromolecules and their interaction with small molecules (Berman et al., 2000). Currently, the PDB contains more than 100k structures of biomacromolecules that include a bound ligand. While the structural data available from the PDB are extremely valuable for the research of biomacromolecules and their interactions with small molecules, these data represent only a very small fraction of (known) interactions.

Sturm et al. (2012) investigated the relationship between the promiscuity of drug-like molecules and the molecular properties of ligands and their binding sites. In order to do so, they compiled a dataset of more than 1,000 protein-ligand complexes in which drug-like molecules are bound to at least two distinct proteins. They identified two major drivers of ligand promiscuity: the structural similarities of ligand binding sites (largely independent of the similarities of the overall protein sequences or folds) and the ability of ligands to adopt distinct binding modes for different proteins. The latter is facilitated by the conformational flexibility of ligands and/or the specific characteristics of their pharmacophoric features. In related work, He et al. (2015), analyzed the structures of 100 pharmaceutically relevant ligands bound to at least two different proteins (to which they bind with comparable *in vitro* affinities). Contrary to the common belief that ligand flexibility and promiscuity are correlated, no evidence for a distinct correlation was found within their dataset. In fact, for 59 out of the 100 investigated ligands, no significant changes between the conformers of ligands bound to different proteins were observed.

The relative abundance of available structural data on the conformation of protein-bound cofactors, and nucleotide cofactors in particular, has made them a primary subject of investigation. For example, Moodie and Thornton (1993) analyzed 65 structures of nucleotides bound to proteins and found them to bind predominantly in an extended conformation. In more recent work, Stockwell and Thornton (2006) analyzed the conformational variability of adenosine triphosphate (ATP), nicotinamide adenine dinucleotide (NAD) and flavin adenine dinucleotide (FAD) in a preprocessed set of more than 2,000 structures extracted from the PDB. Dym and Eisenberg (2001) compiled a set of 150 structures of FAD bound to 32 non-redundant flavoproteins. They found a clear correlation between the FAD-family fold, the shape of the cofactor binding site and the conformation of FAD. Bojovschi et al. (2012) investigated the conformational diversity of ATP/Mg:ATP in motor proteins based on a set of 159 X-ray structures extracted from the PDB. They found that ATP adopts a wide range of different conformations, with a preference for extended conformations in tight binding pockets (e.g., F1-ATPase) and compact conformations in motor proteins such as RNA polymerase and DNA helicase. The incorporation of Mg2+ was found to increase the conformational flexibility of ATP. They clustered the conformations of the individual ligands based on the similarity of their binding pockets and, in the case of ATP for example, identified 27 clusters with a mean intercentroid RMSD of more than 2 Å. The authors concluded that, within the individual protein superfamilies, the investigated ligands generally bind in a fairly conserved manner, although several exceptions were identified. In the case of ATP, most structures were found to have the ligand bound in an extended conformation. In few cases however, a conformation bent such that the terminal phosphate atoms are almost in van der Waals contact with the adenine ring was observed. Stegemann and Klebe (2012) explored the structural properties of six cofactors including an adenosine diphosphate moiety bound to a variety of different proteins with low sequence identity. They found that common binding pocket patterns sometimes only recognize parts of the cofactor and thereby induce similar conformations.

These and further studies have contributed substantially to the understanding of protein-bound ligand conformations. However, a major bottleneck is the limited quality (Liebeschuetz et al., 2012; Reynolds, 2014), quantity and diversity of the structural data that these studies are based on, in particular with respect to the uncertainty of atom coordinates that is inherent to crystallographic structures. Only recently, a robust and fully automated method for the assessment of the support of individual atom coordinates (as well as molecules) by the measured electron density (EDIA) has become available (Meyder et al., 2017). This allowed, for the first time, extraction of a complete subset of high-quality structures of protein-bound ligands from the PDB (Friedrich et al., 2017b). Prior to the development of the EDIA method, time-consuming manual inspection by human experts was required to assure the high quality of structural data, which limited the size of available datasets (see e.g., Warren et al., 2012).

In this work we assess the conformational variability of small molecules based on a complete set of high-quality structures of protein-bound ligands extracted from in the PDB, each of which is represented by at least ten high-quality X-ray structures. In total the conformational variability of 91 approved drugs and cofactors represented by 4,574 protein-bound conformations was assessed. The bioactive conformational space of 17 representative molecules was studied in detail.

MATERIALS AND METHODS

Dataset Compilation

The Sperrylite Dataset was extracted from the PDB using a workflow described previously (Friedrich et al., 2017a). It consists of 10,936 conformers of 4,548 unique small molecules. Ninety-one ligands in this dataset are represented by at least 10 structures, and these served as the basis of this analysis.

To ensure that all ligands with the same PDB ligand ID have identical stereochemistry, their isomeric smiles (generated with UNICON, Sommer et al., 2016) were compared in order to keep only the isomer with the most occurrences. The Approved Drugs subset of DrugBank (Wishart et al., 2017) was used to identify the approved drugs present in the Sperrylite Dataset.

RMSD, Rotatable Bonds and Sequence Identity Calculations

All RMSD values were calculated with NAOMI (Urbaczek et al., 2011), which selects the minimum heavy-atom RMSD for the best superposition of each pair of conformers, taking molecular symmetry into account via complete automorphism enumeration.

The number of rotatable bonds was calculated with RDKit (RDKit: Open-Source Cheminformatics, version 2015.09.1, 2015). The default definition was used, meaning that amide and ester bonds were not counted as rotatable bonds.

All-against-all sequence identity was determined with NCBI BLAST (Altschul et al., 1990; BLAST, version 2.2.31. https://blast.ncbi.nlm.nih.gov (accessed Jan 14, 2018); Camacho et al., 2009) and the sequence identity of individual pairs of proteins was measured with the Molecular Operating Environment (Molecular Operating Environment (MOE), version 2016.08; Chemical Computing Group Inc.: Montreal, QC, 2016) based on sequence and structural alignments.

Principal component analysis (PCA)-derived score plots of the alignments with the minimum median RMSDs were generated with R for each ligand.

Visualization

Visualization of the (i) alignments of ligand conformers (ii) alignments of protein structures and (iii) interactions of proteins and ligands were generated with Maestro (Schrödinger Release 2016-2: Maestro, Schrödinger, LLC, New York, NY, 2016), MOE (Molecular Operating Environment (MOE), version 2016.08; Chemical Computing Group Inc.: Montreal, QC, 2016) and LigandScout (LigandScout, version 4.2; Inte:Ligand GmbH: Vienna, Austria, 2017; Wolber and Langer, 2005), respectively.

For the sake of clarity, all hydrogens, only polar hydrogens or no hydrogens were included in the depictions on a case-by-case basis to avoid overcrowded figures.

RESULTS

The Sperrylite Dataset is a collection of all high-quality X-ray structures of small molecules bound to biomacromolecules that are contained in the PDB. The dataset includes 10,936 structures of 4,548 unique protein-bound ligands and was compiled with a recently developed cheminformatics pipeline that automatically (i) prepares the chemical structures of small molecules by taking into account the protein environment (in order to determine, e.g., the most likely tautomeric and protonation states); (ii) removes undesirable molecules such as crystallization aids as well as structures with topological and/or geometrical errors; and (iii) rejects structures of low quality (Friedrich et al., 2017a,b). Importantly, the procedure not only includes checks for resolution and DPI (Cruickshank, 1999), but also employs the recently developed EDIA method (Meyder et al., 2017) to assess the support of individual atoms of a structure by the electron density.

In this study the diversity of the protein-bound conformations of all ligands represented by at least 10 high-quality structures was investigated. This dataset consists of a total of 4,574 conformations of 91 unique ligands (an overview of all structures is provided in Scheme S1), including more than 30 nucleotides and 20 approved drug molecules. In an all-against-all comparison of the differences in conformation of each ligand as measured by RMSD, 81 of the 91 ligands had at least one conformer with an RMSD above 0.6 Å (which corresponds to the maximum positional uncertainty for atoms in the Sperrylite Dataset), and 69 had at least one conformer above 1 Å, meaning that they are clearly distinct. The correlation observed between the minimum median RMSD measured for all pairs of conformations and the number of rotatable bonds was (very) weak ($R^2 = 0.126$; Figure S1).

This work focuses on the analysis of the bioactive conformational space of a representative set of 17 approved drugs and cofactors (**Tables 1**, **2**; note that there is an overlap between cofactors and approved drugs). This set was compiled with the objective to include the most relevant and best-represented small molecules in a detailed analysis of individual ligands.

Definitions

In the following sections, “high-quality structures” refers to any structures matching the quality criteria defined in previous work (Friedrich et al., 2017b). Importantly, this term only refers to the quality of the protein-bound ligand, not the overall structure of the protein-ligand complex. Four-letter codes refer to PDB entries and three-letter codes in italics refer to PDB ligand identifiers.

Small-Molecule Drugs

Imatinib

Imatinib (*STI*) is an approved anti-cancer drug targeting Bcr-Abl and several other tyrosine kinases. The drug binds to the ATP-binding site, spanning almost the entire width of the protein (Reddy and Aggarwal, 2012). Imatinib locks the protein in a closed conformation, thus arresting the enzyme’s functionality. The PDB lists 11 high-quality structures with imatinib, 10 thereof with the drug bound to one of three different tyrosine kinases (ABL1: 1IEP, 1OPJ, 3K5V, 3MS9, 3MSS, 3PYY; ABL2: 3GVU; c-Src: 2OIQ, 3OEZ) or a synthetic construct of tyrosine kinase AS (4CSV), a common ancestor of Src and Abl.

The accessible conformational space of imatinib, which has seven rotatable bonds, is large. However, the conformations observed for imatinib bound to any of these tyrosine kinases are similar (**Figures 1A,B**), which is reflected by the low maximum pairwise RMSD of just 0.3 Å and is in agreement with the findings of He et al. (2015). This conformational similarity can be explained by the highly conserved nature of the residues that form the ligand binding sites of these tyrosine kinases (the minimum pairwise sequence identity between these proteins is 45%; **Figure 1D**).

One high-quality structure of imatinib is a complex with human quinone reductase 2 (3FW1). This enzyme exists as a dimer with two active sites, each located in a deep pocket at the interface between the monomers (Foster et al., 1999; Winger et al., 2009). Quinone reductase 2 is structurally dissimilar to protein kinases. Imatinib binds to the enzyme active site in proximity to the isoalloxazine ring of the FAD cofactor (**Figure 1C**), thereby adopting a distinct, “horseshoe-like” conformation (Winger et al., 2009) that differs by at least 2.4 Å from any of the conformations observed with tyrosine kinases (**Figure 1A**).

TABLE 1 | Summary of approved drugs investigated in this work.

Name	No. of PDB entries	Protein names: No. of high-quality conformers	No. of Confs.[a]	Major observations
Imatinib (STI)	18	Tyrosine kinases: 10 Quinone reductase 2: 1	2	Conformers for different tyrosine kinases are similar, even for pairs of proteins with low sequence identity. A distinct conformation is observed in a complex with quinone reductase 2
Darunavir (017)	54	HIV-1 protease: 14	1	Conformers are highly similar, also those in complex with various different mutants of this protein
Acetazolamide (AZM)	29	Carboanhydrases: 9 Endochitinase: 1	n.d.[b]	It is likely that the ligand binds in a similar conformation to all proteins covered by the dataset (the experimental data do not allow a definitive conclusion)
Triclosan (TCL)	31	Enoyl-acyl carrier protein reductases: 11	1	All conformers are highly similar. The median RMSD is 0.1 Å and the maximum pairwise RMSD is below 0.6 Å
Ubenimex/bestatin (BES)	28	Aminopeptidases: 9 Leukotriene A-4 hydrolase: 2	3	Conformations observed for most (even distantly) related aminopeptidases and human leukotriene A-4 hydrolases are similar, with the exception of one conformation observed in complex with human aminopeptidase N
Biotin (BTN)	99	Streptavidin: 24 Avidin: 7 Biotin-protein ligase: 6 Others: 6	3	Conformations observed among the different complexes with core streptavidin are very similar. Two distinct conformers are observed in complex with biotin-protein ligase and biotin carboxylase
Sapropterin (H4B)	472	Total: 188	2	All but three conformers are extremely similar to each other (median RMSD smaller than 0.1 Å), even for distantly related proteins
Cholic acid (CHD)	74	Cytochrome c oxidase: 2 Ferrochelatase: 2 Alcohol dehydrogenase: 1 Others: 8	4	Due to the rigid steroid scaffold, the conformations observed for both ligands are all highly similar
Deoxycholic acid (DXC)	29	Cathepsin A: 11 Bet v1: 2 Others: 3	2	

[a] *No. of distinct bioactive conformations.*

[b] *The experimental data are insufficient to allow a definitive conclusion on the number of distinct bioactive conformations.*

Note that imatinib is known to bind to spleen tyrosine kinase (SYK) in an orientation that is different from that observed for Bcr-Abl and other tyrosine kinases (Alton and Lunney, 2008). A crystal structure of the imatinib-SYK complex exists (1XBB; Atwell et al., 2004) but is not part of the Sperrylite Dataset because of a poor electron density support of parts of the ligand facing the bulk water phase (Figure S2). The conformer of imatinib in complex with SYK has an RMSD of 2.5 Å to any of the other kinase-bound conformers but is similar to the imatinib conformation observed in the complex with quinone reductase 2 (RMSD = 1.3 Å).

Darunavir

Darunavir (*017*) is an antiretroviral drug approved for the treatment and prevention of human immunodeficiency virus (HIV) infections. The compound inhibits HIV-1 protease at picomolar concentrations by forming strong polar interactions with the target enzyme (King et al., 2004). Fourteen out of the 54 available structures with darunavir are of high quality, all of them being structures with darunavir bound to wild type or mutant HIV-1 protease. The mutations observed in the 14 high-quality structures introduce only subtle changes to the shape and chemical properties of the ligand binding environment. This is reflected in the high similarity of the protein-bound conformations of darunavir, where, among the high-quality structures, a maximum pairwise RMSD of just 0.2 Å was measured (Figure S3).

Acetazolamide

Acetazolamide (*AZM*) is an inhibitor of carbonic anhydrase and approved for the treatment of glaucoma, cardiac edema, idiopathic intracranial hypertension, epilepsy, and altitude sickness (Chakravarty and Kannan, 1994; Kaur et al., 2002). Ten out of the 29 structures of acetazolamide listed in the PDB are of high quality. Nine of these structures are with acetazolamide bound to one of six different human carbonic anhydrases (isoforms II, VII, IX, XII, XIII, and XIV, represented by PDB entries 3V2J, 3ML5, 3IAI, 1JD0, 3CZV, and 4LU3, respectively) or three different extremophilic bacteria carbonic anhydrases (*Sulfurihydrogenibium* sp., *Thermovibrio ammonificans*, and *Sulfurihydrogenibium azorense*, represented by PDB entries 4G7A, 4UOV, and 4X5S, respectively). The ligand binding pockets of all these carbonic anhydrase isozymes are highly similar (**Figure 2G**) and so are the conformations of acetazolamide observed for these complexes (**Figure 2A**). The protein-ligand complexes are stabilized by hydrogen bonds formed between the acetyl group of acetazolamide and the binding pocket (**Figure 2B**), with one exception, which is a complex with human carbonic anhydrase XII (1JD0). In that structure, the acetyl group of the ligand is rotated by about

TABLE 2 | Summary of cofactors and cofactor analogs investigated in this work.

Name	No. of PDB entries	Protein names: No. of high-quality conformers	No. of Confs.^{a}	Major observations
Sinefungin (SFG)	70	Methyltransferase: 23 Others: 7	6	Three clusters of conformers are observed. The largest group includes 23 highly similar conformers and includes structures bound to proteins that share low sequence identity. The maximum RMSD measured for any of the sinefungin conformers is 3.6 Å
S-adenosylmethionine (SAM)	410	Methyltransferase: 92 Others: 31	24	A wide variety of conformations are observed, with a clear clustering into three distinct groups of conformers. Within these groups, a large number of similar conformers are observed, even when bound to proteins sharing low sequence identity. The maximum pairwise RMSD is 3.3 Å
S-Adenosyl-L-Homocysteine (SAH)	784	Methyltransferase: 284 RNA polymerase: 8 Others: 19	23	
Glutathione (GSH)	360	Glutathione transferase: 46 Others: 28	16	Most of the conformers have a pairwise RMSD between 0.6 and 1.6 Å, but the maximum pairwise RMSD is 3.6 Å
Adenosine monophosphate (AMP)	575	Total: 171	36	A wide variety of different conformers is observed. One distinct, extremely coiled conformer was observed in complex with an adenylate kinase-related protein. The maximum pairwise RMSD is 2.5 Å
Adenosine diphosphate (ADP)	1,810	Total: 462	81	The conformers are similar to those observed for AMP, despite the presence of an additional phosphate group. The median RMSD is 0.9 Å
Adenosine triphosphate (ATP)	1,079	Total: 218	76	ATP is observed in an extended conformation in most structures, but some conformers are extremely bent. The median and the maximum pairwise RMSDs are 1.6 and 3.9 Å, respectively
Flavin mononucleotide (FMN)	919	Total: 367	21	The overall median RMSD is 0.9. The all-against-all comparison revealed four groups of conformers, with peaks in the RMSD distribution at around 0.3, 1.2, 1.7, and 2.4 Å

[a] *No. of distinct bioactive conformations.*

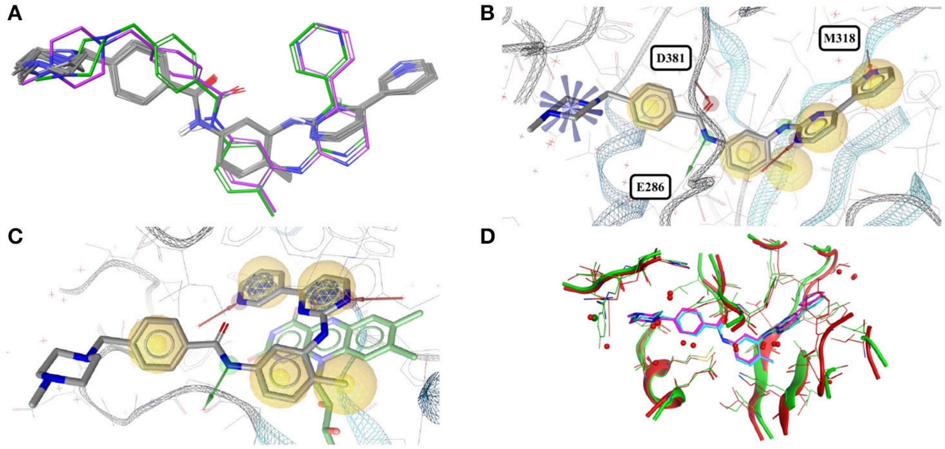


FIGURE 1 | (A) Ligand-based alignment of imatinib conformers observed in complex with three different tyrosine kinases (gray carbon atoms), human quinone reductase 2 (3FW1; violet carbon atoms) and human spleen tyrosine kinase (1XBB; green carbon atoms). **(B)** Imatinib bound to ABL1 (3MS9) in an extended conformation that is characteristic for the drug bound to tyrosine kinases. Red and green vectors indicate hydrogen bond donors and acceptors, respectively. Yellow spheres mark hydrophobic moieties involved in interactions with the protein, and blue astral centers indicate charge interactions involving a positively charged group on the ligand side. **(C)** Imatinib bound to human quinone reductase 2 in a conformation that is different from those characteristic of tyrosine kinases (3FW1; FAD with green carbon atoms). **(D)** Alignment of the binding sites of human ABL1 (3K5V; red) and c-Src (2OIQ; green). Despite a sequence identity of only 45%, the ligand binding sites of both proteins are almost identical.

140° as compared to any of the other structures (RMSD 0.9 Å; **Figure 2C**). A second, distinct conformation of acetazolamide is found in a complex with a different enzyme, endochitinase from *Saccharomyces cerevisiae* (2UY4) with a fundamentally different binding pocket. In that structure, the carbon-sulfur bond of the ligand is rotated by 120° (**Figure 2D**). The moieties in question are oriented toward the bulk water phase, freely rotatable, and not engaged in directed interactions with the protein. Also, the electron density maps do not allow a definitive conclusion on the orientation of these moieties (**Figures 2E,F**). It is therefore entirely possible that in reality all conformers of acetazolamide in the Sperrylite Dataset are nearly identical.

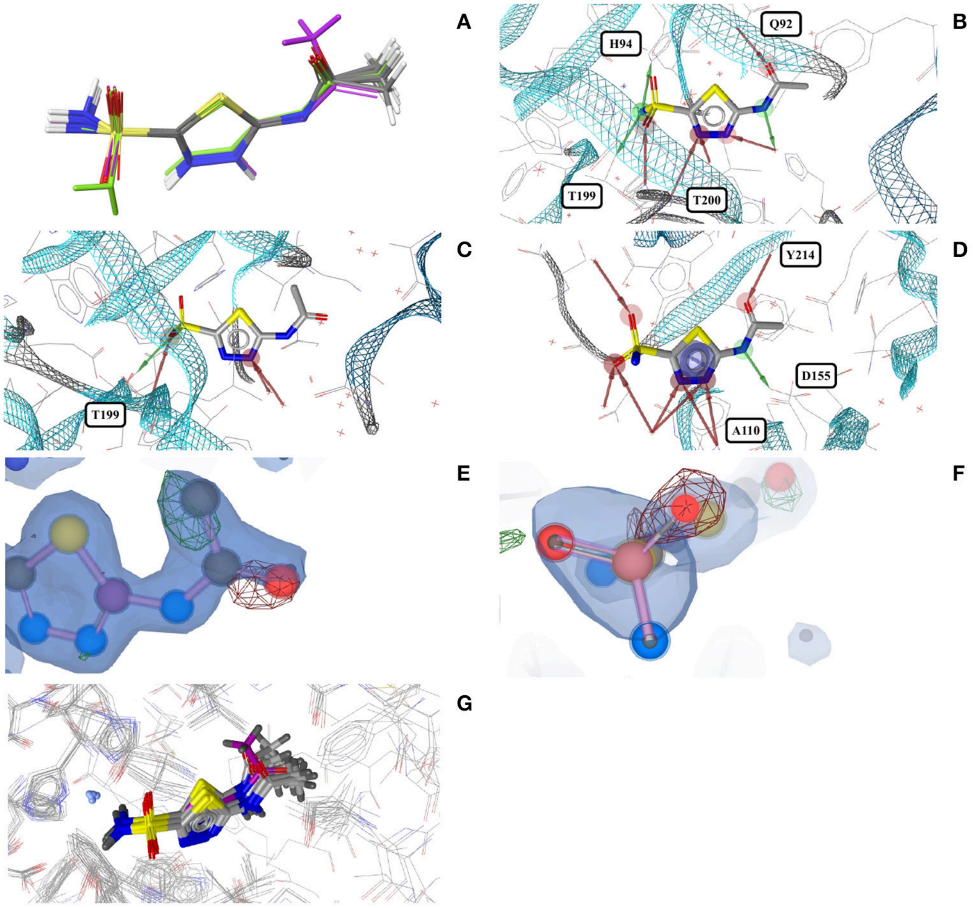


FIGURE 2 | (A) Ligand-based alignment of acetazolamide bound to different carbonic anhydrases (gray carbon atoms, except those of 1JD0, which are violet) and endochitinase (2UY4; green). **(B)** The acetyl group of acetazolamide forms hydrogen bond interactions with some carbonic anhydrases such as isozyme VII (3ML5) depicted here. **(C)** In a complex with human carbonic anhydrase XII (1JD0) the acetyl group of acetazolamide is rotated by about 140°. **(D)** In a complex with endochitinase (2UY4), the sulfonamide moiety of acetazolamide is rotated by about 120°. The support of atom positions by the measured electron density can be quantified by the EDIA score. For some of the atoms of the acetyl **(E)** and sulfonamide groups **(F)** of these structures the EDIA scores are below 0.8, meaning that their exact position is uncertain. The 2Fo-Fc, Fo-Fc(–ve) and Fo-Fc(+ve) sigma maps are shown in blue, red and green, respectively. It can therefore not be excluded that the acetyl group in **(C)** and the sulfonamide moiety in **(D)** are present in the same orientation that is observed in any of the other crystal structures. **(G)** Superposed binding pockets of the nine human and three extremophilic bacterial carbonic anhydrases.

Triclosan

Triclosan (*TCL*) is an antibacterial and antifungal agent inhibiting enoyl-acyl carrier protein reductases (ENR), which are key enzymes in the fatty acid elongation cycle. Its wide use as a disinfectant in cremes and consumer products (e.g., soaps, toothpaste, detergents) is a controversial topic nowadays (Buth et al., 2010; Carey and McNamara, 2014).

In all 31 structures of triclosan contained in the PDB, the ligand is bound to an ENR. The conformers of triclosan observed among the 11 high-quality structures with ENR I and ENR III are very similar (median RMSD 0.1 Å; maximum pairwise RMSD < 0.6 Å; **Figure 3A**). These include the structures of *Plasmodium falciparum* ENR I (2O2Y) and *Bacillus subtilis* ENR III (3OID) which, despite a sequence identity of just 14% and a highly flexible binding site region (when in the unbound state), show almost identical structural features in the presence of triclosan (Kim et al., 2011).

In an X-ray structure of triclosan bound to *Staphylococcus aureus* ENR I (3GR6; not included in the Sperrylite Dataset because of low EDIA scores), the hydroxyl group of all four instances of triclosan is modeled in a different orientation (RMSD 1.4 Å measured to any of the other conformations present in the dataset). The EDIA score for the oxygen atom of the hydroxyl group of the four instances of this conformer is just 0.11–0.27, and visual inspection of the electron density map confirms a lack of support of this conformation (**Figure 3B**). The characteristic hydrogen bonds formed between the phenolic hydroxyl group of triclosan and Y156 as well as NAD(P) (Heath et al., 1999; Levy et al., 1999; **Figure 3C**) are also missing in this model (**Figure 3D**). All of these observations taken together indicate a likely error in this structural model.

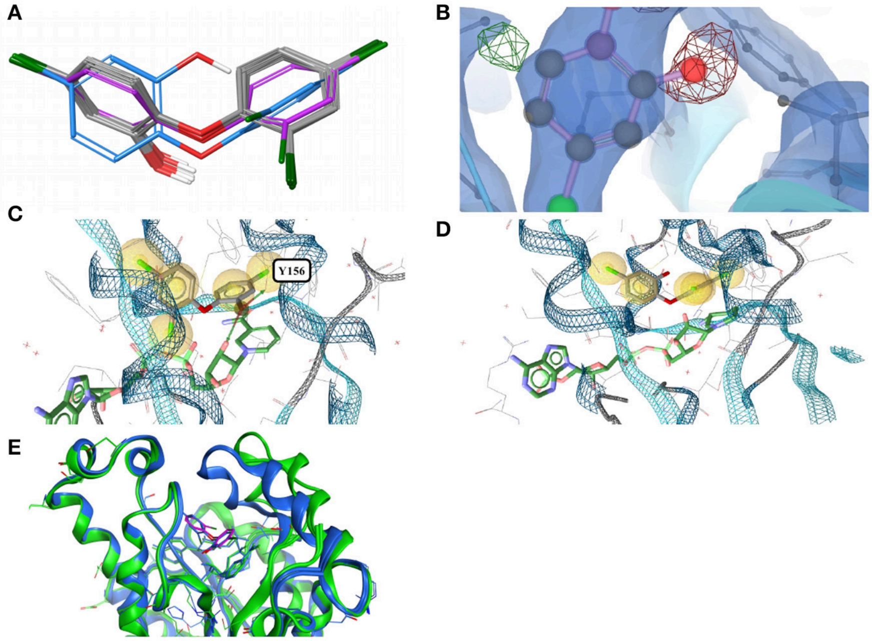


FIGURE 3 | (A) Ligand-based alignment of eleven conformers of triclosan present in the Sperrylite Dataset bound to ENRs, including the drug-resistant G93V mutant of ENR I (3PJF; violet carbon atoms) and an uncommon conformation observed in *Staphylococcus aureus* ENR I (3GR6; blue carbon atoms). The latter is not part of the Sperrylite Dataset because of a lack of support of the structural model by the electron density, as shown in **(B)**, with the 2Fo-Fc, Fo-Fc(–ve) and Fo-Fc(+ve) sigma maps in blue, red and green, respectively. **(C)** Interaction of triclosan and NAD (green carbon atoms), including the characteristic hydrogen bond between both molecules in the binding pocket of *E. coli* ENR I (1QG6). **(D)** Triclosan and NADP (green) bound to *Staphylococcus aureus* ENR I (3GR6). In this structural model, the characteristic hydrogen bond is missing because of the unusual position of the hydroxyl group. However, this conformation of triclosan lacks support by the measured electron density. **(E)** A G93V mutation in ENR I (green protein backbone; ligand with violet carbon atoms) induces a conformational shift of the flexible α-helical turn located in the proximity of triclosan. The complex of the WT protein and triclosan (4M89) is shown with the protein backbone and ligand in blue.

The largest deviations between conformers of triclosan within the Sperrylite Dataset were observed for the complex with a triclosan-resistant G93V mutant (3PJF) of ENR I from *Escherichia coli*. These deviations are related to small conformational changes of a flexible α-helical turn in close proximity to the ligand (**Figure 3E**), resulting in the weakening of some edge-to-face aromatic interactions near the ligand (Singh et al., 2011). The high-level resistance of this mutant is not caused by a substantial loss in binding affinity of the drug but is a consequence of the inability of the G93V mutant to form the high affinity ENR-NAD+-triclosan ternary complex that inhibits the wild type (Heath et al., 1999).

Ubenimex, Bestatin

Ubenimex, also known as bestatin (*BES*), is a competitive protease inhibitor under investigation for the treatment of acute myelocytic leukemia and lymphedema (Tian et al., 2017). The molecule inhibits aminopeptidases and has shown immunomodulatory and host-mediated antitumor activities (Urabe et al., 1993; Inoi et al., 1995; Sakuraya et al., 2000). It has been approved in Japan as an adjunct to chemotherapy agents against acute non-lymphocytic leukemia for decades and has been reported to inhibit the growth of malaria parasites (*Plasmodium falciparum*) *in vitro* (Nankya-Kitaka et al., 1998).

Twenty-eight structures of bestatin are listed in the PDB. All of the 11 high-quality structures are with bestatin bound to aminopeptidases. The ligand conformations observed in eight of these high-quality structures are very similar to each other (maximum pairwise RMSD = 0.8 Å), even though the proteins originate from three different bacteria (*E. coli*, *Pseudomonas putida* and *Vibrio proteolyticus*), the unicellular protozoan parasite *Plasmodium falciparum* and mouse, and their minimum pairwise sequence identity is only 3.3%.

In contrast, the structure of bestatin bound to human aminopeptidase N (4FYR) shows an extended ligand conformation that has an RMSD of 2.0 Å to any of the ligand conformers observed for the bacterial proteins (**Figure 4A**). The conformations of the drug bound to human leukotriene A-4 hydrolase differ only slightly from and have similar binding modes to the characteristic conformation observed for aminopeptidases mentioned above (RMSD = 1.0 Å for both 3FUH and 3FTX; **Figures 4B–D**).

Biotin

Biotin (*BTN*, vitamin B_7) is a water-soluble coenzyme for carboxylase enzymes and an approved drug for the treatment of dietary shortage or imbalance. There are 99 crystal structures including biotin listed in the PDB. The biotin conformers observed for the 43 high-quality structures can be assigned

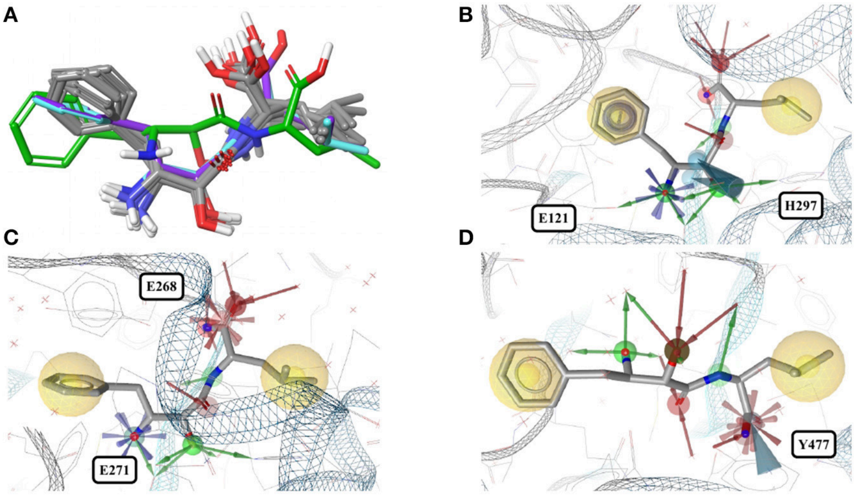


FIGURE 4 | (A) Superposition of all eleven conformers of bestatin in the Sperrylite Dataset. The carbon atoms of the conformers in complex with human aminopeptidase N (4FYR) and human leukotriene A-4 hydrolase (3FUH and 3FTX) are indicated in green, violet and cyan, respectively. The carbon atoms of all other structures are shown in gray. **(B)** Typical conformer of bestatin bound to aminopeptidases N from *E. coli* (2HPT). **(C)** A conformation that differs slightly from the characteristic conformation, observed in complex with human leukotriene A-4 hydrolase (3FUH shown here). **(D)** Uncommon, extended conformation of bestatin observed in complex with the human aminopeptidase N (4FYR).

to three distinct groups, indicated by gray, green and violet carbon atoms in **Figure 5A**. Twenty-four of the 43 structures are complexes with core streptavidin from different bacteria (both wild type and mutants). Streptavidin homotetramers have a very high affinity for biotin, one of the strongest non-covalent interactions known (Kd $\approx 10^{-14}$ to 10^{-16} M) (Laitinen et al., 2006). The protein-ligand complex stands out by a high degree of shape complementarity and an extensive network of hydrogen bonds formed between both binding partners. One of the 24 structures of biotin bound to core streptavidin (4GD9) shows the impact of the cutting of a binding loop on the conformation of the bound ligand (Figure S4; Le Trong et al., 2013). Another structure (2IZJ) shows subtle structural changes of the streptavidin-biotin complex induced by a low pH that stabilizes intersubunit salt bridges (**Figure 5A**; orange carbon atoms; Katz, 1997).

Six crystal structures of avidin from chicken (wild type and mutants) and one of engineered avidin (2C4I) are also included in the dataset. Avidin is loosely related to streptavidin, with an equally high affinity to biotin and a very similar binding site (Figure S4). As expected, biotin binds to this protein in a conformation that is very similar to those predominantly observed for complexes with streptavidin.

Biotin-protein ligase (1WPY, 2EJ9, 2EJF, 2DTH, 2FYK, and 2ZGW) and biotin carboxylase (3G8C) share very low structural similarity with streptavidin and with each other. The conformations observed for biotin bound to biotin-protein ligase (**Figure 5A**; violet carbon atoms) are virtually identical among each other but differ by an RMSD of 1.1 Å from the predominant conformation observed in the Sperrylite Dataset. In particular, the angle of the alkyl chain leaving the ring system differs by around 103° from that observed for biotin bound to streptavidin. A third conformer of biotin is observed in complex with *E. coli* biotin carboxylase (3G8C; **Figure 5A**; green carbon atoms), with an RMSD of 0.9 Å measured against any of the streptavidin-bound conformers. Despite substantial structural differences observed among the various different biotin-binding proteins, the non-covalent interactions formed between biotin and the target protein are largely conserved (**Figures 5B–D**).

Sapropterin

Sapropterin (tetrahydrobiopterin, *H4B*) is an approved drug for the treatment of tetrahydrobiopterin deficiency. It is an essential cofactor for the synthesis of nitric oxide and the hydroxylation of phenylalanine, tyrosine and tryptophan. The PDB counts 472 complexes with sapropterin, 188 of which are of high quality.

Of the high-quality conformers of sapropterin, all but three are extremely similar to each other (median RMSD of less than 0.1 Å; Figure S5A). All of these highly similar sapropterin conformers are bound to nitric oxide synthase, from five different species (human, rat, mouse, cattle and *Bacillus subtilis*). The exceptions are the conformers bound to human phenylalanine hydroxylase (1MMK, 1MMT and 1J8U), and differ by an RMSD of 0.7 Å from the conformer in human nitric oxide synthase (4D1N, Figure S5B). The sequence identity between human phenylalanine hydroxylase and human nitric oxide synthase is less than 15%. The slightly different conformer bound to phenylalanine hydroxylase is stabilized by hydrophobic interactions (Figure S5C).

Cholic Acid

Cholic acid (*CHD*) is one of the major bile acids produced from cholesterol in the liver. It is approved for the treatment of bile acid synthesis disorders and as an adjunctive treatment of peroxisomal disorders.

Thirteen of the 74 available crystal structures that include cholic acid are of high quality. Twelve thereof are from eukaryotic

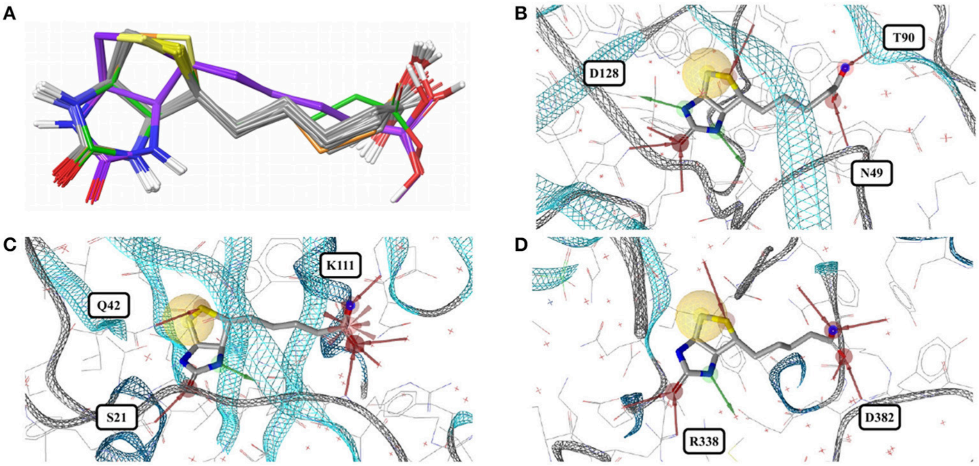


FIGURE 5 | (A) Superposition of 43 structures of biotin (BTN) bound to core streptavidin (gray carbon atoms), *E. coli* biotin carboxylase (3G8C; green carbon atoms), biotin-protein ligase (1WPY; violet carbon atoms), and streptavidin-biotin at low pH (2IZJ; orange carbon atoms). The binding modes observed for biotin in complex with **(B)** core streptavidin from *Streptomyces avidinii* (3WYP), **(C)** biotin-protein ligase from *Pyrococcus horikoshii* (1WPY) and **(D)** *E. coli* biotin carboxylase (3G8C) are very similar.

proteins, including alcohol dehydrogenase, ferrochelatase, cytochrome c oxidase and bile acid-binding proteins; one structure is of choloylglycine hydrolase from *Clostridium perfringens* (2RLC).

Some pockets of cholic acid-binding proteins can accommodate more than a single cholic acid molecule, as observed e.g., in structures of the chicken liver basic fatty acid-binding protein (1TW4) and the zebrafish liver bile acid-binding proteins (2QO5).

Given the rigid scaffold of steroids it is not surprising that, despite in part low sequence identity between the cholic acid-binding proteins, the observed ligand conformations (i.e., those bound to the deepest part of their respective binding pocket) are highly similar (median RMSD = 0.6 Å; **Figure 6A**). The maximum pairwise RMSD of 1.6 Å was measured between the conformation of cholic acid in the crystal structure of the G55R mutant of zebrafish liver bile acid-binding protein (2QO6) and in human mitochondrial ferrochelatase (3W1W).

Deoxycholic Acid

Deoxycholic acid (*DXC*), a metabolic byproduct of intestinal bacteria, is a steroid acid commonly found in the bile of mammals (Ridlon et al., 2016). Deoxycholic acid is a detergent that disturbs the integrity of biological membranes and is used to isolate membrane-associated proteins. Deoxycholic acid is approved for submental fat reduction, as a safer and less invasive alternative to surgical procedures for the treatment of lipomas (Duncan and Rotunda, 2011) and for improvements of aesthetic appearance.

Of the 29 entries deposited in the PDB, 18 are of high quality. Eleven of those structures are deoxycholic acid bound to cathepsin A and have a maximum pairwise RMSD of just 0.1 Å. Because of the rigid ligand core, deoxycholic acid also binds to structurally distinct proteins in very similar conformations (**Figure 6B**). Examples from the Sperrylite Dataset include two structures of *Betula pendula* Bet v1 (a major pollen allergen; 4A81 and 4A84), a structure of subunits I and II of cytochrome c oxidase (3DTU) from *Rhodobacter sphaeroides*, a structure of choloylglycine hydrolase from *Clostridium perfringens* (2BJF), a structure of the multidrug transporter MdfA (4ZP0) from *E. coli*, and even a conformer of deoxycholic acid bound to the interface of a dimer of the cell invasion protein SipD from *Salmonella enterica* (3O01; Chatterjee et al., 2011) The maximum pairwise RMSD (0.9 Å) was measured for the ligand conformers bound to a K9E mutant of cathepsin A (4HAJ) and salmonella invasion protein D (3O01), indicated by violet carbon atoms in **Figure 6B**.

Cofactors and Cofactor Analogs

The most abundant small molecules in the Sperrylite Dataset are cofactors and their analogs. The cofactors represented by at least 10 high-quality structures can roughly be grouped into three categories: sinefungin and its analogs (S-adenosylmethionine, SAM, and S-adenosylhomocysteine, SAH; **Figure 7**), adenosine phosphates (AMP, ADP, ATP; **Figure 8**), and three cofactors without analogs listed in the dataset (glutathione, flavin mononucleotide and sapropterin). The RMSD distributions (all-against-all comparisons) for the most relevant cofactors are reported in **Figure 9**.

Sinefungin and Analogs

Sinefungin

Sinefungin (*SFG*), an analog of the cofactor substrate SAM, inhibits a wide range of methyltransferases, thereby interfering with DNA synthesis (Pugh et al., 1978). It is an antifungal antibiotic and also a known effective inhibitor of the transformation of chick embryo fibroblasts by the cancer-causing *Rous sarcoma* virus (Vedel et al., 1978).

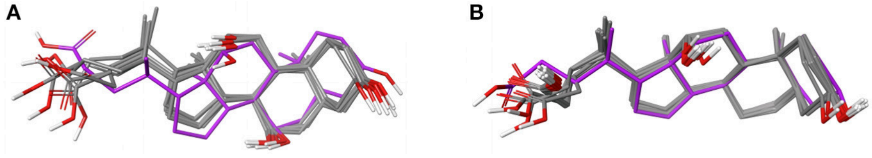


FIGURE 6 | (A) Ligand-based alignment of 13 structures of cholic acid bound to different eukaryotic proteins and choloylglycine hydrolase from *Clostridium perfringens* (2RLC; gray carbon atoms) and human mitochondrial ferrochelatase (3W1W; violet carbon atoms). **(B)** Ligand-based alignment of 16 structures of deoxycholic acid bound to structurally distinct proteins, including salmonella invasion protein D (3O01; violet carbon atoms).

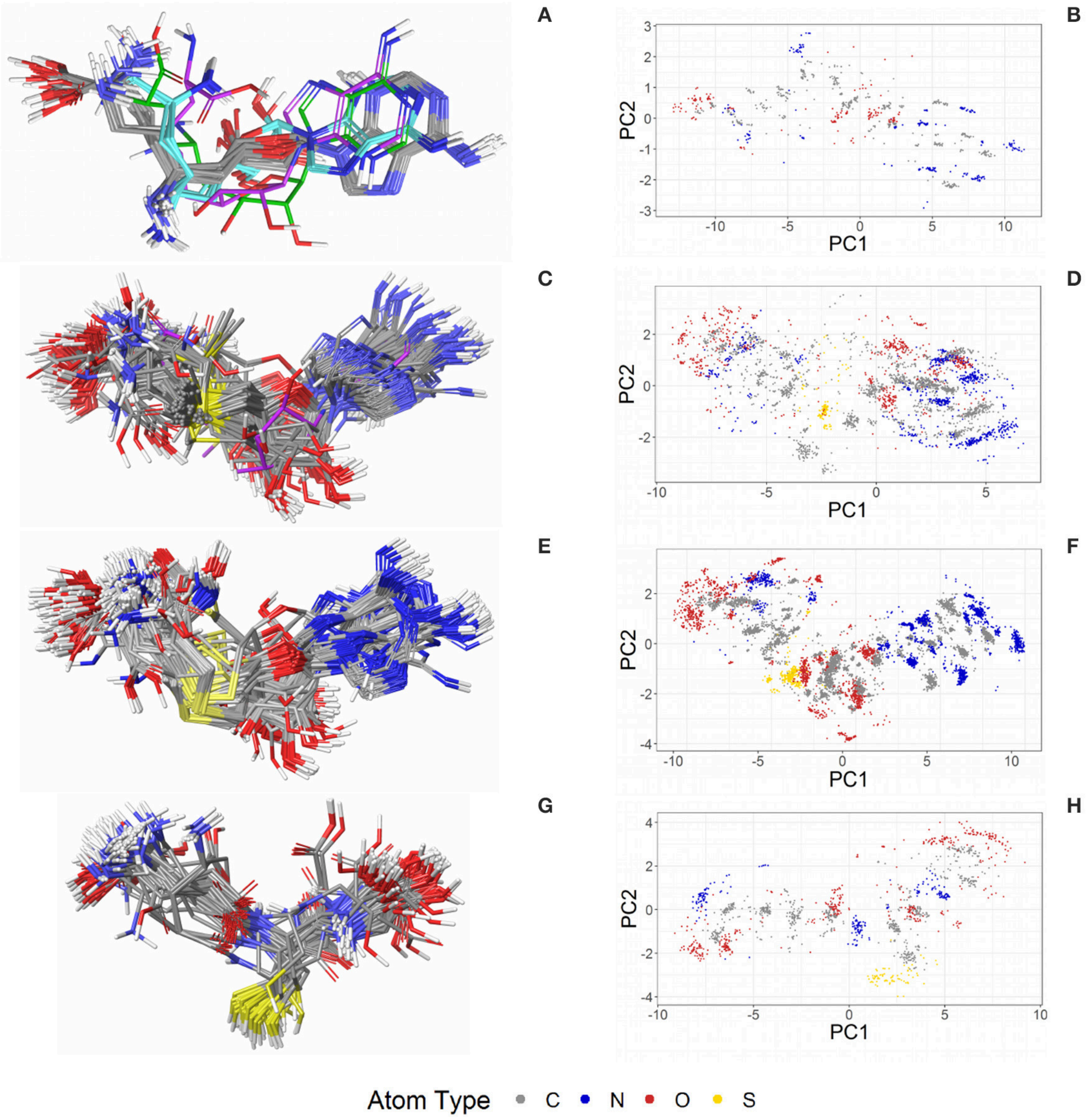


FIGURE 7 | Ligand-based alignment (left) and PCA-derived score plots (right) of **(A,B)** 30 structures of sinefungin bound to different methyltransferases (gray carbon atoms; these and all further color definitions in this caption are referring to the left panels only), ribosomal RNA small subunit methyltransferase NEP1 (3BBH; violet carbon atoms), tRNA (guanine-N(1)-)-methyltransferase (4YVH; green carbon atoms), SMYDs and SET7 lysine methyltransferase (3CBP, 3PDN, 3N71, 3QWW and 3RU0; cyan carbon atoms); **(C,D)** 123 structures of SAM bound to different methyltransferases (gray carbon atoms), tRNA(m1G37)methyltransferase (1UAK; violet carbon atoms) and yeast ribosome synthesis factor Emg1 (2V3K; green carbon atoms); **(E,F)** 311 structures of SAH and **(G,H)** 74 conformers of glutathione (GSH).

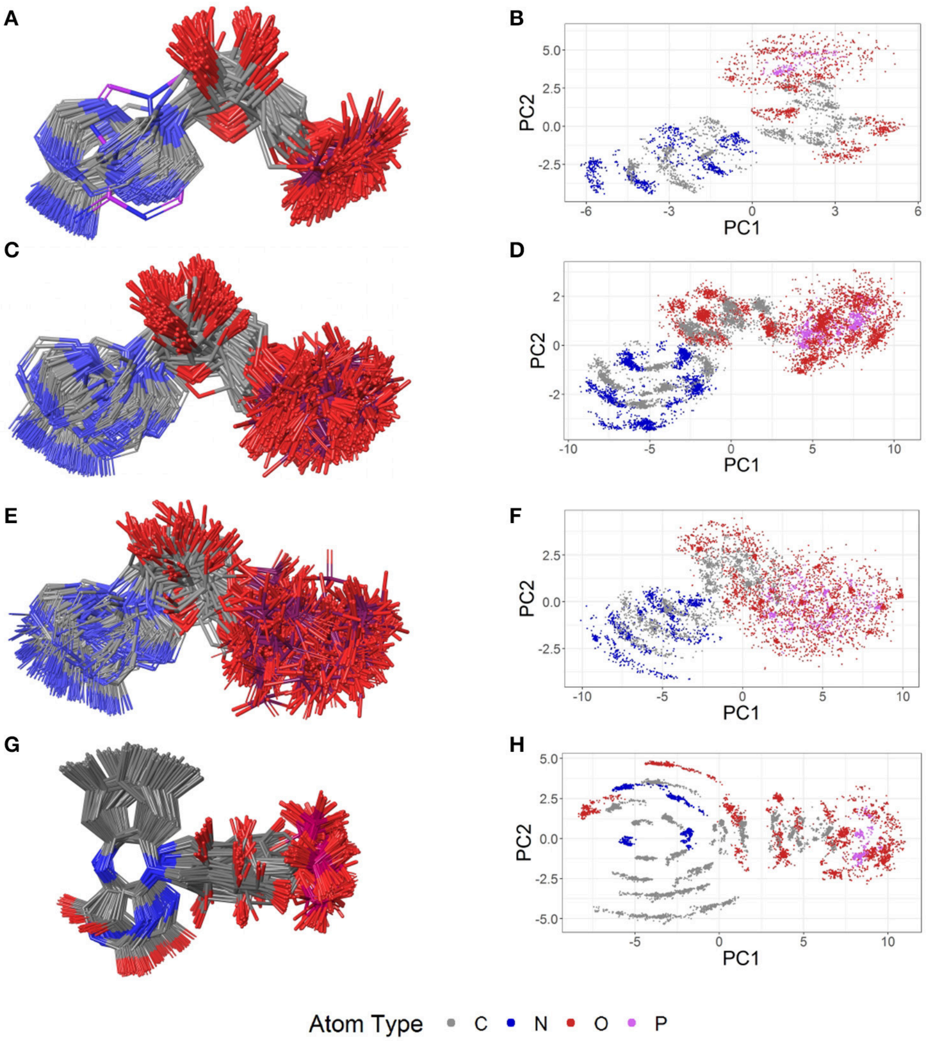


FIGURE 8 | Ligand-based alignment (left) and PCA-derived score plots (right) of **(A,B)** 171 conformers of AMP (conformer bound to adenylate kinase-related protein from *Sulfolobus solfataricus* in **(A)** with violet carbon atoms; 3LW7), **(C,D)** 462 conformers of ADP, **(E,F)** 218 conformers of ATP, and **(G,H)** 367 conformers of FMN.

The PDB lists 70 structures of sinefungin, all of them bound to methyltransferases. Thirty of these structures are of high quality. The observed conformers of sinefungin can be classified into three groups by an all-against-all comparison of their RMSDs (**Figure 9**). The largest group (**Figure 7A**; gray carbon atoms) includes 23 highly similar conformers (a representative example is given in **Figure 10A**) with a median RMSD of 0.5 Å, even though some of the proteins that these sinefungin molecules are bound to share low sequence identity (e.g., 30% for murine protein arginine N-methyltransferase 6 and the ribosomal protein L11 methyltransferase of *Thermus thermophilus*).

The second largest group consists of sinefungin conformers bound to the murine SET and MYND domains (SMYD) 1 (3N71) and 2 (3QWW), the human SMYD 3 (3PDN, 3RU0) and the SET7 lysine methyltransferase (3CBP), with RMSDs between 1.7 and 1.8 Å measured against the conformations representing the largest group (**Figure 7A**; cyan carbon atoms). Murine SMYD 1 (3N71) and human SET7 lysine methyltransferase (3CBP) have less than 15% sequence identity but bind sinefungin in very similar conformations (RMSD 0.3 Å).

Distinct conformations of sinefungin are observed for a complex with *Haemophilus influenzae* tRNA (guanine-N(1)-)methyltransferase (4YVH; **Figure 7A**; green atoms) and a complex with the ribosomal RNA small subunit methyltransferase NEP1 (3BBH, **Figure 7A**; violet carbon atoms, and **Figure 10B**) from *Methanocaldococcus jannaschii,* with RMSDs measured to the most abundantly observed conformation of 3.1 and 3.6 Å, respectively. In both cases the ligand conformation is stabilized by a hydrogen bond formed between the ligand's carboxyl group and the protein backbone.

S-Adenosylmethionine

SAM (*SAM*) is a cofactor that functions as a methyl donor in methyltransferases. It is essential for the methylation of proteins, DNA, lipids and small molecules. The bulk of SAM is generated in the liver, but all mammalian cells use it as an intermediate in the methionine-homocysteine cycle (Mato et al., 2013). SAM is also involved in the synthesis of many other endogenous metabolites. It has wide-ranging anti-inflammatory activity (Pfalzer et al., 2014) and, since its synthesis is depressed in chronic liver diseases, there has been considerable interest in its therapeutic use (Anstee and Day, 2012; Guo et al., 2015). S-adenosylmethionine is used as a drug for the treatment of depression, liver disorders, fibromyalgia, and osteoarthritis.

Four hundred ten structures listed in the PDB contain SAM. For example, almost all crystal structures of flavivirus methyltransferases contain SAM (because the molecule co-purifies with the enzymes (Noble et al., 2014). There are 119 high-quality SAM-containing structures present in the Sperrylite Dataset. Many of these conformers are similar, with an overall median RMSD of 0.6 Å (**Figures 7C,D**). Even conformers bound to proteins sharing a low sequence identity (e.g., 19% in the case of *Aeropyrum pernix* fibrillarin, 4DF3, and human NSUN5, 2B9E), have RMSDs of just 0.5 Å. The all-against-all RMSD comparison shows a partitioning into three groups that are mainly determined by the torsion angles between the adenine and the ribose and to the torsion angles including the sulfonium linkage (**Figure 7**). The highest RMSD measured between any pair of SAM conformers is 3.3 Å, which was measured for the ligand in complex with *Haemophilus influenzae* tRNA(m1G37)methyltransferase (1UAK; **Figure 7C**; violet carbon atoms) and with SAM methyltransferase from *Ruegeria pomeroyi* (3IHT).

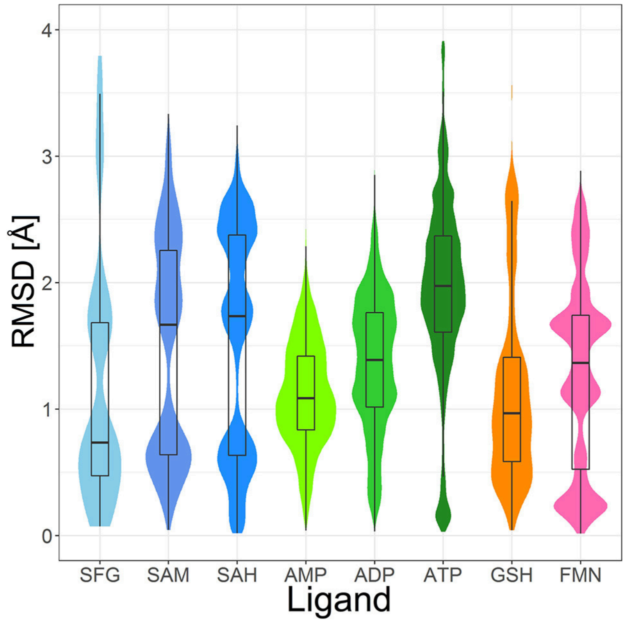


FIGURE 9 | Violin plot including box plots of the RMSD distributions of high-quality, protein-bound conformations of sinefungin (SFG), SAM, SAH, AMP, ADP, ATP, GSH and FMN. The width of each violin plot for a certain RMSD value indicates how often the specific value occurs in the pairwise comparison of all conformers.

S-Adenosyl-L-homocysteine

The strong product inhibitor SAH (*SAH*) is released in all SAM-dependent methyltransferase reactions (Tehlivets et al., 2013). The ratio of SAM to SAH controls the activity of methyltransferase enzymes ("methylation ratio"; Schatz et al., 1977).

The PDB lists 784 structures including SAH, of which an unusually high proportion (40%; 311 structures) is of high quality (**Figure 7**). These represent a highly diverse set of proteins from all three domains of organisms in nature. Most of the structures are of human (73 structures) and *Pyrococcus horikoshii* (72 structures) proteins.

Many of the SAH conformations are highly similar, with an overall median RMSD of 0.6 Å. The all-against-all RMSD comparison shows three groups of conformations and an overall spread very similar to that observed for SAM (**Figure 9**). As shown in **Figure 7**, the conformations observed for SAM and SAH are similar. Also, all conformations of sinefungin are closely represented by at least one conformation of SAM and SAH.

The largest difference observed among the SAH conformations was measured between a coiled conformer bound to *Haemophilus influenzae* tRNA (Guanine-N(1)-)-methyltransferase (1UAL) and a mostly stretched conformer bound to *E. coli* ribosomal RNA large subunit methyltransferase L (3V97) with an RMSD of 3.2 Å.

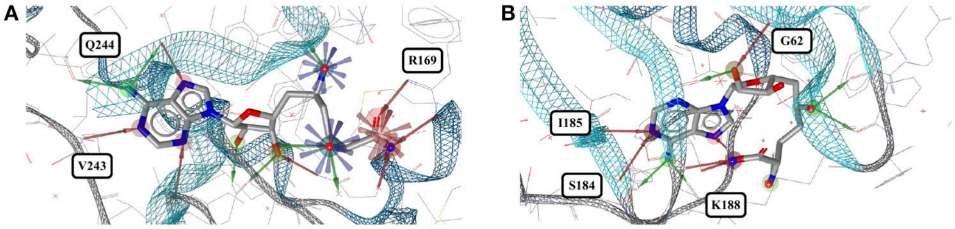


FIGURE 10 | (A) A typical conformer of sinefungin bound to human histone-arginine methyltransferase CARM1 (2Y1W) and **(B)** the coiled conformer in the ribosomal RNA small subunit methyltransferase NEP1 from *Methanocaldococcus jannaschii* (3BBH).

Glutathione

The tripeptide glutathione (GSH; *GSH*) is a cofactor of various different enzymes and a defensive reagent against toxic xenobiotics. Of the 360 entries with glutathione listed in the PDB, 74 structures are of high quality. These high-quality structures cover glutathione bound to 10 different proteins (**Figures 7G,H**). Most of the GSH conformers have a pairwise RMSD between 0.6 and 1.6 Å (**Figure 9**). The two most distinct conformers of glutathione observed in the Sperrylite Dataset are an unusually stretched conformer bound to a putative branched-chain amino acid ABC transporter from *Chromobacterium violaceum* (4PYR, **Figure 11A**) and an extremely coiled conformer bound to human mPGES-1 (4YL1, **Figure 11B**), with an RMSD of 3.6 Å. Nevertheless, their interaction patterns show similarities. Glutathione transferases are represented by 46 high-quality structures. These are mostly similar and have a median RMSD of less than 0.5 Å (**Figures 7G,H**).

Adenosine Phosphates

ATP functions as the most important molecule for intracellular storage and transport of chemical energy. It has many crucial roles in metabolism and is also a neurotransmitter. During metabolic processes, ATP is converted into adenosine diphosphate (ADP) and, subsequently, adenosine monophosphate (AMP), thereby releasing the stored energy.

Adenosine monophosphate

Out of the 575 complexes with AMP (*AMP*) found in the PDB, 171 conformers are of high quality. AMP has four rotatable bonds and the median RMSD measured between all high-quality conformers is 0.8 Å. The all-against-all comparison of AMP conformers results in a wide spread of the RMSD values (**Figure 9**). The flexibility of the molecule is mostly limited to the phosphate group (**Figures 8A,B**). The maximum RMSD of 2.5 Å was measured between an extremely coiled conformer bound to an adenylate kinase-related protein from *Sulfolobus solfataricus* (3LW7; **Figure 8A**, violet carbon atoms; **Figure 12A**) and the stretched conformer bound to NTPDase1 from *Legionella pneumophila* (4BRN; **Figure 12B**).

Adenosine diphosphate

Out of the 1,810 entries including ADP (*ADP*) in the PDB, 462 conformers are of high quality. Despite an additional phosphate group and a total of six rotatable bonds, the conformational space covered by ADP is very similar to that covered by AMP (**Figures 8C,D**). This similarity is reflected in the median RMSD of 0.9 Å between the conformers of ADP and a similar overall spread in the all-against-all comparison (**Figure 9**). The two most different ADP conformers in the Sperrylite Dataset are those bound to tryptophanyl-tRNA synthetase from *Campylobacter jejuni* (3TZL; **Figure 13A**) and an Stt7 homolog from *Micromonas algae* (4IX6; **Figure 13B**), with an RMSD of 2.9 Å.

Adenosine triphosphate

Only 218 conformers out of the 1,079 structures of the PDB containing ATP (*ATP*) were of high quality. In all structures of ATP included in the Sperrylite Dataset, the N-glycosidic bond is found in an anti-orientation. With its eight rotatable bonds ATP is more flexible than the previously discussed adenosine

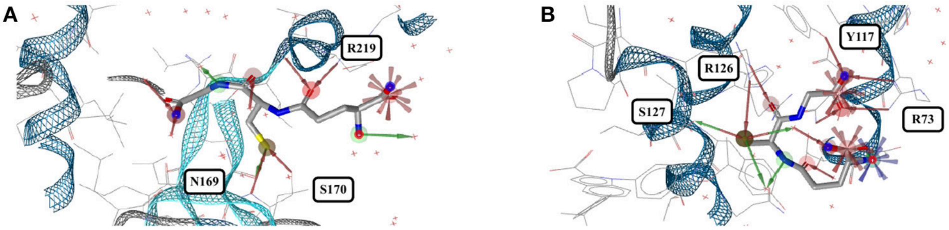


FIGURE 11 | (A) The most stretched conformer of glutathione bound to an ABC transporter from *Chromobacterium violaceum* (4PYR) and **(B)** an unusually coiled conformer of glutathione bound to human mPGES-1 (4YL1).

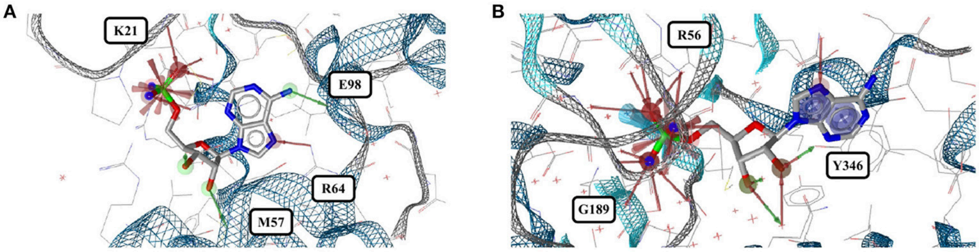


FIGURE 12 | (A) Unusually coiled conformer of AMP bound to adenylate kinase-related protein of *Sulfolobus solfataricus* (3LW7) and **(B)** the most stretched conformer in *Legionella pneumophila* NTPDase1 (4BRN).

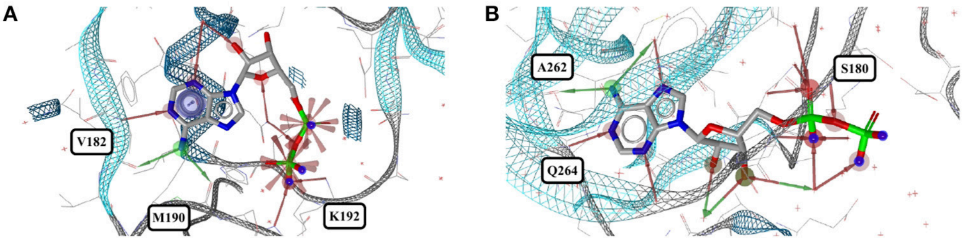


FIGURE 13 | The most distinct conformers of ADP in the Sperrylite Dataset are the coiled conformer from **(A)** tryptophanyl-tRNA synthetase from *Campylobacter jejuni* (3TZL; sodium ion in light blue) and **(B)** the stretched conformer from a Stt7 homolog from *Micromonas algae* (4IX6).

phosphates. This results in a median RMSD of 1.6 Å among the ATP structures of the Sperrylite dataset (as compared to a median RMSD of 0.9 Å measured for ADP) and a distinct spread of the RMSD values in the all-against-all comparison (**Figure 9**). The maximum pairwise RMSD was 3.9 Å, measured between ATP conformers from human lysyl-tRNA synthetase (3BJU) and *Drosophila melanogaster* Wiskott-Aldrich syndrome protein homology 2 (3MN6).

ATP is observed in an extended conformation in most structures (**Figures 8G,H**), which is in agreement with earlier studies (Moodie and Thornton, 1993; Stockwell and Thornton, 2006; Bojovschi et al., 2012; Stegemann and Klebe, 2012). As reported also by Stockwell and Thornton (Stockwell and Thornton, 2006), some conformers are bent to an extent that the terminal phosphate atoms are almost in van der Waals contact with the adenine ring. Examples of ATP in bent conformations include complexes with the aspartyl-tRNA synthetase from *Pyrococcus kodakaraensis* (1B8A; Figure S6) and the ribonucleotide reductase protein R1 from *E. coli* (3R1R).

Flavin Mononucleotide

Flavin mononucleotide (FMN; *FMN*) is the prosthetic group of various oxidoreductases (including NADH dehydrogenase), as well as a cofactor in biological blue-light photoreceptors (Froehlich et al., 2002; Schwerdtfeger and Linden, 2003). Blue-light receptors in plants (phototropins), for example, employ flavin mononucleotide as the chromophore for their light sensing function (He, 2002).

Its frequent occurrence as a prosthetic group and a cofactor result in flavin mononucleotide's presence in 919 structures deposited in the PDB, among which 367 conformers of FMN are of high quality. Despite having seven rotatable bonds, most structures show extended, similar conformations (**Figures 8G,H**), with a median RMSD of 0.9 Å. The all-against-all comparison reveals four groups of conformers, with peaks observed in the RMSD distribution around 0.3, 1.2, 1.7, and 2.4 Å (**Figure 9**). These peaks correspond to an accumulation of conformers with similar torsion angles of the side chain. The maximum RMSD of 2.9 Å was observed between the conformation of FMN in *E. coli* pyridoxine 5′-phosphate oxidase (1JNW) and in human glycolate oxidase (2RDU), with the sidechain bent into opposing directions.

CONCLUSIONS

The Sperrylite Dataset presented in this work is a complete subset of high-quality conformations of protein-bound ligands extracted from the PDB. This dataset resulted from a multi-step data processing and filtering procedure that, most importantly, also includes an automated approach for the evaluation of the support of individual atom positions by the electron density. The Sperrylite Dataset consists of a total of 10,936 high-quality structures of 4,548 unique ligands. Ninety-one of those ligands are each represented by a minimum of ten structures, and among these only a (very) weak correlation was observed between the number of rotatable bonds of a molecule and its overall variability (measured as the minimum median RMSD; $R^2 = 0.126$). Sixty-nine out of the 91 ligands had at least two distinct conformations (defined as RMSD above 1 Å).

A representative subset of 17 approved drugs and cofactors was analyzed in detail to determine the conformational variability of protein-bound conformations of small molecules. For all of the analyzed small-molecule drugs and some of the cofactors, a clear trend for the formation of few clusters of highly similar conformers was observed. Similar conformers were observed for proteins with similar binding sites, mostly independent of the overall protein sequence identity (which is in agreement with the findings of, e.g., Sturm et al., 2012). A particularly interesting example is imatinib, which was found to adopt highly similar conformations when binding to different tyrosine kinases (even to those sharing low overall sequence identity) but to adopt a distinct conformation upon binding to quinone reductase 2. For cofactors, a clear trend for extended conformations was observed, which is in agreement with previous works (Moodie and Thornton, 1993; Stockwell and Thornton, 2006; Bojovschi et al., 2012; Stegemann and Klebe, 2012). A few cases of strongly coiled conformers of cofactors were also observed. This result is well in line with earlier reports (Stockwell and Thornton, 2006).

It is clear that the currently available structural data on protein-bound ligands is still too limited to allow us to gain a full understanding of the bioactive space of small molecules. However, for several cofactors a large number of conformers observed in complex with dozens of proteins are available to date and provide valuable insight into the bioactive conformational space and the prevalence of bioactive conformations of small molecules. With an automated workflow for the extraction of high-quality ligand structures from the PDB in place, it is expected that the ever increasing amount of data will allow a more detailed understanding of, e.g., conformational preferences, ligand promiscuity, or the relationship between the bioactive conformational space of small molecules and the structural diversity of binding pockets.

DATA AVAILABILITY

The dataset generated for this study can be found at: http://www.zbh.uni-hamburg.de/sperrylite_dataset.

AUTHOR CONTRIBUTIONS

JK and N-OF: conceived the work; N-OF and MS: conducted the computational studies. All authors contributed to the interpretation of the data and the writing of the manuscript. All authors have given approval to the final version of the paper.

ACKNOWLEDGMENTS

The authors thank Christina de Bruyn Kops for discussion and proofreading of the manuscript. MS was supported by the Erasmus+ Programme of the European Commission.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fchem.2018.00068/full#supplementary-material

REFERENCES

Alton, G. R., and Lunney, E. A. (2008). Targeting the unactivated conformations of protein kinases for small molecule drug discovery. *Expert Opin. Drug Discov*. 3, 595–605. doi: 10.1517/17460441.3.6.595

Altschul, S. F., Gish, W., Miller, W., Myers, E. W., and Lipman, D. J. (1990). Basic local alignment search tool. *J. Mol. Biol*. 215, 403–410. doi: 10.1016/S0022-2836(05)80360-2

Anstee, Q. M., and Day, C. P. (2012). S-adenosylmethionine (SAMe) therapy in liver disease: a review of current evidence and clinical utility. *J. Hepatol*. 57, 1097–1109. doi: 10.1016/j.jhep.2012.04.041

Atwell, S., Adams, J. M., Badger, J., Buchanan, M. D., Feil, I. K., Froning, K. J., et al. (2004). A novel mode of Gleevec binding is revealed by the structure of spleen tyrosine kinase. *J. Biol. Chem*. 279, 55827–55832. doi: 10.1074/jbc.M409792200

Berman, H. M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T. N., Weissig, H., et al. (2000). The protein data bank. *Nucleic Acids Res*. 28, 235–242. doi: 10.1093/nar/28.1.235

Bojovschi, A., Liu, M. S., and Sadus, R. J. (2012). Conformational dynamics of ATP/Mg:ATP in motor proteins via data mining and molecular simulation. *J. Chem. Phys*. 137:075101. doi: 10.1063/1.4739308

Boström, J. (2001). Reproducing the conformations of protein-bound ligands: a critical evaluation of several popular conformational searching tools. *J. Comput. Aided Mol. Des*. 15, 1137–1152. doi: 10.1023/A:1015930826903

Boström, J., Norrby, P. O., and Liljefors, T. (1998). Conformational energy penalties of protein-bound ligands. *J. Comput. Aided Mol. Des*. 12, 383–396. doi: 10.1023/A:1008007507641

Brameld, K. A., Kuhn, B., Reuter, D. C., and Stahl, M. (2008). Small molecule conformational preferences derived from crystal structure data. A medicinal chemistry focused analysis. *J. Chem. Inf. Model*. 48, 1–24. doi: 10.1021/ci7002494

Buth, J. M., Steen, P. O., Sueper, C., Blumentritt, D., Vikesland, P. J., Arnold, W. A., et al. (2010). Dioxin photoproducts of triclosan and its chlorinated derivatives in sediment cores. *Environ. Sci. Technol*. 44, 4545–4551. doi: 10.1021/es1001105

Camacho, C., Coulouris, G., Avagyan, V., Ma, N., Papadopoulos, J., Bealer, K., et al. (2009). BLAST+: architecture and applications. *BMC Bioinformatics* 10:421. doi: 10.1186/1471-2105-10-421

Carey, D. E., and McNamara, P. J. (2014). The impact of triclosan on the spread of antibiotic resistance in the environment. *Front. Microbiol*. 5:780. doi: 10.3389/fmicb.2014.00780

Chakravarty, S., and Kannan, K. K. (1994). Drug-protein interactions. Refined structures of three sulfonamide drug complexes of human carbonic anhydrase I enzyme. *J. Mol. Biol*. 243, 298–309. doi: 10.1006/jmbi.1994.1655

Chatterjee, S., Zhong, D., Nordhues, B. A., Battaile, K. P., Lovell, S., and De Guzman, R. N. (2011). The crystal structures of the Salmonella type III secretion system tip protein SipD in complex with deoxycholate and chenodeoxycholate. *Protein Sci*. 20, 75–86. doi: 10.1002/pro.537

Cruickshank, D. W. (1999). Remarks about protein structure precision. *Acta Crystallogr. D Biol. Crystallogr*. 55, 583–601. doi: 10.1107/S0907444998012645

Duncan, D., and Rotunda, A. M. (2011). Injectable therapies for localized fat loss: state of the art. *Clin. Plast. Surg*. 38, 489–501, vii. doi: 10.1016/j.cps.2011.02.005

Dym, O., and Eisenberg, D. (2001). Sequence-structure analysis of FAD-containing proteins. *Protein Sci*. 10, 1712–1728. doi: 10.1110/ps.12801

Foster, C. E., Bianchet, M. A., Talalay, P., Zhao, Q., and Amzel, L. M. (1999). Crystal structure of human quinone reductase type 2, a metalloflavoprotein. *Biochemistry* 38, 9881–9886. doi: 10.1021/bi990799v

Friedrich, N.-O., de Bruyn Kops, C., Flachsenberg, F., Sommer, K., Rarey, M., and Kirchmair, J. (2017a). Benchmarking commercial conformer ensemble generators. *J. Chem. Inf. Model*. 57, 2719–2728. doi: 10.1021/acs.jcim.7b00505

Friedrich, N.-O., Meyder, A., de Bruyn Kops, C., Sommer, K., Flachsenberg, F., Rarey, M., et al. (2017b). High-quality dataset of protein-bound ligand conformations and its application to benchmarking conformer ensemble generators. *J. Chem. Inf. Model*. 57, 529–539. doi: 10.1021/acs.jcim.6b00613

Froehlich, A. C., Liu, Y., Loros, J. J., and Dunlap, J. C. (2002). White Collar-1, a circadian blue light photoreceptor, binding to the frequency promoter. *Science* 297, 815–819. doi: 10.1126/science.1073681

Günther, S., Senger, C., Michalsky, E., Goede, A., and Preissner, R. (2006). Representation of target-bound drugs by computed conformers: implications for conformational libraries. *BMC Bioinformatics* 7:293. doi: 10.1186/1471-2105-7-293

Guo, T., Chang, L., Xiao, Y., and Liu, Q. (2015). S-adenosyl-L-methionine for the treatment of chronic liver disease: a systematic review and meta-analysis. *PLoS ONE* 10:e0122124. doi: 10.1371/journal.pone.0122124

He, M. W., Lee, P. S., and Sweeney, Z. K. (2015). Promiscuity and the conformational rearrangement of drug-like molecules: insight from the protein data bank. *Chem. Med. Chem*. 10, 238–244. doi: 10.1002/cmdc.201402389

He, Q. (2002). White collar-1, a DNA binding transcription factor and a light sensor. *Science* 297, 840–843. doi: 10.1126/science.1072795

Heath, R. J., Rubin, J. R., Holland, D. R., Zhang, E., Snow, M. E., and Rock, C. O. (1999). Mechanism of triclosan inhibition of bacterial fatty acid synthesis. *J. Biol. Chem*. 274, 11110–11114. doi: 10.1074/jbc.274.16.11110

Inoi, K., Goto, S., Nomura, S., Isobe, K., Nawa, A., Okamoto, T., et al. (1995). Aminopeptidase inhibitor ubenimex (bestatin) inhibits the growth of human choriocarcinoma in nude mice through its direct cytostatic activity. *Anticancer Res*. 15, 2081–2087.

Katz, B. A. (1997). Binding of biotin to streptavidin stabilizes intersubunit salt bridges between Asp61 and His87 at low pH. *J. Mol. Biol*. 274, 776–800. doi: 10.1006/jmbi.1997.1444

Kaur, I. P., Smitha, R., Aggarwal, D., and Kapil, M. (2002). Acetazolamide: future perspective in topical glaucoma therapeutics. *Int. J. Pharm.* 248, 1–14. doi: 10.1016/S0378-5173(02)00438-6

Kim, K.-H., Ha, B. H., Kim, S. J., Hong, S. K., Hwang, K. Y., and Kim, E. E. (2011). Crystal structures of Enoyl-ACP reductases I (FabI) and III (FabL) from *B. subtilis*. *J. Mol. Biol.* 406, 403–415. doi: 10.1016/j.jmb.2010.12.003

King, N. M., Prabu-Jeyabalan, M., Nalivaika, E. A., Wigerinck, P., de Béthune, M.-P., and Schiffer, C. A. (2004). Structural and thermodynamic basis for the binding of TMC114, a next-generation human immunodeficiency virus type 1 protease inhibitor. *J. Virol.* 78, 12012–12021. doi: 10.1128/JVI.78.21.12012-12021.2004

Laitinen, O. H., Hytönen, V. P., Nordlund, H. R., and Kulomaa, M. S. (2006). Genetically engineered avidins and streptavidins. *Cell. Mol. Life Sci.* 63, 2992–3017. doi: 10.1007/s00018-006-6288-z

Le Trong, I., Chu, V., Xing, Y., Lybrand, T. P., Stayton, P. S., and Stenkamp, R. E. (2013). Structural consequences of cutting a binding loop: two circularly permuted variants of streptavidin. *Acta Crystallogr. D Biol. Crystallogr.* 69, 968–977. doi: 10.1107/S0907444913003855

Levy, C. W., Roujeinikova, A., Sedelnikova, S., Baker, P. J., Stuitje, A. R., Slabas, A. R., et al. (1999). Molecular basis of triclosan activity. *Nature* 398, 383–384. doi: 10.1038/18803

Liebeschuetz, J., Hennemann, J., Olsson, T., and Groom, C. R. (2012). The good, the bad and the twisted: a survey of ligand geometry in protein crystal structures. *J. Comput. Aided Mol. Des.* 26, 169–183. doi: 10.1007/s10822-011-9538-6

Mato, J. M., Martínez-Chantar, M. L., and Lu, S. C. (2013). S-adenosylmethionine metabolism and liver disease. *Ann. Hepatol.* 12, 183–189.

Meyder, A., Nittinger, E., Lange, G., Klein, R., and Rarey, M. (2017). Estimating electron density support for individual atoms and molecular fragments in X-ray structures. *J. Chem. Inf. Model.* 57, 2437–2447. doi: 10.1021/acs.jcim.7b00391

Moodie, S. L., and Thornton, J. M. (1993). A study into the effects of protein binding on nucleotide conformation. *Nucleic Acids Res.* 21, 1369–1380. doi: 10.1093/nar/21.6.1369

Nankya-Kitaka, M. F., Curley, G. P., Gavigan, C. S., Bell, A., and Dalton, J. P. (1998). *Plasmodium chabaudi* chabaudi and *P. falciparum:* inhibition of aminopeptidase and parasite growth by bestatin and nitrobestatin. *Parasitol. Res.* 84, 552–558. doi: 10.1007/s004360050447

Nicklaus, M. C., Wang, S., Driscoll, J. S., and Milne, G. W. (1995). Conformational changes of small molecules binding to proteins. *Bioorg. Med. Chem.* 3, 411–428. doi: 10.1016/0968-0896(95)00031-B

Noble, C. G., Li, S.-H., Dong, H., Chew, S. H., and Shi, P.-Y. (2014). Crystal structure of dengue virus methyltransferase without S-adenosyl-L-methionine. *Antiviral Res.* 111, 78–81. doi: 10.1016/j.antiviral.2014.09.003

Perola, E., and Charifson, P. S. (2004). Conformational analysis of drug-like molecules bound to proteins: an extensive study of ligand reorganization upon binding. *J. Med. Chem.* 47, 2499–2510. doi: 10.1021/jm030563w

Pfalzer, A. C., Choi, S.-W., Tammen, S. A., Park, L. K., Bottiglieri, T., Parnell, L. D., et al. (2014). S-adenosylmethionine mediates inhibition of inflammatory response and changes in DNA methylation in human macrophages. *Physiol. Genomics* 46, 617–623. doi: 10.1152/physiolgenomics.00056.2014

Pugh, C. S., Borchardt, R. T., and Stone, H. O. (1978). Sinefungin, a potent inhibitor of virion mRNA(guanine-7-)-methyltransferase, mRNA(nucleoside-2'-)-methyltransferase, and viral multiplication. *J. Biol. Chem.* 253, 4075–4077.

Reddy, E. P., and Aggarwal, A. K. (2012). The ins and outs of bcr-abl inhibition. *Genes Cancer* 3, 447–454. doi: 10.1177/1947601912462126

Reynolds, C. H. (2014). Protein-ligand cocrystal structures: we can do better. *ACS Med. Chem. Lett.* 5, 727–729. doi: 10.1021/ml500220a

Ridlon, J. M., Harris, S. C., Bhowmik, S., Kang, D.-J., and Hylemon, P. B. (2016). Consequences of bile salt biotransformations by intestinal bacteria. *Gut Microbes* 7, 22–39. doi: 10.1080/19490976.2015.1127483

Sakuraya, M., Tamura, J., Itoh, K., Kubota, K., and Naruse, T. (2000). Aminopeptidase inhibitor ubenimex inhibits the growth of leukaemic cell lines and myeloma cells through its cytotoxicity. *J. Int. Med. Res.* 28, 214–221. doi: 10.1177/147323000002800503

Schatz, R. A., Vunnam, C. R., and Sellinger, O. Z. (1977). S-Adenosyl-L-homocysteine in brain: regional concentrations, catabolism, and the effects of methionine sulfoximine. *Neurochem. Res.* 2, 27–38. doi: 10.1007/BF00966019

Schwerdtfeger, C., and Linden, H. (2003). VIVID is a flavoprotein and serves as a fungal blue light photoreceptor for photoadaptation. *EMBO J.* 22, 4846–4855. doi: 10.1093/emboj/cdg451

Seeliger, D., and de Groot, B. L. (2010). Conformational transitions upon ligand binding: holo-structure prediction from apo conformations. *PLoS Comput. Biol.* 6:e1000634. doi: 10.1371/journal.pcbi.1000634

Singh, N. J., Shin, D., Lee, H. M., Kim, H. T., Chang, H.-J., Cho, J. M., et al. (2011). Structural basis of triclosan resistance. *J. Struct. Biol.* 174, 173–179. doi: 10.1016/j.jsb.2010.11.008

Sommer, K., Friedrich, N.-O., Bietz, S., Hilbig, M., Inhester, T., and Rarey, M. (2016). UNICON: a powerful and easy-to-use compound library converter. *J. Chem. Inf. Model.* 56, 1105–1111. doi: 10.1021/acs.jcim.6b00069

Stegemann, B., and Klebe, G. (2012). Cofactor-binding sites in proteins of deviating sequence: comparative analysis and clustering in torsion angle, cavity, and fold space. *Proteins* 80, 626–648. doi: 10.1002/prot.23226

Stockwell, G. R., and Thornton, J. M. (2006). Conformational diversity of ligands bound to proteins. *J. Mol. Biol.* 356, 928–944. doi: 10.1016/j.jmb.2005.12.012

Sturm, N., Desaphy, J., Quinn, R. J., Rognan, D., and Kellenberger, E. (2012). Structural insights into the molecular basis of the ligand promiscuity. *J. Chem. Inf. Model.* 52, 2410–2421. doi: 10.1021/ci300196g

Tehlivets, O., Malanovic, N., Visram, M., Pavkov-Keller, T., and Keller, W. (2013). S-adenosyl-L-homocysteine hydrolase and methylation disorders: yeast as a model system. *Biochim. Biophys. Acta* 1832, 204–215. doi: 10.1016/j.bbadis.2012.09.007

Tian, W., Rockson, S. G., Jiang, X., Kim, J., Begaye, A., Shuffle, E. M., et al. (2017). Leukotriene B4 antagonism ameliorates experimental lymphedema. *Sci. Transl. Med.* 9:eaal3920. doi: 10.1126/scitranslmed.aal3920

Urabe, A., Mutoh, Y., Mizoguchi, H., Takaku, F., and Ogawa, N. (1993). Ubenimex in the treatment of acute nonlymphocytic leukemia in adults. *Ann. Hematol.* 67, 63–66. doi: 10.1007/BF01788128

Urbaczek, S., Kolodzik, A., Fischer, J. R., Lippert, T., Heuser, S., Groth, I., et al. (2011). NAOMI: on the almost trivial task of reading molecules from different file formats. *J. Chem. Inf. Model.* 51, 3199–3207. doi: 10.1021/ci200324e

Vedel, M., Lawrence, F., Robert-Gero, M., and Lederer, E. (1978). The antifungal antibiotic sinefungin as a very active inhibitor of methyltransferases and of the transformation of chick embryo fibroblasts by *Rous sarcoma* virus. *Biochem. Biophys. Res. Commun.* 85, 371–376. doi: 10.1016/S0006-291X(78)80052-7

Warren, G. L., Do, T. D., Kelley, B. P., Nicholls, A., and Warren, S. D. (2012). Essential considerations for using protein-ligand structures in drug discovery. *Drug Discov. Today* 17, 1270–1281. doi: 10.1016/j.drudis.2012.06.011

Winger, J. A., Hantschel, O., Superti-Furga, G., and Kuriyan, J. (2009). The structure of the leukemia drug imatinib bound to human quinone reductase 2 (NQO2). *BMC Struct. Biol.* 9:7. doi: 10.1186/1472-6807-9-7

Wishart, D. S., Feunang, Y. D., Guo, A. C., Lo, E. J., Marcu, A., Grant, J. R., et al. (2017). DrugBank 5.0: a major update to the DrugBank database for 2018. *Nucleic Acids Res.* 46, D1074–D1082. doi: 10.1093/nar/gkx1037

Wolber, G., and Langer, T. (2005). LigandScout: 3-D pharmacophores derived from protein-bound ligands and their use as virtual screening filters. *J. Chem. Inf. Model.* 45, 160–169. doi: 10.1021/ci049885e

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Conformator: A Novel Method for the Generation of Conformer Ensembles

Reprinted with permission from
Friedrich, N.-O.; Flachsenberg, F.; Meyder, A.; Sommer, K.; Kirchmair, J.; Rarey, M. Conformator: A Novel Method for the Generation of Conformer Ensembles. *J. Chem. Inf.* Model. **2019**, 59 (2), 731–742.

DOI: 10.1021/acs.jcim.8b00704

JCIM JOURNAL OF CHEMICAL INFORMATION AND MODELING

Article

Cite This: J. Chem. Inf. Model. 2019, 59, 731−742

pubs.acs.org/jcim

Conformator: A Novel Method for the Generation of Conformer Ensembles

Nils-Ole Friedrich,[†] Florian Flachsenberg,[†] Agnes Meyder,[†] Kai Sommer,[†] Johannes Kirchmair,[*,†,‡,§] and Matthias Rarey[*,†]

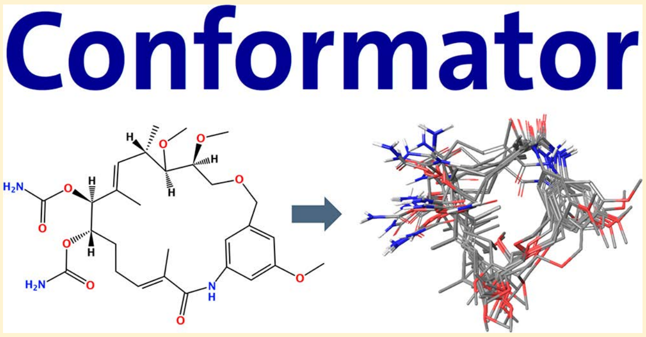
[†]Center for Bioinformatics, Universität Hamburg, Bundesstrasse 43, 20146 Hamburg, Germany

[‡]Department of Chemistry, University of Bergen, N-5020 Bergen, Norway

[§]Computational Biology Unit (CBU), University of Bergen, N-5020 Bergen, Norway

Supporting Information

ABSTRACT: Computer-aided drug design methods such as docking, pharmacophore searching, 3D database searching, and the creation of 3D-QSAR models need conformational ensembles to handle the flexibility of small molecules. Here, we present Conformator, an accurate and effective knowledge-based algorithm for generating conformer ensembles. With 99.9% of all test molecules processed, Conformator stands out by its robustness with respect to input formats, molecular geometries, and the handling of macrocycles. With an extended set of rules for sampling torsion angles, a novel algorithm for macrocycle conformer generation, and a new clustering algorithm for the assembly of conformer ensembles, Conformator reaches a median minimum root-mean-square deviation (measured between protein-bound ligand conformations and ensembles of a maximum of 250 conformers) of 0.47 Å with no significant difference to the highest-ranked commercial algorithm OMEGA and significantly higher accuracy than seven free algorithms, including the RDKit DG algorithm. Conformator is freely available for noncommercial use and academic research.



INTRODUCTION

Computational methods for 3D virtual screening, drug design, and other applications depend on the ability of algorithms to represent the conformations that small molecules adopt upon binding to biomacromolecules. In particular, fast tools such as pharmacophore-based and shape-focused screening engines make use of precalculated, multiconformational databases composed of compounds represented by (preferably small) conformer ensembles.[1−4]

The generation of representative conformer ensembles of small molecules poses significant challenges. Small molecules can have a substantial number of conformational degrees of freedom.[5] Upon binding, they may adopt conformations that are distinct from the low-energy conformations observed in the gas phase and in solution, such as strained conformations related to transition states.[6−9] On top of that, what constitutes the most appropriate algorithm for conformer ensemble generation depends on the specific purpose of use: fast algorithms may be preferred for sampling large molecular libraries for use with, for example, coarse virtual screening approaches such as pharmacophore models, whereas more time-consuming but more accurate algorithms are generally preferred for sampling small sets of molecules to be used e.g. for 3D QSAR. In consequence, a large number of conformer ensemble generators based on various algorithmic approaches are available today. They are based, among others, on random and systematic search algorithms, molecular dynamics (MD) simulations, genetic algorithms (GA), distance geometry (DG), and knowledge-based approaches.[10] Two recent studies from our laboratories[11,12] directly compare the performance of seven free (the RDKit DG algorithm[13] and the Experimental-Torsion basic Knowledge Distance Geometry algorithm (ETKDG),[14] Confab,[15] Frog2,[16] Multiconf-DOCK,[17] and the Balloon DG and GA algorithms[18]) and eight commercial (ConfGen,[19] ConfGenX,[20] cxcalc,[21] iCon,[22] MOE LowModeMD,[23] MOE Stochastic, MOE Conformation Import, and OMEGA[24]) conformer ensemble generators. These studies were the first to employ comprehensive sets of high-quality structures of protein-bound ligands for benchmarking. In particular, a newly developed cheminformatics pipeline was utilized for the fully automated extraction and curation of a complete set of 10 936 high-quality structures of protein-bound ligands (Sperrylite Dataset[5]) from a total of over 350k ligand conformations (from structures deposited in the PDB). The support of the individual atoms of all ligands by the measured electron density was quantified by the electron density score for individual atoms (EDIA[25]). On the basis of

Received: October 9, 2018
Published: February 12, 2019

DOI: 10.1021/acs.jcim.8b00704

the Sperrylite Dataset, a diverse subset of 2859 high-quality structures of unique ligands bound to their biomacromolecular targets (Platinum Diverse Dataset[12]) was compiled and provided to the scientific community for benchmarking. The outcomes of these studies show that commercial algorithms generally obtain accuracy and robustness higher than those of their free counterparts. OMEGA was confirmed as the leading commercial algorithm, with the distance geometry approach of RDKit and its knowledge-based counterpart, ETKDG, as the best-performing free alternatives.[11,12] Importantly, for all of the tested free algorithms, severe geometrical errors related to wrong bond lengths and bond angles, as well as out-of-plane errors, were detected in the generated conformations. In contrast, for most of the tested commercial algorithms, only a few instances of anomalous geometries were observed. For OMEGA and iCon, no geometric errors were identified.

In this work, we introduce Conformator as a new conformer ensemble generator that is free for noncommercial use and academic research and which addresses several of the limitations shared by most of the existing free algorithms. Conformator is a knowledge-based conformer ensemble generator that builds on concepts of the previously introduced CONFECT algorithm.[26] Major conceptual advancements of Conformator over CONFECT include a novel approach to sampling the conformational space of macrocycles, a new efficient clustering algorithm, an extended set of rules for sampling torsion angles, and capabilities for handling SMILES and InChI input. Together with the revised and extended torsion angle library of Guba et al.,[27] these advancements make Conformator a highly accurate and effective algorithm that stands out by its robustness with respect to input formats, molecular geometries, and the handling of macrocycles.

METHODS

Conformer Generation Algorithm. Conformator is a conformer ensemble generator built on established concepts of incremental construction of conformers. At its core, Conformator consists of a torsion driver enhanced by an elaborate algorithm for the assignment of torsion angles to rotatable bonds, plus a new clustering component that compiles ensembles efficiently by taking advantage of the fact that the lists of generated conformers are partially presorted. The clustering algorithm minimizes the number of comparisons between pairs of conformers that are required to effectively derive individual root-mean-square deviation (RMSD) thresholds for molecules and to compile the ensemble.

Conformator features two conformer ensemble generation modes, “Fast” and “Best”. As their names suggest, the emphasis of Fast is on computational efficiency, whereas that of Best is on accuracy. Both modes include checks that ensure chemically correct bond lengths and bond angles as well as the planarity of conjugated systems including rings.

Conformator reads molecular structures from SD and MOL2 files as well as from SMILES and InChI notations. By default, Conformator generates a new set of 3D atom coordinates as a starting point for conformation generation. Thus, Conformator does not rely on input coordinates and generates a canonicalized order of atoms and bonds (similar to canonical SMILES).[28] This representation serves as a unique and independent starting point for conformer ensemble generation (Figure 1).

After parsing, the molecule is compartmentalized at any acyclic, nonterminal single bond that is not connected to a methyl, trifluoromethyl, or nitrile group (following the concept of rigid rotor approximation). Each of these single bonds are assigned all torsion angle values of matching fragments recorded in the torsion angle library developed by Schärfer et al.[29] and revised by Guba et al.[27] As part of the construction of conformers, optimal bond angles based on the Valence Shell Electron Pair Repulsion (VSEPR) model are assigned.[30,31] Bond lengths of acyclic adjacent atoms used in the construction of conformers are calculated from the sum of covalent radii. They are adjusted for different atom types, taking into account the local molecular environment (e.g., delocalization). Details on the exact procedure and exceptions are reported in ref 26.

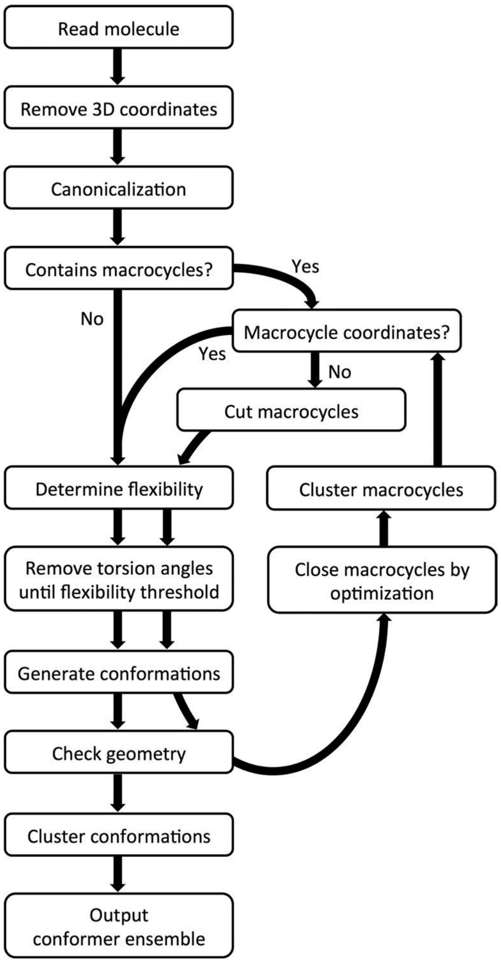


Figure 1. Schematic depiction of the conformer ensemble generation approach followed by Conformator. The boxes show the major algorithmic steps, including the loop for macrocycle conformer generation.

Once all possible torsion angles have been assigned based on this SMARTS pattern matching procedure,[32] individual torsion angle values are removed during an iterative process until the maximum number of possible conformers (based on the combination of all assigned torsion angles, neglecting potential clashes) no longer exceeds the maximum number of generated candidate conformers for clustering. The number of torsion angles assigned to a rotatable bond depends on the bond’s centricity in the molecule, the overall flexibility of the molecule, and the sampling parameters defined by the user (such as the maximum ensemble size). The centricity is estimated from the topological distance of the rotatable bond

DOI: 10.1021/acs.jcim.8b00704
J. Chem. Inf. Model. 2019, 59, 731−742

to the farthest atoms calculated on the molecular graph with the Floyd−Warshall algorithm.[33] Rotatable bonds located at the center of a molecule are assigned more alternative torsion angle values compared to rotatable bonds of terminal fragments. This is because fragments close to the center of a molecule are more likely to have a determinant effect on the overall conformation. More specifically, fragments located at the center of a molecule keep many if not all torsion angles recorded for a specific SMARTS pattern in the torsion angle library, whereas fragments located away from the center of the molecule are assigned only a few of the most frequently observed torsion angles. The overall aim of this procedure is the reduction of the number of conformers to be generated and analyzed during the clustering process (typically hundreds of thousands or even millions of conformations) by two to three orders of magnitude. The flexibility of a molecule is estimated based on the maximum number of possible conformations resulting from the enumeration of all torsion angle values stored in the library (without the consideration of potential clashes). The maximum number of generated candidate conformers for clustering is the product of the maximum allowed ensemble size (user-adaptable parameter; in this study 50 or 250) and a factor of 10 (Fast) or 20 (Best).

Once all torsion angles for conformer enumeration have been selected, the conformer generation process is initiated, starting from the most central fragment and following a standard incremental construction approach.[34] Initially, a depth-first search of the most likely torsion angles is carried out to ensure that the most relevant torsion angles are represented in the conformer ensemble and that the conformer generation produces the conformers which are likely most relevant. Provided that the number of conformers resulting from this depth-first search does not exceed the maximum number of candidate conformers for clustering, breadth-first search (starting again from the most central fragment) is carried out iteratively to explore all selected torsion angles and, hence, generate additional candidate conformers.

During conformer generation, topological symmetry classes of each heavy atom of the molecule are calculated in a canonical way using a variant of the CANON algorithm.[28] On the basis of these, local symmetries are detected and considered during torsion angle enumeration to avoid the generation of duplicate conformers. Because local symmetry detection depends on the used central fragment, not all symmetries can be detected and a final symmetry clustering via complete automorphism enumeration is performed to remove similar conformers due to global symmetries.

Conformations for rings formed by up to nine heavy atoms are calculated using conformations from a ring template library embedded in NAOMI[35] as described by Schärfer et al.[26] Ring systems are incrementally constructed from individual ring conformers. Following the concept of unique ring families (URFs) reported by Kolodzik et al.[36] (a recent reimplementation by Flachsenberg et al.[37] was used for Conformator), at most one relevant cycle (RC) per URF is selected for ring system conformation generation. Starting from the RC with the highest connectivity, the remaining cycles are attached while considering atom geometries according to VSEPR and taking into account the available stereo information. Within a tailored optimizer, simplified force field terms for bond distortion, angle bending, and torsion energy are used for evaluating the deviations of molecular geometries from the ideal values and for assessing steric clashes. The tailored optimizer subsequently relaxes the assembled ring system conformation.

This optimizer is also used to generate additional low-energy conformations based on initial template conformations to generate an ensemble of ring system conformations. Rings formed by more than nine atoms are handled by a new algorithm for sampling the conformations of macrocycles (see Conformer Generation for Macrocycles).

Conformations causing clashes are rejected as early as possible during the incremental construction process. Intramolecular clashes are defined as overlaps of more than 30% of the van der Waals radii of 1−4-connected (or more distant) heavy atom pairs that are not part of the same ring system. Alternatively, users can choose for Conformator to include hydrogen atoms in the clash calculation.

The configuration of any defined stereogenic center is preserved by the algorithm, whereas the configuration of any undefined *R*/*S*-stereogenic center is arbitrarily chosen once per molecule. Undefined *E*/*Z*-stereogenic centers are enumerated (limited only by steric hindrances and the maximum ensemble size). In the case of undefined stereogenic centers, the macrocycle conformation generation (see section Conformer Generation for Macrocycles) may produce a mix of stereoisomers (*R*/*S* and *E*/*Z*). Arbitrarily selecting one stereoisomer could prevent the algorithm from finding any reasonable result, especially in the case of *E*/*Z* isomers.

Clustering of Conformers. A new algorithm based on sphere exclusion clustering[38,39] was developed as part of Conformator for the efficient assembly of conformer ensembles (Algorithm S1, Figure S1). The clustering algorithm is the final step of the conformer ensemble generation. It aims to reduce the number of computationally expensive geometric comparisons of pairs of conformers required for the assembly of ensembles of a defined maximum size by exploiting the fact that sequentially generated conformers are likely to be highly similar to each other. To an outside observer, the list of conformers generated by Conformator will appear to be the result of a systematic search which explores valid torsion angles for one rotatable bond after the other. Geometric deviations between pairs of sequentially generated conformers are likely small because they often differ only by one torsion angle. Large deviations are less common and are often related to clashes which, when occurring during early stages of the search, can result in the rejection of whole branches of the search tree. The number of comparisons (RMSD calculations) between conformers is heavily reduced by traversing the list of conformers forward and the list of cluster centers backward. This increases the probability of similar conformers being compared early. When a similar enough conformer (defined by a RMSD threshold) is identified, the conformer is removed from the list of candidates and not compared to any further conformers.

During clustering, Conformator adjusts the minimum RMSD distance between conformers and determines an appropriate RMSD threshold for each individual molecule to generate ensembles that do not exceed the maximum ensemble size. This RMSD threshold depends on the maximum ensemble size and quality level as well as the size and flexibility of the molecule. The algorithm is heuristic but deterministic, i.e., it produces the same result given the same list of conformations (note that, unless the user requests that input coordinates be used as a starting point for conformer generation, the list of conformations generated during each run is identical for a given molecule).

Conformator does not rank conformers explicitly (although the first conformers generated by the algorithm are more likely based on the most commonly observed torsion angles). The conformers of an ensemble of small size (e.g., 5 conformers) will not necessarily be part of an ensemble of larger size (e.g., 50 conformers) because for small ensembles Conformator may prioritize conformers of high diversity over conformers with more commonly observed torsion angles. It is also unlikely that the first few conformers of an ensemble of larger size are those that would be included in an ensemble of small size. For this reason, to obtain ensembles of desired size, users are advised to not extract individual conformers but to define an adequate maximum ensemble size prior to ensemble generation.

The clustering algorithm (illustrated in Figure S1 and reported as pseudo code in Algorithm S1) involves the following key steps (with *radius* and *increase* having the values 0.1 and 0.05 Å for Best and 0.5 and 0.5 Å for Fast):

1. An empty list of cluster centers is created.
2. The first conformation becomes the first cluster center.
3. Each conformer in the list of conformers is compared to the reversed list of cluster centers.
4. If the conformer is (a) similar to an existing cluster center (RMSD smaller than *radius*), then the conformer is immediately discarded, or (b) dissimilar to any of the existing cluster centers, then the conformer is added to the list of cluster centers.
5. If the number of cluster centers reaches the maximum ensemble size, *radius* is increased as specified by the *increase* parameter, and the clustering process is restarted with an empty list of cluster centers and the list of remaining conformers.
6. When all conformers are assigned to a cluster center and the ensemble size is equal to or below the maximum ensemble size, the list of cluster centers is reported as the conformer ensemble.

Conformer Generation for Macrocycles. Conformers for macrocyclic ring systems are generated using a novel algorithm. First, all macrocycles are sliced by cutting bonds until no macrocycles are left. Next, conformations are generated for these structures without macrocycles, which serve as starting points for the rebuilding of the macrocycles by a local optimization algorithm. The following sections describe these processes in detail. Schematics of the conformer generation algorithm for macrocycles are provided in Figure S2.

Preprocessing of Macrocyclic Structures for Conformer Generation. In the following, all rings formed by more than nine atoms are termed *macrocycle*; all others are termed *small rings*. This distinction is necessary because conformations for small rings are covered by the ring template library (see Conformer Generation Algorithm). The concept of unique ring families (URFs)[36,37] is used to consider one ring family at a time instead of processing individual rings. URFs are a unique, chemically meaningful, and polynomial description of the rings in a molecule.

First, all URFs of the molecule are identified.[36,37] A URF is called macrocyclic if it contains at least one ring with more than nine atoms. All ring systems are processed independently. All *macrocyclic* URFs in a ring system are iteratively cut at one single bond outside of *small rings* until the resulting ring system no longer contains any *macrocycles*. In case a molecule contains exactly one *macrocycle*, this process results in the cutting of one bond. By choosing exactly one bond to be cut during each iteration, the molecule remains connected. The single bond to be cut is chosen by prioritizing carbon−carbon and then carbon-incident bonds. If no such bond exists, the same priority rule is applied to bonds in conjugated systems. Bonds that are not adjacent to small rings are favored in the selection process. Double bonds, triple bonds, and bonds that are part of small rings are not cut. Macrocycles consisting entirely of small rings are incrementally constructed from individual ring conformers. Following the cutting of a bond, new single bonds equal in length to the original bond are introduced by attaching two dummy atoms.

Generation of Conformers for Preprocessed Macrocyclic Structures. Diverse conformations of the preprocessed macrocyclic structures are generated with Conformator's standard algorithm following the exact same procedure as described above (see Conformer Generation Algorithm; Figure 1).

Rebuilding the Macrocycles by Numerical Optimization. The conformations generated during the previous process are used as starting points for a gradient-based numerical optimization procedure that aims to reconstitute macrocycles by superimposing the dummy atoms with the atoms they replaced during the cutting step. Note that the initial conformations already have valid geometries at this point, obviously with the exception of the part where the macrocyclic bond is to be reintroduced. The optimization is performed employing internal coordinates, namely the torsion angles and bond angles in the macrocycles. By this strategy, the number of parameters is reduced down to at most one bond angle per atom and one torsion angle per bond.

Local optimization is performed using a reimplementation of the BFGS-B algorithm,[40,41] which was modified to not allow any atoms to move by more than 0.5 Å per iteration. This modification, inspired by recent work on the refinement of the positions of water molecules in protein crystal structures,[42] was made to increase the locality of the optimization method and avoid unreasonably large changes in geometry. The local optimization is performed only on the atoms of the macrocycle (all other atoms of the molecule are not considered), and no part of the macrocycle is fixed (except for individual atoms in small rings, which are moved as a unit).

The here introduced macrocyclic optimization score (MCOS, see eq 1) is used to reconstruct the macrocycle. It includes several well-known components from common force fields and some components specific to the optimization of macrocycles. The formulas of the terms in eq 1 are provided in the Figures S3−S9; the weights were determined empirically and are provided in Table S1. Please note that the MCOS and the individual score contributions are dimensionless and are not genuine energy terms.

$$\begin{aligned}\text{MCOS} = {} & w_{\text{overlay}}S_{\text{overlay}} + w_{\text{bond}}S_{\text{bond}} + w_{\text{angle}}S_{\text{angle}} + w_{\text{limit}}S_{\text{limit}} \\ & + w_{\text{torsion}}S_{\text{torsion}} + w_{\text{torsion,conjugated}}S_{\text{torsion,conjugated}} \\ & + w_{\text{clash}}S_{\text{clash}} \qquad (1)\end{aligned}$$

The overlay score given in eq 2 is the central part of the scoring function.

$$\begin{aligned}S_{\text{overlay}} = {} & \sum_{\{i,j\}\in \text{cut bonds}} \frac{1}{2}(\text{distance}(i,\ \text{dummy}(i))^2 \\ & + \text{distance}(j,\ \text{dummy}(j))^2) \qquad (2)\end{aligned}$$

where $\{i,j\}$ is a cut bond and dummy(j) is the dummy atom replacing atom j as a terminal atom adjacent to atom i.

S_{overlay} scores the distance between the dummy atoms and the atoms in the original macrocycle they replaced. Ideally, this distance should be close to 0 (see Figure S3). The overlay score ensures that the bond angle and bond length across the cut bond will be restored during local optimization. It also supports the preservation of local stereochemistry.

The bond angle term S_{angle} uses a harmonic potential (calculated on the angle cosine, see Figure S4) to account for deviations from the ideal values (see Conformer Generation Algorithm and ref 26). It is calculated only for bond angles directly altered during optimization (i.e., angles involving bonds along the macrocycle that are optimization parameters) and the angles involving the cut bonds. During local optimization, bond angles are box-constrained such that no bond angle may be set to values greater than 179 degrees (if the atom does not have linear VSEPR geometry) and smaller than 0 degrees. This is to prevent unreasonable bond angle changes or even inversions of the local stereochemistry as bond angles usually stay rather close to the respective ideal values. The bond angle constraints are further supported by the penalty S_{limit} in the scoring function for bond angles in macrocycles, which leads to a preference of bond angles between 30 and 150 degrees (see Figure S5). Both terms S_{angle} and S_{limit} are multiplied by a function (see Figure S7) that reduces the scores to 0 in cases where any bond length adjacent to the angle approaches 0 Å. This is necessary because bond angles are not defined in cases where two defining atoms are placed on top of each other.

In addition, the bond length term S_{bond} uses a harmonic potential (see Figure S6) to account for deviations from ideal values (see Conformer Generation Algorithm and ref 26. Only the bond lengths of the cut bonds are scored.

The torsion angle score for bonds within ($S_{\text{torsion,conjugated}}$) and outside ($S_{\text{torsion}}$) of conjugated systems is calculated using the same torsion angle potential but different weights. The (continuous) torsion angle potential is based solely on torsion angle peaks recorded in a freely available torsion angle library derived from the CSD.[27] It uses the von Mises function as the kernel for curve approximation[43] with a tailored equation for kappa. We estimate the curve width through connecting the second peak tolerance and the peak score from the torsion library with the measure of concentration of the von Mises function (kappa). Due to the numerical optimization steps in continuous torsion space, torsional angles may differ from the angles stored in the torsion library (note that the angles start from those stored in the torsion angle library).

The torsion angle potential is multiplied by a function (see Figure S8) that reduces the torsion angle score to 0 in cases where any bond angle along that torsion bond is either close to 0 or 180 deg (such bond angle values may be observed for cut bonds where the bond angle is not directly modified and therefore not subject to the box constraints). This is necessary because the torsion angle, as a function of the four atom coordinates, has a discontinuity when three consecutive atoms are collinear. The torsion angle potential is furthermore multiplied by the same function described above for S_{angle} and S_{limit} that reduces the score to 0 in cases where bond lengths are close to 0 Å (Figure S7).

To prevent intramolecular clashes, the clash term S_{clash} was added to the MCOS. S_{clash} is a quadratic function that penalizes van der Waals overlaps between 1 and 4-connected (or further away) heavy atoms that exceed the threshold level of 30% (see Figure S9).

Postprocessing and Filtering of Macrocyclic Structures for Conformer Generation. Following the optimization procedure, the cut bonds are reintroduced to close the macrocycle conformations again, and the dummy atoms are removed. In the rare event that the resulting macrocycle has assigned a configuration that does not correspond to the conformation of the input structure, the conformer is rejected. The geometry of all atoms forming macrocycles is then checked and, if required, optimized to resemble VSEPR geometries by adjusting the position of the macrocycle substituents.

All *macrocycle* conformations are then checked for bond lengths and angles that deviate strongly from the known optimal value.[26] The optimal values for bond length and bond angles were the same as used for the optimization; for allowed deviations see ref 44. Furthermore, the planarity of conjugated macrocycles (e.g., protoporphyrin IX, *PP9*) is tested by checking their bonds for torsion angles deviating from 0 or 180°. Because macrocycles can adopt highly strained conformations a maximum deviation of 20 degrees of torsion angles in conjugated macrocycles is allowed. Only in cases where no (approximately) planar conjugated system can be generated, nonplanar alternative conformations are considered.

Before utilizing the macrocycle conformations for ensemble generation, the conformations are sorted by their final MCOS and subjected to one iteration of clustering utilizing the identical clustering algorithm (see Clustering of Conformers) with an RMSD threshold of 0.1 Å. The sorting step prior to the clustering step ensures that for each cluster the best-scored conformation is selected.

Output Summary. In addition to any warnings and errors, Conformator prints out a single-line summary for each processed molecule. The summary includes information on the name of the molecule, the number of generated conformers, and stereochemistry. The user may request additional output, such as the minimum pairwise RMSD between a generated conformer and the input conformer, and the minimum pairwise RMSD between any generated conformers. Note that these options may lead to substantially longer runtimes.

Benchmarking Conformer Ensemble Generators. *Preparation of the Benchmark Dataset for Computation.* The Platinum Diverse Dataset used for benchmarking conformer ensemble generators is a representative subset of the Platinum Dataset.[45] Both datasets were compiled according to the method described in ref 11, with the improvements described in ref 12 and downloaded from ref 46.

Conformer Ensemble Generation. In our previous benchmark studies, standard 3D structures (SDF format) generated from SMILES with NAOMI served as input for conformer ensemble generation for the RDKit DG algorithm and OMEGA. The same structures were used as input for CONFECT[26] in the present work. Conformator was benchmarked with both SMILES and 3D structures as input. Conformer ensembles were calculated with the parameters described in the Results section and summarized in Table 1.

RMSD Calculations, Geometry Checks, and Runtime Measurements. The RMSD between pairs of conformers was calculated with NAOMI.[35] NAOMI determines the RMSD based on the best superposition of a pair of conformers, taking into account molecular symmetry via complete automorphism enumeration.

Table 1. Parameter Sets Applied to Conformer Ensemble Generation

algorithm	mode[a]	clustering[b]	force field
Conformator	Best (default)	RMSD	n/MCOS[c]
Conformator	Fast	RMSD	n/MCOS[c]
CONFECT	3[d]	TFD[e]	TrAmber[f]
RDKit DG[g]	n/a	RMSD	UFF[47]
OMEGA[g]	default	RMSD	mmff94s_NoEstat[h]

[a] Parameter sets and search modes supplied by the developers of the respective algorithms. [b] Distance measure for clustering conformers to form ensembles. Default values were applied. [c] Macrocycle Optimization Score (MCOS). Only used for macrocycle optimization. [d] Setting recommended by the developers.[48] [e] Torsion fingerprint distance.[49] [f] TrAmber is a hybrid force field partly based on TAFF[50] and used for resolving clashes by small rotations of torsion angles. [g] Best-performing parameter set in our previous study.[12] [h] MMFF94 variant that includes all MMFF94s terms except those for Coulomb interactions.

NAOMI was also utilized to determine the deviation of atom angles and bond lengths from known optimal values as well as the divergence of aromatic rings and ring systems (up to six bonds per relevant cycle) from planarity.[44] Runtimes of conformer ensemble generation were measured for SD files containing single molecules.

Statistical Analysis. The Mann−Whitney U test was used to test for statistical significance at $\alpha = 0.05$ and $\alpha = 0.01$, with the Holm−Bonferroni method[51] applied to control the familywise error rate. The p-values are reported for pairwise comparisons of the conformer ensemble generators at maximum ensemble sizes 250 and 50 in the Supporting Information (Tables S2 and S3).

Hardware Setup. All calculations were performed single-threaded on Linux workstations running openSUSE 42.2 and equipped with Intel Xeon processors (2.2−2.7 GHz) and 126 GB of main memory (Conformator typically uses less than 1 GB of memory).

■ RESULTS

Benchmarking Conformator. The accuracy and efficiency of Conformator in representing protein-bound ligand conformations was assessed using the same dataset[45] and following the same testing procedure[12] previously applied to the benchmarking of the commercial algorithms ConfGen,[19] ConfGenX,[20] cxcalc,[21] iCon,[22] MOE,[23] and OMEGA.[24] In a second, earlier published study,[11] we compared the performance of the free conformer ensemble generators Balloon (two different algorithms),[18] the RDKIT DG[13] and ETKDG[14] algorithms, Confab,[15] Frog2,[16] and Multiconf-DOCK.[17] This study also followed the identical testing protocol but utilized an earlier version of the Platinum Diverse Dataset.[52] We have previously shown[12] that the marginal differences in the composition of both versions of the Platinum Dataset have no significant impact on any study outcomes. This means that all results presented in the current work can be directly compared to the results reported in either of our previous studies.

The following sections report on key performance figures computed for Conformator and CONFECT, some of which are summarized in Figure 2 and Table 2. In support of the discussions, results obtained as part of our previous study with the best-performing parameter sets (Table 1) for the RDKit DG algorithm (the best-performing free algorithm) and OMEGA (the best-performing commercial algorithm) are recited in the figures and tables of the current work. Results of the Mann−Whitney U test for statistical significance for maximum ensemble sizes of 250 and 50 are provided in the Supporting Information (Tables S2 and S3). In the following sections, four-letter codes refer to PDB entries, and three-letter codes in italics refer to PDB ligand identifiers.

Accuracy and Ensemble Size. This study, like most benchmark studies (including ours[11,12]), defines the accuracy of conformer ensemble generators by the minimum RMSD in Å measured between the experimentally determined protein-bound conformation and any conformer of the computed ensemble. Accuracy is, to some extent, a function of ensemble

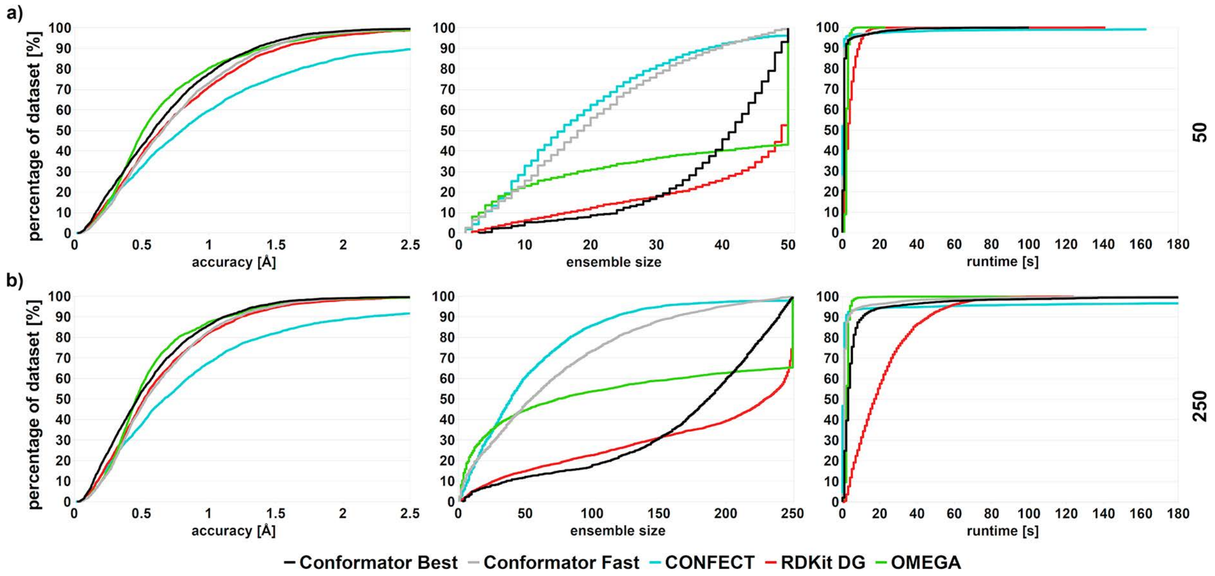


Figure 2. Percentage of protein-bound ligand conformations of the Platinum Diverse Dataset reproduced by the different algorithms within a certain accuracy (left), ensemble size (middle), and runtime per molecule (right) at maximum ensemble sizes (a) 50 and (b) 250 conformers. Steeper curves indicate better performance with respect to all three criteria.

Table 2. Comparison of the Performance of Conformer Ensemble Generators on the Platinum Diverse Dataset[a]

	maximum ensemble size 50		maximum ensemble size 250	
algorithm	mean	median	mean	median
	RMSD (Å)			
Conformator Best	**0.68**	**0.58**	**0.57**	**0.47**
Conformator Fast	0.75	0.66	0.64	0.53
CONFECT	0.92	0.74	0.78	0.67
RDKit DG	0.82	0.64	0.64	0.52
OMEGA	**0.67**	**0.51**	**0.57**	**0.46**
	ensemble size			
Conformator Best	38	42	166	187
Conformator Fast	20	19	70	54
CONFECT	**18**	**15**	**50**	**38**
RDKit DG	42	49	180	229
OMEGA	34	50	118	74
	runtime (s)			
Conformator Best	**2**	**1**	7	3
Conformator Fast	**2**	**1**	**3**	**1**
CONFECT	**2**	**1**	4	**1**
RDKit DG	4	3	18	14
OMEGA	**2**	2	**3**	2

[a] The best values obtained for RMSD (considering statistical significance), ensemble size, and runtime by any of the tested algorithms are marked in bold.

size.[53] This is because ensembles are generally designed to consist of diverse conformers, which means that chances for one of these conformers to closely resemble the experimentally observed conformation generally increase with the number of generated conformers. Unless stated otherwise, all results presented in the following sections refer to ensembles with a maximum of 250 conformers.

Conformator Best represented the protein-bound ligand conformations with a median RMSD of 0.47 Å at a median ensemble size of 187. Its accuracy was significantly better than that of the RDKit DG algorithm (median RMSD 0.52 Å), even though the RDKit DG algorithm produces larger ensembles (median 229 conformers). The accuracy of Conformator Best was also competitive with that of OMEGA (RMSD 0.47 vs 0.46 Å; difference not statistically significant) at, however, the expense of a substantially larger median ensemble size (187 vs 74 conformers). Run at a maximum ensemble size of 250, Conformator Best tends to produce larger ensembles than OMEGA for molecules with four or fewer rotatable bonds (Figure 3a). The opposite trend is observed for more flexible molecules, for which OMEGA generally produces more conformers than Conformator Best. Whereas only 0.8% of all ensembles generated with Conformator Best consisted of the maximum allowed number of conformers (i.e., 250), this figure was 34% for OMEGA. The R^2 for the correlation between the number of rotatable bonds and the size of conformer ensembles was 0.27 for Conformator Best. This weak correlation is a result of the rules for sampling torsion angles for rotatable bonds and of the clustering algorithm, both of which bias the ensembles toward more diversity, meaning that even if for a rotatable bond multiple preferred torsion angles are known, few representative torsion angles are utilized to comply with the maximum allowed ensemble size.

For a maximum ensemble size of 50 conformers, Conformator Best produced smaller ensembles (median 42 conformers) than OMEGA (median 50 conformers) and the RDKit DG algorithm (median 49 conformers). In this setup, no statistically significant difference in the accuracy of Conformator Best (median 0.58 Å) and OMEGA (median 0.51 Å) was observed (Table S3). Again, the accuracy of Conformator Best was significantly higher than that of RDKit DG (median 0.64 Å). At a maximum ensemble size of 50 conformers, Conformator Best generated larger ensembles than OMEGA for molecules with less than four rotatable bonds but smaller-sized ensembles for molecules with more than four rotatable bonds (Figure 3b). Only 7% of all conformers generated with Conformator Best, but 56% of all conformers generated with OMEGA had the maximum ensemble size of 50 conformers (Figure 2a).

At a maximum ensemble size of 250 conformers, Conformator Fast reproduced the experimentally observed conformations with equal accuracy as the RDKit DG algorithm (median RMSD 0.53 vs 0.52 Å; difference not statistically significant), despite much smaller ensembles (median 54 vs 229 conformers). CONFECT produced the smallest ensembles but also was the least accurate among all tested algorithms (median 38 conformers per ensemble; median RMSD 0.67 Å).

In addition, we quantified the accuracy of conformer ensemble generators as the percentage of experimentally observed conformations represented below RMSD thresholds of 0.5, 1.0, 1.5, and 2.0 Å (Table 3). In this assessment, Conformator Best and OMEGA showed comparable performance, with 53 and 56% of all experimental conformations represented with an RMSD below 0.5 Å, and 97 and 96% represented with an RMSD below 1.5 Å, respectively (maximum ensemble size 250 conformers). The success rates

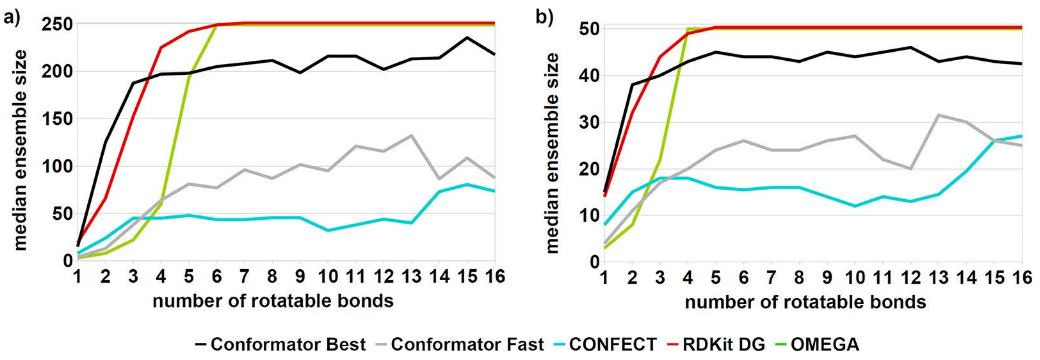


Figure 3. Median ensemble size vs number of rotatable bonds for ensembles of a maximum of (a) 250 and (b) 50 conformers. Lower curves indicate better performance with respect to ensemble size.

Table 3. Percentage of Structures of the Platinum Diverse Dataset Successfully Reproduced within a Specified RMSD Threshold[a]

algorithm	maximum ensemble size 50				maximum ensemble size 250			
	RMSD threshold (Å)							
	0.5	1.0	1.5	2.0	0.5	1.0	1.5	2.0
Conformator Best	42	78	**94**	**98**	53	86	**97**	**99**
Conformator Fast	37	73	91	**98**	46	83	95	**99**
CONFECT	32	60	76	85	37	62	82	88
RDKit DG	38	71	89	96	47	82	95	98
OMEGA	**49**	**80**	92	97	**56**	**87**	96	**99**

[a] The values of the best-performing algorithms per column are marked in bold.

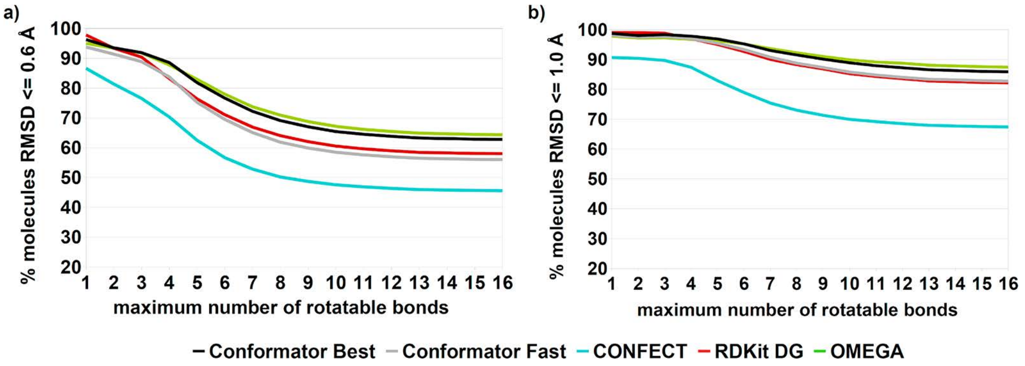


Figure 4. Percentage of molecules of the Platinum Diverse Dataset reproduced by the tested algorithms with a maximum RMSD of (a) 0.6 and (b) 1.0 Å as a function of the maximum number of rotatable bonds. The maximum ensemble size was set to 250 conformers.

of Conformator Fast were comparable with those of the RDKit DG algorithm. For ensembles of a maximum of 50 conformers at an RMSD threshold below 0.5 Å, the success rate of OMEGA was higher than that of Conformator Best (49 vs 42%) and any other tested algorithm.

As a third way of assessing the accuracy of conformer ensemble generators, we quantified the percentage of molecules represented with an RMSD below 0.6 (the maximum positional uncertainty for atoms in the Platinum Dataset)[11] and below 1.0 Å (below which docking poses are commonly deemed sufficiently accurate) with respect to the complexity of their conformational space, represented (in part) by the number of rotatable bonds (Figure 4). At both RMSD thresholds (maximum ensemble size 250 conformers), Conformator Best performed comparably to OMEGA and Conformator Fast comparably to the RDKit DG algorithm. Both Conformator Best and OMEGA, however, performed substantially better than Conformator Fast, the RDKit DG algorithm, and CONFECT at both RMSD thresholds. The success rates of representing experimental structures below an RMSD of 0.6 Å were 63−96% for Conformator Best, 64−95% for OMEGA, and 58−98% for the RDKit DG algorithm. Likewise, the success rates of representing experimental structures below an RMSD of 1.0 Å were 86−99% for Conformator Best, 87−98% for OMEGA, and 82−99% for the RDKit DG algorithm.

Among all tested algorithms, the accuracy of ensembles generated with OMEGA was least dependent on the number of rotatable bonds. At an RMSD cutoff of 0.6 Å, OMEGA successfully represented 88% of all molecules with up to four rotatable bonds and 71% of all molecules with up to eight rotatable bonds. These figures were 89% and 69% for Conformator Best, respectively.

The diversity of the ensembles generated with Conformator strongly depends on the specific molecular structure in question. In general, the diversity of ensembles increases with the number of rotatable bonds. The R^2 for the correlation between the median pairwise RMSD of all conformers and the number of rotatable bonds was 0.60 (default settings; Figure S10). Two outliers were observed, which are the highly symmetrical ligands *B3P* (Figure S10A) and *5MY* (Figure S10B), for which the symmetry-corrected RMSD was lower than expected based on the number of rotatable bonds. The R^2 for the correlation between the minimum pairwise RMSD and the number of rotatable bonds was 0.50 (default settings; Figure S11). Note that the RMSD also depends on the size of the molecule and that the clustering threshold is not adjusted if the initially generated conformer ensemble is smaller than the maximum allowed ensemble size. Also, during each round of clustering, the radius is incrementally increased by a defined value (i.e., 0.1 Å for Fast and 0.05 Å for Best), for which reason the maximum allowed ensemble size is often not reached.

For a subset of 987 molecules of the Platinum Diverse Dataset (all of them have a maximum of six rotatable bonds), we were able to generate complete conformer ensembles without clustering and without a set maximum ensemble size (maximum allowed runtime of 72 h per molecule; Table S4). For 92% of all molecules in this subset (84% with default settings), the complete ensembles included a conformer with an RMSD lower than 0.5 Å and for 99% (98% with default settings) a conformer with an RMSD lower than 1 Å. Use of complete conformer ensembles instead of the (default) ensembles of a maximum size of 250 improved the RMSD by 0.5 Å or more in only 14 out of 987 cases. The maximum ensemble size measured was 185 112 conformers; the mean ensemble size 12 024. These results demonstrate the efficiency of the clustering procedure implemented in Conformator.

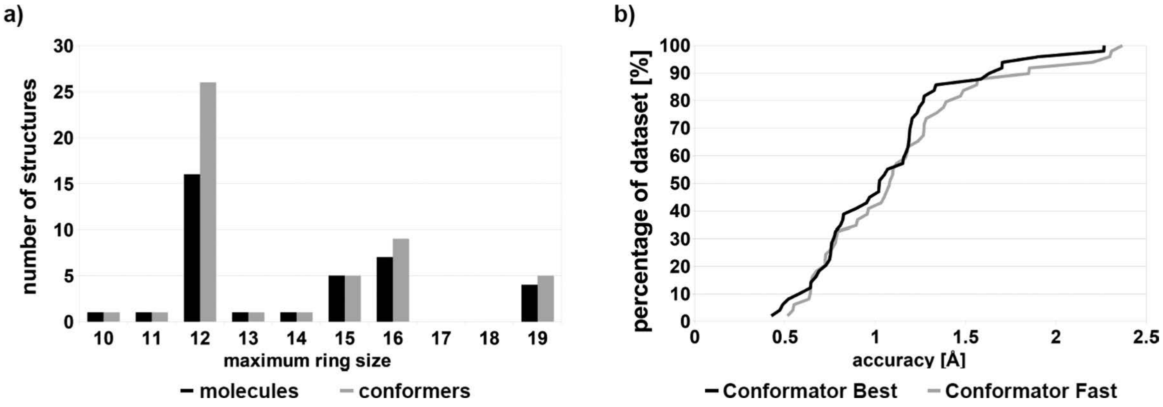


Figure 5. Sperrylite Dataset contains 49 protein-bound structures of compounds including at least one macrocycle formed by ten or more atoms. (a) Distribution of the maximum ring sizes (number of atoms in a ring) of these macrocycles and their conformations. (b) Cumulative percentage of these structures reproduced by Conformator below a defined maximum RMSD threshold (maximum ensemble size 250 conformers).

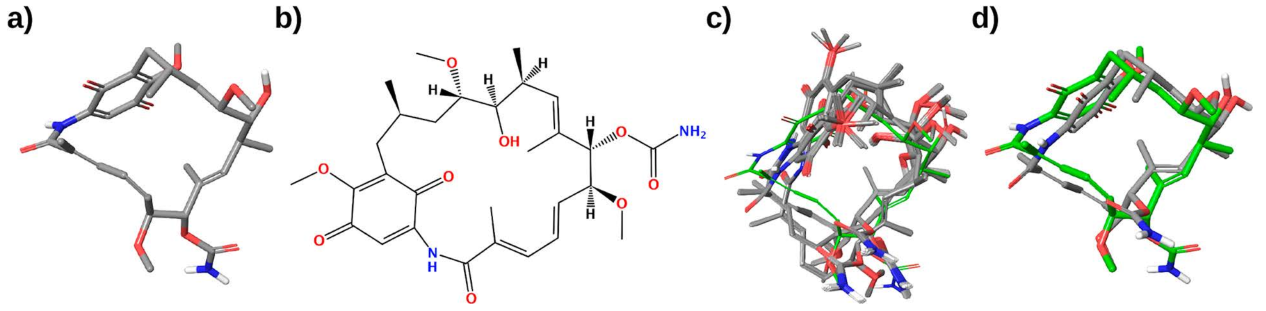


Figure 6. Visualization of structures of geldanamycin. (a) The conformer from the Sperrylite Dataset (*GDM* in 3C11; input for the validation of Conformator), (b) 2D representation of geldanamycin, (c) an ensemble of conformers generated by Conformator Best and superposed with original conformer (green carbon atoms), and (d) the closest conformer generated with Conformator Best and superposed with the original conformer (green carbon atoms).

Success Rates in Processing Molecules. With the exception of CONFECT (success rate 93.4%), all ensemble generators successfully produced ensembles for more than 99% of all tested molecules (Conformator Best and Fast 100.0%; OMEGA 99.6%; RDKit DG algorithm 99.9%). Conformator and OMEGA are designed to handle both 2D and 3D input and produce identical results with either type of information. In the case of SMILES input, Conformator was able to successfully process all molecules with the exception of three molecules with small, bridged rings (i.e., *HUX*, *SAW*, *TSA*). If valid input coordinates are given and the option to generate new 3D coordinates is not set, these three molecules can also be successfully processed by Conformator.

Runtimes. For ensembles consisting of a maximum of 250 conformers, the median runtimes for Conformator Fast and Best were 1 and 3 s, respectively (for individual molecules, repeated runtime measurements differed by less than 5%). Hence Conformator was much faster than the RDKit DG algorithm (median 14 s) and approximately as fast as OMEGA (median 2 s). For ensembles consisting of a maximum of 50 conformers, no substantial differences in the median runtimes were observed: calculations with Conformator Fast and Best had a median runtime of 1 s, with OMEGA 2 s and with the RDKit DG algorithm 3 s. Note that in previous tests,[11] the RDKit ETKDG and DG algorithms produced conformers of comparable quality, with the ETKDG algorithm being 25% faster.

Case Studies on the Reproduction of Experimentally Observed Conformations of Macrocycles. In recent years, macrocycles have emerged as one of the most promising categories of drug candidates for multiple indications.[54−57] Macrocyclic systems are restricted in their rotational and conformational freedom. While this property is actively exploited in the design of highly effective and specific compounds, the interdependency of rotatable bonds and other features such as bridged rings pose significant challenges to conformer ensemble generation. New conformer ensemble generators and extensions, in particular to commercial algorithms, have recently been reported to specifically address these issues.[58−65]

The dedicated algorithm for macrocycle conformer generation, which is part of Conformator, cuts all macrocycles and generates conformers for these open ring structures with Conformator's standard algorithm. In contrast to DG approaches (which usually start from random coordinates), the conformers used as a starting point for cyclization are already geometrically valid.

We tested the ability of Conformator to represent the experimentally observed, protein-bound conformations of macrocyclic compounds. For this purpose, we extracted from the Sperrylite Dataset all 49 structures of compounds including at least one ring formed by 10 or more atoms (29 of these structures are also part of the Platinum Diverse Dataset). Seven of the molecules included in this dataset are represented by more than one experimental structure: latrunculin A (*LAR*; 6 conformers), 6-deoxyerythronolide B (*DEB*; 4 conformers), and geldanamycin (*GDM*), *LAB*, *LY4*, *PP9*, and *S1A* (2 conformers). The dataset contains rings of eight different sizes

(Figure 5a). It is dominated by 16 molecules (26 conformers) with rings consisting of 12 atoms and 7 molecules (nine conformers) with rings consisting of 16 atoms.

Conformator Best successfully processed all 49 macrocyclic structures and obtained a median RMSD of 1.0 Å (Figure 5b). The maximum RMSD measured was 2.3 Å for both structures of geldanamycin (PDB complexes 3C11 and 4XDM; Figure 6). Geldanamycin is a particularly challenging molecule. It consists of 40 heavy atoms and a macrocycle formed by 19 atoms. Its conformation is strongly bent and includes several torsion angles that according to Conformator's torsion angle library are unlikely.

All further (47) macrocyclic structures were reproduced with RMSD values of less than 2.0 Å. Conformator Best reproduced the experimentally observed conformation of macbecin (*BC2*; 2VWC) and valerjesomycin (*VJ6*; 4JQL), both including macrocycles formed by 19 atoms, with RMSDs of 1.9 and 0.8 Å, respectively. For 27 macrocyclic structures (55%), Conformator Best generated at least one conformer with an RMSD not higher than 1.0 Å. At a maximum ensemble size of 250 conformers, the median size of ensembles generated with Conformator Best for the 49 macrocycles was 197 conformers and the average runtime was 104 s (median 88 s) per molecule. Given the limited amount of high-quality structural data on protein-bound macrocycles available to date, no statistically sound conclusions can be drawn on which of the two algorithms performs better.

Comparison of Conformator's Clustering Algorithm with k-Medoids Clustering. To assess the performance of the new clustering algorithm implemented in Conformator, we produced a version of Conformator Best with the new clustering algorithm replaced by the k-medoids clustering algorithm (the partitioning around medoids method).[66,67] With a maximum of 25 iterations, Conformator in combination with the k-medoids clustering algorithm reached median and mean accuracy values identical to those of the original version of Conformator (median RMSD 0.47 Å; mean RMSD 0.57 Å). However, the median and mean runtimes were substantially longer for the k-medoids clustering algorithm variant (14 and 272 s per molecule, respectively) as compared to the original version of Conformator (median 3 s; mean 7 s per molecule, respectively). The longest runtime observed for the k-medoids clustering variant was 12.1 h as compared to 512 s for the original version of Conformator. The ensembles generated by the k-medoids clustering variant had a median ensemble size of 250 conformers (mean ensemble size 205) as compared to 187 conformers (mean ensemble size 166) for the original version of Conformator. With k-medoids clustering, 58% of all generated ensembles were of the maximum allowed size (250), whereas this was the case for only 7% of all ensembles generated with the original version of Conformator. The high percentage of large ensembles generated by the k-medoids clustering variant is not surprising because reaching the maximum ensemble size is a defined objective of this clustering algorithm.

■ CONCLUSION

Conformator is an efficient knowledge-based algorithm for the generation of conformer ensembles of small molecules. One of the key features of Conformator is its new clustering algorithm for the compilation of representative conformer ensembles that exploits the partial presorting of consecutively generated conformers. Conformer ensembles generated with Conformator are independent of input geometries and formats because the input coordinates are not considered; the new cluster algorithm introduced here is deterministic, and the atom order of the molecule is canonized prior to conformer generation. Furthermore, we present a novel algorithm for the generation of conformations for macrocyclic ring systems. The algorithm is robust, widely applicable, and makes use of the sophisticated technology for acyclic conformer generation. A novel numeric optimizer working hand in hand with a differentiable scoring function MCOS is responsible for low-energy conformations even in complex, macrocyclic ring systems.

Conformator reaches a level of accuracy and efficiency that is comparable to that of OMEGA. The new algorithm performs particularly well with molecules composed of five or more rotatable bonds, for which it reaches competitive performance while keeping ensemble sizes low. OMEGA, on the other hand, is still ahead in sampling molecules with fewer than five rotatable bonds (which account for more than half of all molecules of the benchmarking dataset), for which it obtains the best accuracy among all tested algorithms even with small ensembles. Preference for either algorithm will depend on the specific application, such as the composition and size of the molecular libraries to be processed. From the outcomes of this study, however, it is clear that in direct comparison with other free algorithms, Conformator obtains very good performance and is the only algorithm for which no significant geometric errors were detected in any of the generated conformations. Conformator successfully processes 99.9% of all input structures, is capable of handling different types of 2D and 3D input, and requires only moderate computing resources. In contrast to many other approaches, Conformator does not use any PDB data for deriving geometric parameters like bond lengths, bond angles, torsion angles, or ring conformations. Therefore, the performance measured on the basis of the Platinum Dataset gives a realistic picture of the algorithm's practical performance.

Software availability: Conformator is free for noncommercial use and academic research. It is part of the software tool UNICON, a universal converter able to create 2D and 3D conformations on the fly. Conformator and UNICON are standalone command-line tools within the NAOMI ChemBio Suite[35] available from https://uhh.de/naomi.

■ ASSOCIATED CONTENT

Ⓢ Supporting Information

The Supporting Information is available free of charge on the ACS Publications website at DOI: 10.1021/acs.jcim.8b00704.

> Additional figures and tables: Results of the Mann−Whitney U tests and *p*-values adjusted with the Holm−Bonferroni method for ensembles with a maximum of 250 conformers; pseudo code for the cluster algorithm; visualization of Conformator's clustering algorithm by an example; empirically determined weights for the MCOS and functions of its individual score contributions (PDF)

■ AUTHOR INFORMATION

Corresponding Authors

*E-mail: rarey@zbh.uni-hamburg.de; Tel.: +49 40 42838-7351.

*E-mail: kirchmair@zbh.uni-hamburg.de; Tel.: +49 40 42838-7303.

ORCID

Nils-Ole Friedrich: 0000-0002-8983-388X
Florian Flachsenberg: 0000-0001-7051-8719
Agnes Meyder: 0000-0001-8519-5780
Kai Sommer: 0000-0003-1866-8247
Johannes Kirchmair: 0000-0003-2667-5877
Matthias Rarey: 0000-0002-9553-6531

Author Contributions

N.F., J.K., and M.R. conceived the work. N.F. developed the algorithmic concepts, implemented the software, and tested it. F.F. contributed to the development of the algorithmic concepts, implemented the macrocycle optimization, and contributed to the testing of Conformator. K.S. contributed to the implementation and testing of the algorithm and provided improvements. A.M. developed the tailored equation for kappa of the Mises function as kernel for curve approximation for the (continuous) torsion angle potential in the calculation of the torsion angle score. M.R. supervised the method development and J.K. the validation of Conformator. All authors contributed to writing of the manuscript and have given approval to the final version of the paper.

Funding

J.K. is supported by the Bergen Research Foundation (BFS) Grant BFS2017TMT01.

Notes

The authors declare the following competing financial interest(s): M.R. declares a potential financial interest in the event that the Conformator software is licensed for a fee to non-academic institutions in the future.

■ ACKNOWLEDGMENTS

The authors thank Christina de Bruyn Kops from the Center for Bioinformatics (ZBH) of the University of Hamburg for discussion and proofreading of the manuscript. Conformator is built on the NAOMI platform developed by many colleagues at BioSolveIT GmbH and the ZBH. The authors thank the whole team for their help in software development.

■ REFERENCES

(1) Güner, O.; Clement, O.; Kurogi, Y. Pharmacophore Modeling and Three Dimensional Database Searching for Drug Design Using Catalyst: Recent Advances. *Curr. Med. Chem.* **2004**, *11*, 2991−3005.

(2) Wolber, G.; Langer, T. LigandScout: 3-D Pharmacophores Derived from Protein-Bound Ligands and Their Use as Virtual Screening Filters. *J. Chem. Inf. Model.* **2005**, *45*, 160−169.

(3) Chen, I.-J.; Foloppe, N. Conformational Sampling of Druglike Molecules with MOE and Catalyst: Implications for Pharmacophore Modeling and Virtual Screening. *J. Chem. Inf. Model.* **2008**, *48*, 1773−1791.

(4) Dixon, S. L.; Smondyrev, A. M.; Knoll, E. H.; Rao, S. N.; Shaw, D. E.; Friesner, R. A. PHASE: A New Engine for Pharmacophore Perception, 3D QSAR Model Development, and 3D Database Screening: 1. Methodology and Preliminary Results. *J. Comput.-Aided Mol. Des.* **2006**, *20*, 647−671.

(5) Friedrich, N.-O.; Simsir, M.; Kirchmair, J. How Diverse Are the Protein-Bound Conformations of Small-Molecule Drugs and Cofactors? *Front. Chem.* **2018**, *6*, 68.

(6) Perola, E.; Charifson, P. S. Conformational Analysis of Drug-like Molecules Bound to Proteins: An Extensive Study of Ligand Reorganization upon Binding. *J. Med. Chem.* **2004**, *47*, 2499−2510.

(7) Iuzzolino, L.; Reilly, A. M.; McCabe, P.; Price, S. L. Use of Crystal Structure Informatics for Defining the Conformational Space Needed for Predicting Crystal Structures of Pharmaceutical Molecules. *J. Chem. Theory Comput.* **2017**, *13*, 5163−5171.

(8) Govind, N.; Petersen, M.; Fitzgerald, G.; King-Smith, D.; Andzelm, J. A Generalized Synchronous Transit Method for Transition State Location. *Comput. Mater. Sci.* **2003**, *28*, 250−258.

(9) Jacobson, L. D.; Bochevarov, A. D.; Watson, M. A.; Hughes, T. F.; Rinaldo, D.; Ehrlich, S.; Steinbrecher, T. B.; Vaitheeswaran, S.; Philipp, D. M.; Halls, M. D.; et al. Automated Transition State Search and Its Application to Diverse Types of Organic Reactions. *J. Chem. Theory Comput.* **2017**, *13*, 5780−5797.

(10) Hawkins, P. C. D. Conformation Generation: The State of the Art. *J. Chem. Inf. Model.* **2017**, *57*, 1747−1756.

(11) Friedrich, N.-O.; Meyder, A.; de Bruyn Kops, C.; Sommer, K.; Flachsenberg, F.; Rarey, M.; Kirchmair, J. High-Quality Dataset of Protein-Bound Ligand Conformations and Its Application to Benchmarking Conformer Ensemble Generators. *J. Chem. Inf. Model.* **2017**, *57*, 529−539.

(12) Friedrich, N.-O.; de Bruyn Kops, C.; Flachsenberg, F.; Sommer, K.; Rarey, M.; Kirchmair, J. Benchmarking Commercial Conformer Ensemble Generators. *J. Chem. Inf. Model.* **2017**, *57*, 2719−2728.

(13) *RDKit*; Open-Source Cheminformatics, Version 2017.03.2, 2017.

(14) Riniker, S.; Landrum, G. A. Better Informed Distance Geometry: Using What We Know To Improve Conformation Generation. *J. Chem. Inf. Model.* **2015**, *55*, 2562−2574.

(15) O'Boyle, N. M. *Confab*, Version 1.0.1, 2011.

(16) Miteva, M. A.; Guyon, F.; Tufféry, P. Frog2: Efficient 3D Conformation Ensemble Generator for Small Compounds. *Nucleic Acids Res.* **2010**, *38*, W622−W627.

(17) Sauton, N.; Lagorce, D.; Villoutreix, B. O.; Miteva, M. A. MS-DOCK: Accurate Multiple Conformation Generator and Rigid Docking Protocol for Multi-Step Virtual Ligand Screening. *BMC Bioinf.* **2008**, *9*, 184.

(18) Vainio, M. J.; Johnson, M. S. Generating Conformer Ensembles Using a Multiobjective Genetic Algorithm. *J. Chem. Inf. Model.* **2007**, *47*, 2462−2474.

(19) Watts, K. S.; Dalal, P.; Murphy, R. B.; Sherman, W.; Friesner, R. A.; Shelley, J. C. ConfGen: A Conformational Search Method for Efficient Generation of Bioactive Conformers. *J. Chem. Inf. Model.* **2010**, *50*, 534−546.

(20) *ConfGenX*, Version 2016-2, Part of the Schrödinger Small-Molecule Drug Discovery Suite; Schrödinger: New York, NY, 2016.

(21) *cxcalc*, Version 15.8.31.0, Part of the Discovery Toolkit; ChemAxon: Budapest, Hungary, 2015.

(22) Poli, G.; Seidel, T.; Langer, T. Conformational Sampling of Small Molecules With iCon: Performance Assessment in Comparison With OMEGA. *Front. Chem.* **2018**, *6*, 229.

(23) *Molecular Operating Environment (MOE)*, Version 2016.08; Chemical Computing Group: Montreal, QC, 2017.

(24) Hawkins, P. C. D.; Skillman, A. G.; Warren, G. L.; Ellingson, B. A.; Stahl, M. T. Conformer Generation with OMEGA: Algorithm and Validation Using High Quality Structures from the Protein Databank and Cambridge Structural Database. *J. Chem. Inf. Model.* **2010**, *50*, 572−584.

(25) Meyder, A.; Nittinger, E.; Lange, G.; Klein, R.; Rarey, M. Estimating Electron Density Support for Individual Atoms and Molecular Fragments in X-Ray Structures. *J. Chem. Inf. Model.* **2017**, *57*, 2437−2447.

(26) Schärfer, C.; Schulz-Gasch, T.; Hert, J.; Heinzerling, L.; Schulz, B.; Inhester, T.; Stahl, M.; Rarey, M. CONFECT: Conformations from an Expert Collection of Torsion Patterns. *ChemMedChem* **2013**, *8*, 1690−1700.

(27) Guba, W.; Meyder, A.; Rarey, M.; Hert, J. Torsion Library Reloaded: A New Version of Expert-Derived SMARTS Rules for Assessing Conformations of Small Molecules. *J. Chem. Inf. Model.* **2016**, *56*, 1−5.

(28) Weininger, D.; Weininger, A.; Weininger, J. L. SMILES. 2. Algorithm for Generation of Unique SMILES Notation. *J. Chem. Inf. Model.* **1989**, *29*, 97−101.

DOI: 10.1021/acs.jcim.8b00704

(29) Schärfer, C.; Schulz-Gasch, T.; Ehrlich, H.-C.; Guba, W.; Rarey, M.; Stahl, M. Torsion Angle Preferences in Druglike Chemical Space: A Comprehensive Guide. *J. Med. Chem.* **2013**, *56*, 2016−2028.

(30) Sutton, L. E. Interatomic Distances Supplment. *Soil Sci.* **1965**, *100*, 76.

(31) Gillespie, R. J. The Electron-Pair Repulsion Model for Molecular Geometry. *J. Chem. Educ.* **1970**, *47*, 18.

(32) Ehrlich, H.-C.; Henzler, A. M.; Rarey, M. Searching for Recursively Defined Generic Chemical Patterns in Nonenumerated Fragment Spaces. *J. Chem. Inf. Model.* **2013**, *53*, 1676−1688.

(33) Floyd, R. W. Algorithm 97: Shortest Path. *Commun. ACM* **1962**, *5*, 345.

(34) Rarey, M.; Kramer, B.; Lengauer, T.; Klebe, G. A Fast Flexible Docking Method Using an Incremental Construction Algorithm. *J. Mol. Biol.* **1996**, *261*, 470−489.

(35) Urbaczek, S.; Kolodzik, A.; Fischer, J. R.; Lippert, T.; Heuser, S.; Groth, I.; Schulz-Gasch, T.; Rarey, M. NAOMI: On the Almost Trivial Task of Reading Molecules from Different File Formats. *J. Chem. Inf. Model.* **2011**, *51*, 3199−3207.

(36) Kolodzik, A.; Urbaczek, S.; Rarey, M. Unique Ring Families: A Chemically Meaningful Description of Molecular Ring Topologies. *J. Chem. Inf. Model.* **2012**, *52*, 2013−2021.

(37) Flachsenberg, F.; Andresen, N.; Rarey, M. RingDecomposerLib: An Open-Source Implementation of Unique Ring Families and Other Cycle Bases. *J. Chem. Inf. Model.* **2017**, *57*, 122−126.

(38) Hudson, B. D.; Hyde, R. M.; Rahr, E.; Wood, J.; Osman, J. Parameter Based Methods for Compound Selection from Chemical Databases. *Quant. Struct.-Act. Relat.* **1996**, *15*, 285−289.

(39) Butina, D. Unsupervised Data Base Clustering Based on Daylight's Fingerprint and Tanimoto Similarity: A Fast and Automated Way To Cluster Small and Large Data Sets. *J. Chem. Inf. Comput. Sci.* **1999**, *39*, 747−750.

(40) Zhu, C.; Byrd, R. H.; Lu, P.; Nocedal, J. Algorithm 778: L-BFGS-B: Fortran Subroutines for Large-Scale Bound-Constrained Optimization. *ACM Trans. Math. Softw.* **1997**, *23*, 550−560.

(41) Morales, J. L.; Nocedal, J. Remark on "algorithm 778: L-BFGS-B: Fortran Subroutines for Large-Scale Bound Constrained Optimization. *ACM Trans. Math. Softw.* **2011**, *38*, 1−4.

(42) Nittinger, E.; Flachsenberg, F.; Bietz, S.; Lange, G.; Klein, R.; Rarey, M. Placement of Water Molecules in Protein Structures: From Large-Scale Evaluations to Single-Case Examples. *J. Chem. Inf. Model.* **2018**, *58*, 1625−1637.

(43) McCabe, P.; Korb, O.; Cole, J. Kernel Density Estimation Applied to Bond Length, Bond Angle, and Torsion Angle Distributions. *J. Chem. Inf. Model.* **2014**, *54*, 1284−1288.

(44) Meyder, A.; Kampen, S.; Sieg, J.; Fährrolfes, R.; Friedrich, N.-O.; Flachsenberg, F.; Rarey, M. StructureProfiler: An All-in-One Tool for 3D Protein Structure Profiling. *Bioinformatics* **2018**, DOI: 10.1093/bioinformatics/bty692.

(45) Platinum Diverse Dataset (version 2017_01). http://www.zbh.uni-Hamburg.de/platinum_dataset (accessed June 27, 2017).

(46) The Platinum Datasets. http://www.zbh.uni-hamburg.de/platinum_dataset (accessed June 27, 2017).

(47) Rappe, A. K.; Casewit, C. J.; Colwell, K. S.; Goddard, W. A.; Skiff, W. M. UFF, a Full Periodic Table Force Field for Molecular Mechanics and Molecular Dynamics Simulations. *J. Am. Chem. Soc.* **1992**, *114*, 10024−10035.

(48) Schärfer, C. Universität Hamburg, Hamburg, Germany, *Personal Communication*, May, 2014.

(49) Schulz-Gasch, T.; Schärfer, C.; Guba, W.; Rarey, M. TFD: Torsion Fingerprints as a New Measure to Compare Small Molecule Conformations. *J. Chem. Inf. Model.* **2012**, *52*, 1499−1512.

(50) Clark, M.; Cramer, R. D.; Van Opdenbosch, N. Validation of the General Purpose Tripos 5.2 Force Field. *J. Comput. Chem.* **1989**, *10*, 982−1012.

(51) Holm, S. A Simple Sequentially Rejective Multiple Test Procedure. *Scand. J. Stat.* **1979**, *6*, 65−70.

(52) Platinum Diverse Dataset (version 2016_01). http://www.zbh.uni-Hamburg.de/platinum_dataset (accessed June 27, 2017).

(53) Griewel, A.; Kayser, O.; Schlosser, J.; Rarey, M. Conformational Sampling for Large-Scale Virtual Screening: Accuracy versus Ensemble Size. *J. Chem. Inf. Model.* **2009**, *49*, 2303−2311.

(54) Marsault, E.; Peterson, M. L. Macrocycles Are Great Cycles: Applications, Opportunities, and Challenges of Synthetic Macrocycles in Drug Discovery. *J. Med. Chem.* **2011**, *54*, 1961−2004.

(55) Krahn, D.; Ottmann, C.; Kaiser, M. Macrocyclic Proteasome Inhibitors. *Curr. Med. Chem.* **2011**, *18*, 5052−5060.

(56) Churpek, J. E.; Pro, B.; van Besien, K.; Kline, J.; Conner, K.; Wade, J. L., 3rd; Hagemeister, F.; Karrison, T.; Smith, S. M. A Phase 2 Study of Epothilone B Analog BMS-247550 (NSC 710428) in Patients with Relapsed Aggressive Non-Hodgkin Lymphomas. *Cancer* **2013**, *119*, 1683−1689.

(57) Dougherty, P. G.; Qian, Z.; Pei, D. Macrocycles as Protein-Protein Interaction Inhibitors. *Biochem. J.* **2017**, *474*, 1109−1125.

(58) Bonnet, P.; Agrafiotis, D. K.; Zhu, F.; Martin, E. Conformational Analysis of Macrocycles: Finding What Common Search Methods Miss. *J. Chem. Inf. Model.* **2009**, *49*, 2242−2259.

(59) Chen, I.-J.; Foloppe, N. Tackling the Conformational Sampling of Larger Flexible Compounds and Macrocycles in Pharmacology and Drug Discovery. *Bioorg. Med. Chem.* **2013**, *21*, 7898−7920.

(60) Watts, K. S.; Dalal, P.; Tebben, A. J.; Cheney, D. L.; Shelley, J. C. Macrocycle Conformational Sampling with MacroModel. *J. Chem. Inf. Model.* **2014**, *54*, 2680−2696.

(61) Coutsias, E. A.; Lexa, K. W.; Wester, M. J.; Pollock, S. N.; Jacobson, M. P. Exhaustive Conformational Sampling of Complex Fused Ring Macrocycles Using Inverse Kinematics. *J. Chem. Theory Comput.* **2016**, *12*, 4674−4687.

(62) Cleves, A. E.; Jain, A. N. ForceGen 3D Structure and Conformer Generation: From Small Lead-like Molecules to Macrocyclic Drugs. *J. Comput.-Aided Mol. Des.* **2017**, *31*, 419−439.

(63) Sindhikara, D.; Spronk, S. A.; Day, T.; Borrelli, K.; Cheney, D. L.; Posy, S. L. Improving Accuracy, Diversity, and Speed with Prime Macrocycle Conformational Sampling. *J. Chem. Inf. Model.* **2017**, *57*, 1881−1894.

(64) Kamenik, A. S.; Lessel, U.; Fuchs, J. E.; Fox, T.; Liedl, K. R. Peptidic Macrocycles - Conformational Sampling and Thermodynamic Characterization. *J. Chem. Inf. Model.* **2018**, *58*, 982−992.

(65) OMEGA v3.0.0 released. https://www.eyesopen.com/news/omega-v3.0.0-released (accessed September 7, 2018).

(66) Kaufman, L.; Rousseeuw, P. Clustering by Means of Medoids. In *Statistical Data Analysis Based on the L1−Norm and Related Methods*; Birkhäuser: Basel, 1987, pp 405−416.

(67) Jin, X.; Han, J. K-Medoids Clustering. In *Encyclopedia of Machine Learning*; Sammut, C., Webb, G. I., Eds.; Springer US: Boston, MA, 2010; p 22.

C.2 Supporting Information of Published Journal Articles

Supporting Information

High-Quality Dataset of Protein-Bound Ligand Conformations and Its Application to Benchmarking Conformer Ensemble Generators

*Nils-Ole Friedrich,[1] Agnes Meyder,[1] Christina de Bruyn Kops,[1] Kai Sommer,[1] Florian Flachsenberg,[1] Matthias Rarey,[1] Johannes Kirchmair[1]**

[1]University of Hamburg, ZBH – Center for Bioinformatics, Bundesstraße 43, Hamburg 20146, Germany

*corresponding author
kirchmair@zbh.uni-hamburg.de
tel. +49 (0)40 42838 7303

TABLE OF CONTENTS

BENCHMARK RESULTS

Figure S1: Platinum Dataset (4626 ligands) 3
Figure S2: Platinum Diverse Dataset (2912 ligands) 4
Figure S3: Success rates with cutoffs of 0.6 Å and 1 Å 5

CORRELATION

Figure S4: Accuracy of RDKit and ETKDG for individual molecules 6
Figure S5: Runtime and accuracy 7
Figure S6: Ensemble size and accuracy 8

STATISTICAL ANALYSIS

Table S1: Mann-Whitney U Test Results 9
Table S2: Arithmetic Mean, Median and Interquartile Range of RMSD Values 10
Table S3: Arithmetic Mean, Median and Interquartile Range of Ensemble Sizes 10
Table S4: Arithmetic Mean, Median and Interquartile Range of Runtimes 11

BENCHMARKING RESULTS

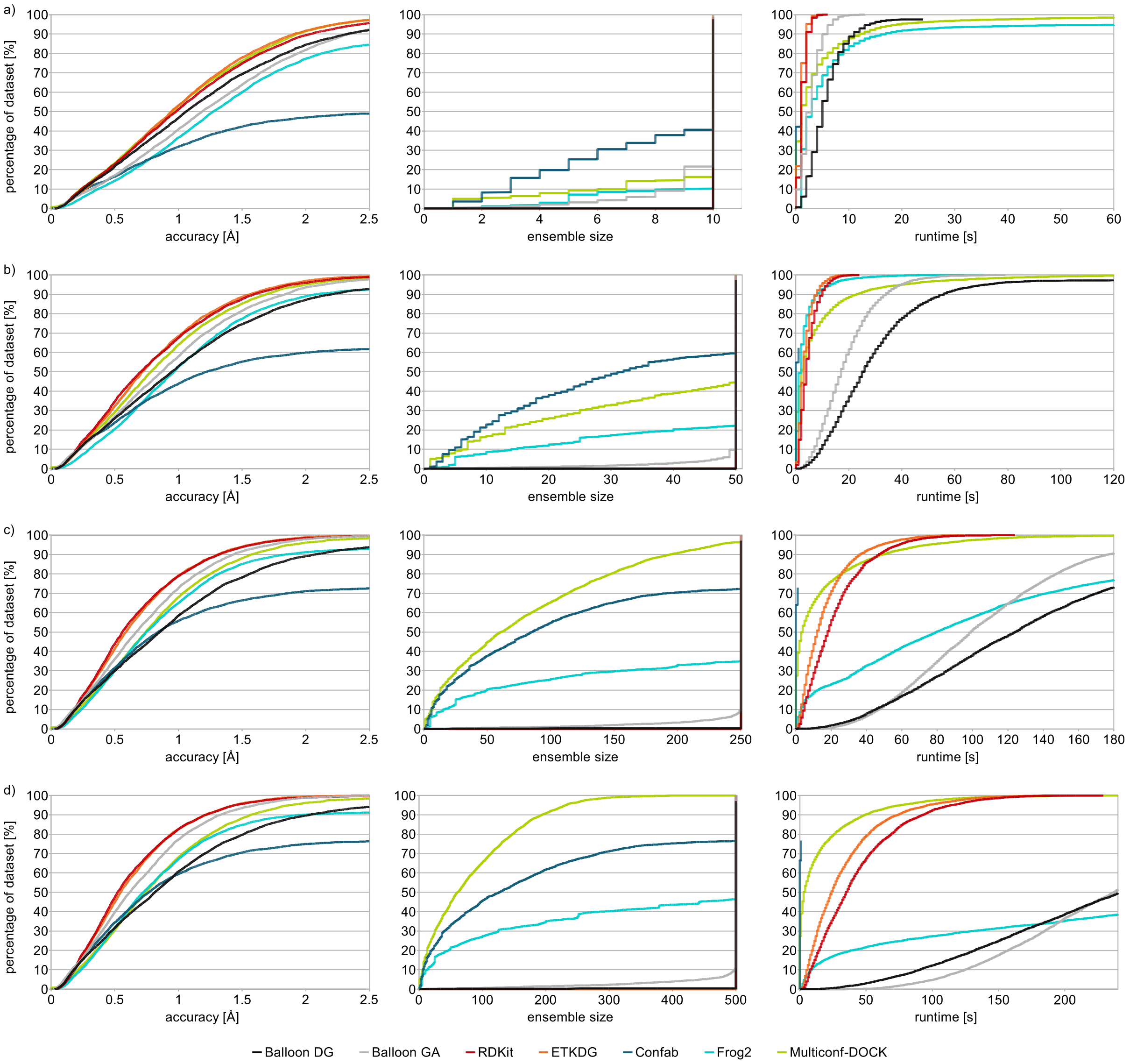


Figure S1. Percentage of protein-bound ligand conformations of the Platinum Dataset reproduced by different conformer ensemble generators vs. accuracy, ensemble size and runtime. Maximum ensemble size (a) 10, (b) 50, (c) 250 and (d) 500 conformations.

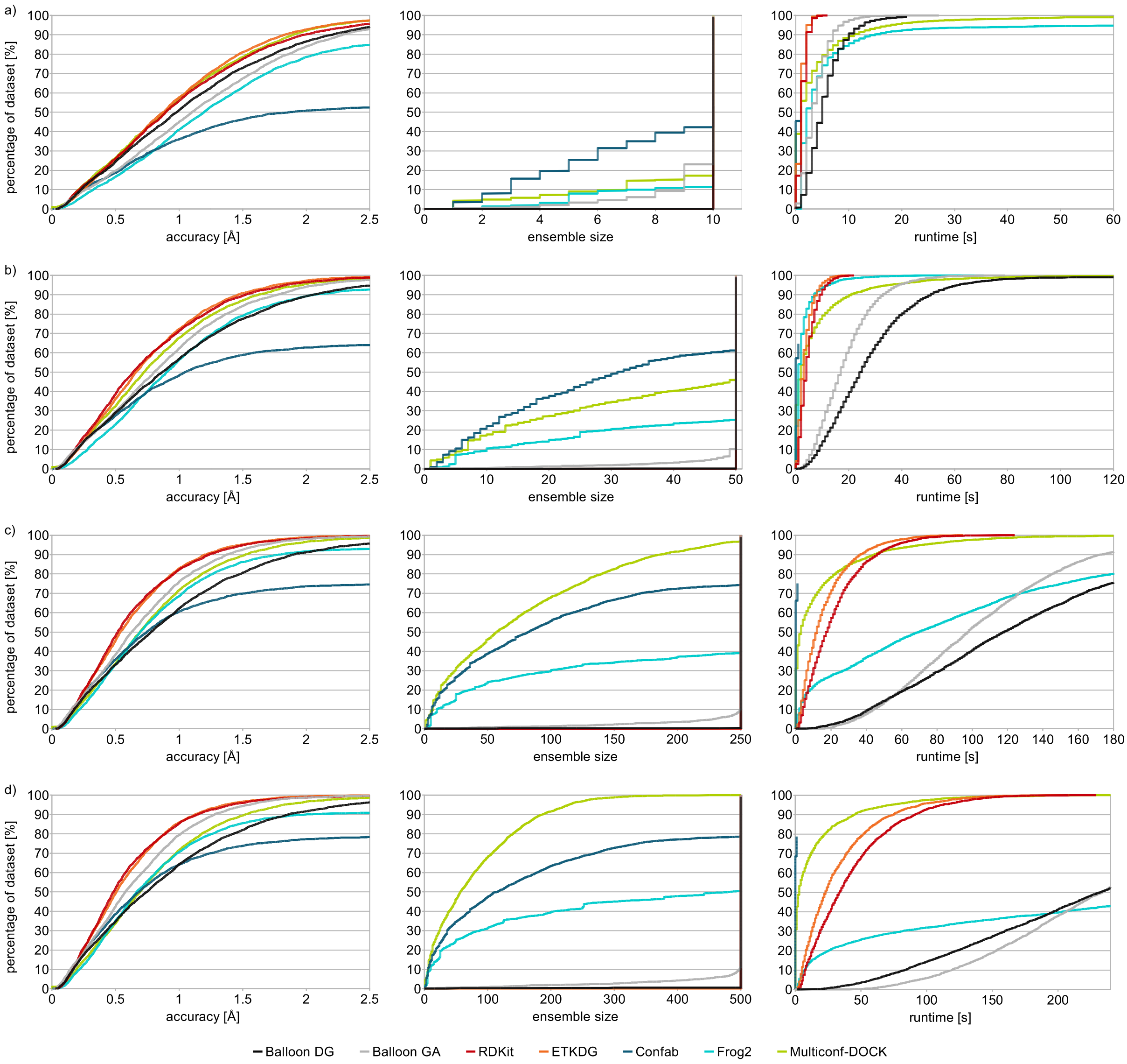


Figure S2. Percentage of protein-bound ligand conformations of the Platinum Diverse Dataset reproduced by different conformer ensemble generators vs. accuracy, ensemble size and runtime. Maximum ensemble size (a) 10, (b) 50, (c) 250 and (d) 500 conformations.

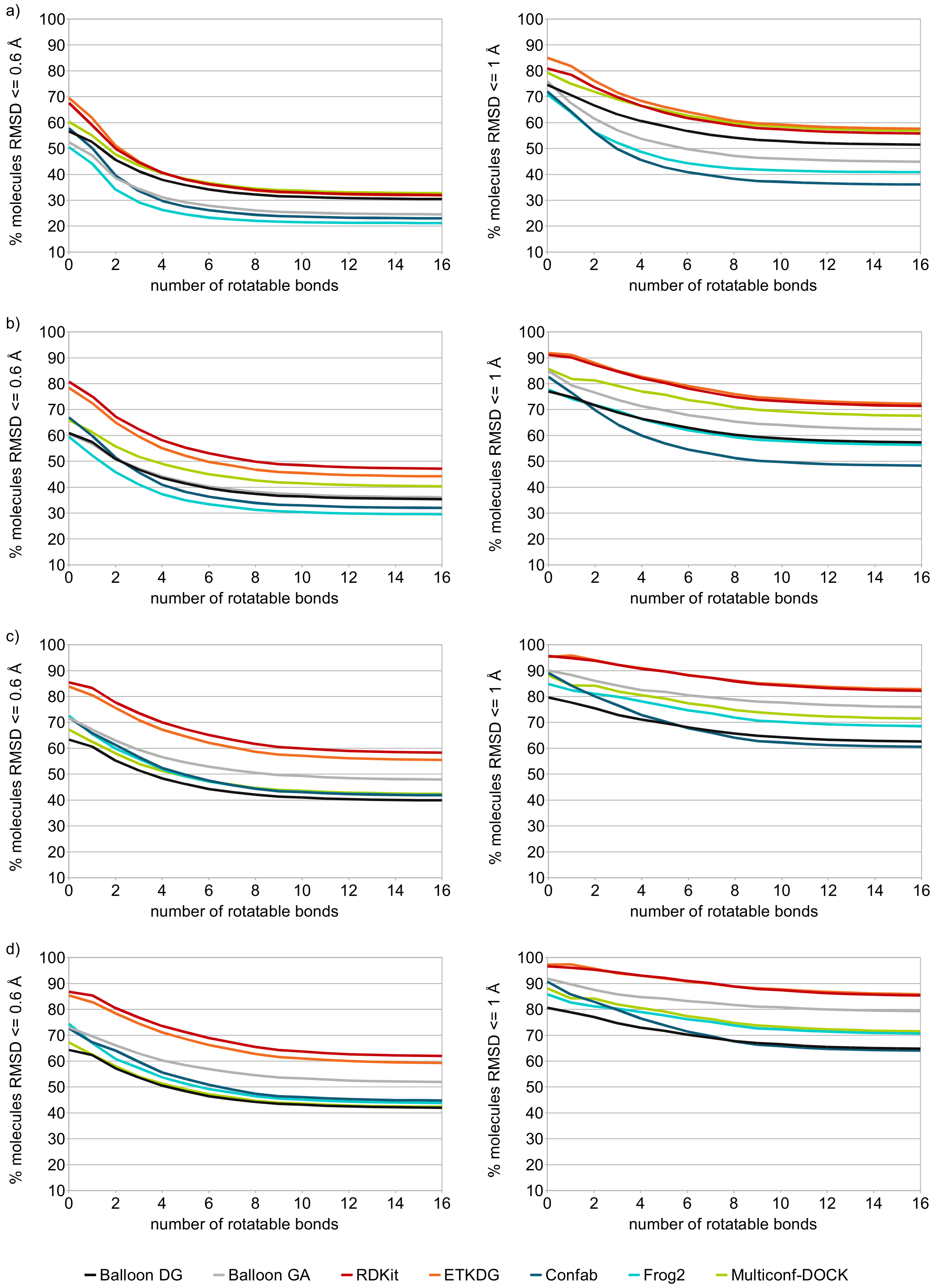


Figure S3. Percentage of molecules of the Platinum Diverse Dataset reproduced with a defined maximum RMSD as a function of the number of rotatable bonds. Maximum ensemble size (a) 10, (b) 50, (c) 250 and (d) 500 conformations.

CORRELATION

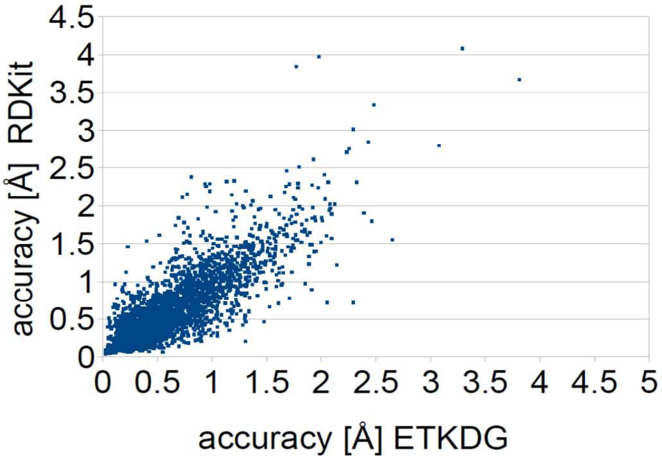


Figure S4. Accuracy of conformer ensembles generated with RDKit vs. ETKDG for individual molecules of the Platinum Diverse Dataset. Maximum ensemble size 250 conformations.

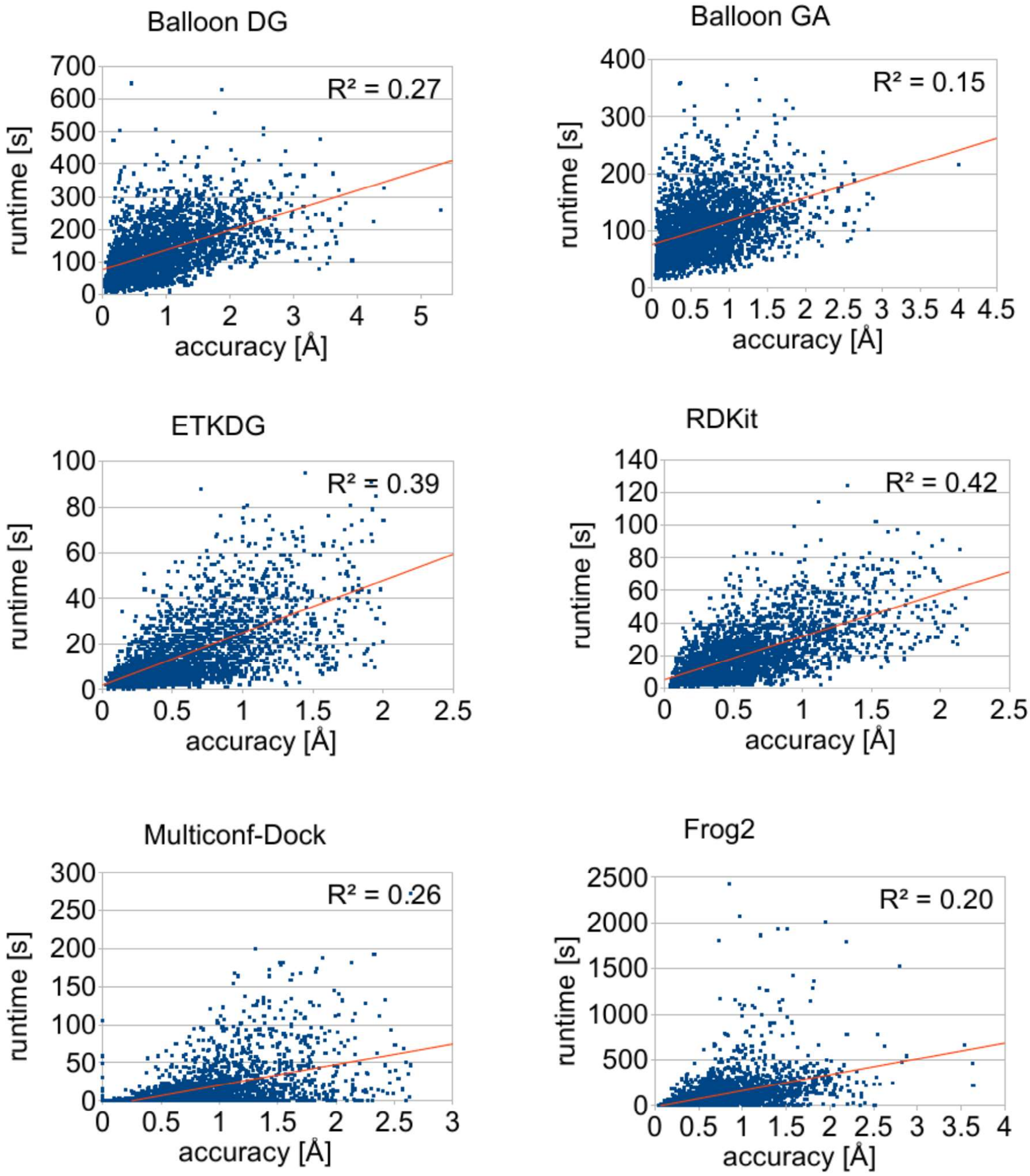


Figure S5. Correlation between runtime and accuracy for individual molecules of the Platinum Diverse Dataset. Maximum ensemble size 250 conformations. Confab is not reported because its runtimes are often below measurement accuracy.

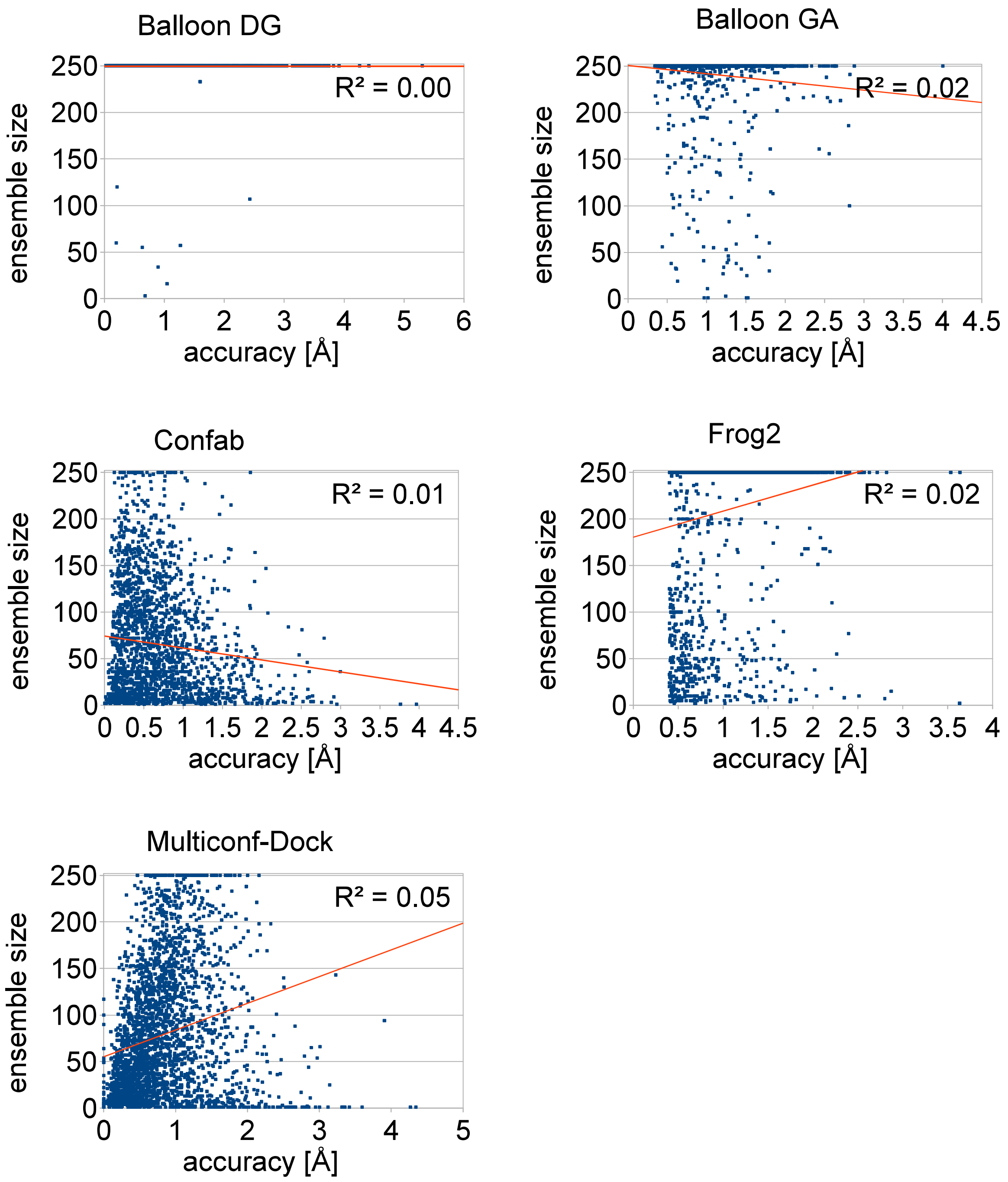


Figure S6. Correlation between ensemble size and accuracy for individual molecules of the Platinum Diverse Dataset. Maximum ensemble size 250 conformations. RDKit and ETKDG are not reported because the size of their ensembles is equal to the maximum ensemble size in all cases.

STATISTICAL ANALYSIS

Table S1: Results of the Mann-Whitney U Test for RMSD Values [Å] Obtained for the Platinum Diverse Dataset.[a]

Tool	**Value**	**Balloon GA**	**RDKit**	**ETKDG**	**Confab**	**Frog2**	**Multiconf-DOCK**
Balloon DG	U	3354288	3217357	3248646	2435626	3516498	3890709
	Z	-13.21	-15.36	-14.79	-13.90	-6.60	-4.80
	p	< 0.00048	< 0.00048	< 0.00048	< 0.00048	< 0.00048	< 0.00048
Balloon GA	U	-	3774070	3823893	2850699	3772593	3921345
	Z	-	-7.20	-6.34	-6.40	-2.97	-4.92
	p	-	< 0.00048	< 0.00048	< 0.00048	< 0.00048	< 0.00048
RDKit	U	-	-	4160167	3194398	3318314	3458100
	Z	-	-	-1.11	0.19	-10.45	-12.17
	p	-	-	0.27	0.85	< 0.00048	< 0.00048
ETKDG	U	-	-	-	3117179	3396693	3504441
	Z	-	-	-	-1.21	-9.08	-11.36
	p	-	-	-	0.22	< 0.00048	< 0.00048
Confab	U	-	-	-	-	2511059	2619925
	Z	-	-	-	-	-9.36	-10.87
	p	-	-	-	-	< 0.00048	< 0.00048
Frog2	U	-	-	-	-	-	3831875
	Z	-	-	-	-	-	-2.04
	p	-	-	-	-	-	0.02

[a] A significance level of < 0.00048 corresponds to a family-wise error rate (FWER) of 0.01. Based on this test, the performance of RDKit and ETKDG is not significantly different. Any statistically significant results for Confab should be considered in the context of its limited ability to produce conformer ensembles (Table 6).

Table S2: Arithmetic Mean, Median and Interquartile Range of RMSD Values for Ensembles of a Maximum of 250 Conformers Determined for the Platinum Diverse Dataset.

	mean [Å]	median [Å]	interquartile range [Å]
Balloon DG	0.92	0.77	0.91
Balloon GA	0.72	0.63	0.42
RDKit	0.63	0.52	0.55
ETKDG	0.63	0.54	0.56
Confab	0.65	0.53	0.62
Frog2	0.75	0.65	0.64
Multiconf-DOCK	0.80	0.69	0.69

Table S3: Arithmetic Mean, Median and Interquartile Range of Ensemble Sizes for Ensembles of a Maximum of 250 Conformers Determined for the Platinum Diverse Dataset.

	mean	median	interquartile range
Balloon DG	249	250	0
Balloon GA	244	250	204
RDKit	250	250	0
ETKDG	250	250	0
Confab	65	48	89
Frog2	176	250	195
Multiconf-DOCK	78	57	107

Table S4: Arithmetic Mean, Median and Interquartile Range of Runtimes for Ensembles of a Maximum of 250 Conformers Determined for the Platinum Diverse Dataset.

	mean [s]	median [s]	interquartile range [s]
Balloon DG	132	117	105
Balloon GA	105	98	90
RDKit	22	18	21
ETKDG	16	12	17
Confab	<1	<1	0
Frog2	128	67	133
Multiconf-DOCK	15	3	16

Supporting Information

Benchmarking Commercial Conformer Ensemble Generators

*Nils-Ole Friedrich,[1] Christina de Bruyn Kops,[1] Florian Flachsenberg,[1] Kai Sommer,[1] Matthias Rarey,[1] Johannes Kirchmair[1]**

[1] Universität Hamburg, Center for Bioinformatics, Bundesstr. 43, Hamburg, 20146, Germany

*corresponding author
kirchmair@zbh.uni-hamburg.de
tel. +49 (0)40 42838 7303

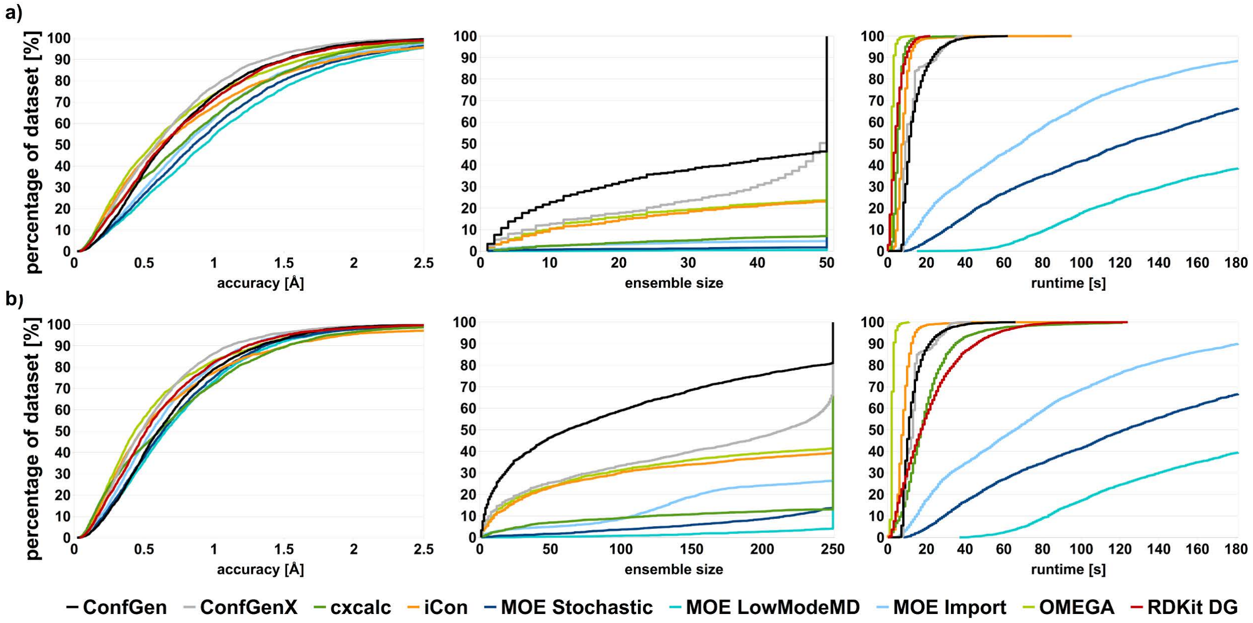


Figure S1. Percentage of protein-bound ligand conformations of the Platinum Diverse Dataset reproduced by the different tools within a certain accuracy (left), ensemble size (middle) and runtime (right) at maximum ensemble sizes a) 50 and b) 250. Clustering by RMSD deactivated for iCon, MOE Stochastic, MOE LowModeMD, OMEGA and the RDKit DG algorithm (UFF).

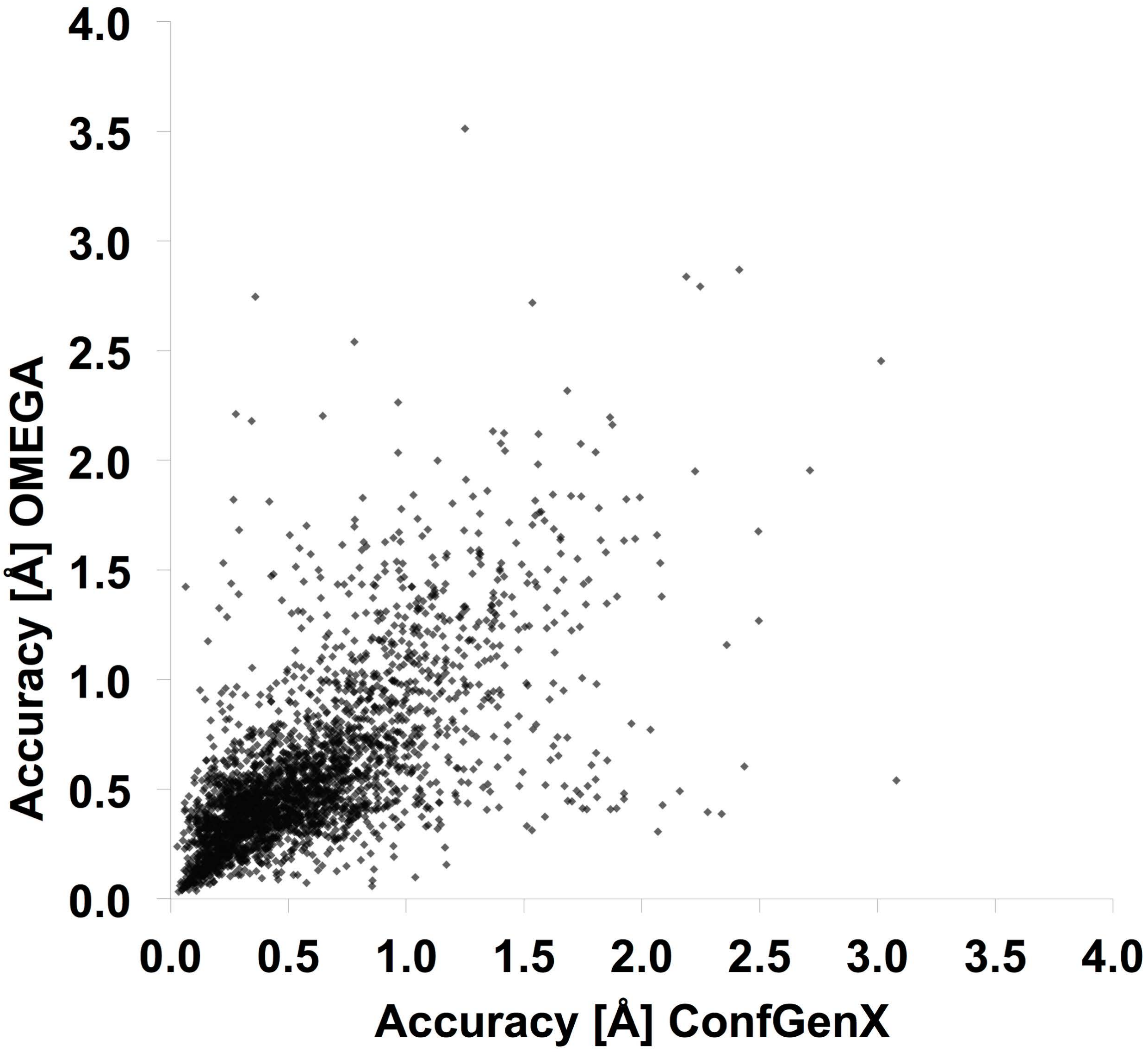


Figure S2. Accuracy of conformer ensembles generated with ConfGenX vs. OMEGA for individual molecules of the Platinum Diverse Dataset. Maximum ensemble size 250 conformations.

Supplementary Material

How Diverse are the Protein-Bound Conformations of Small-Molecule Drugs and Cofactors?

Nils-Ole Friedrich, Méliné Simsir, Johannes Kirchmair*

*** Correspondence:** J. Kirchmair
E-mail: kirchmair@zbh.uni-hamburg.de
Tel.: +49 (0)40 42838 7303

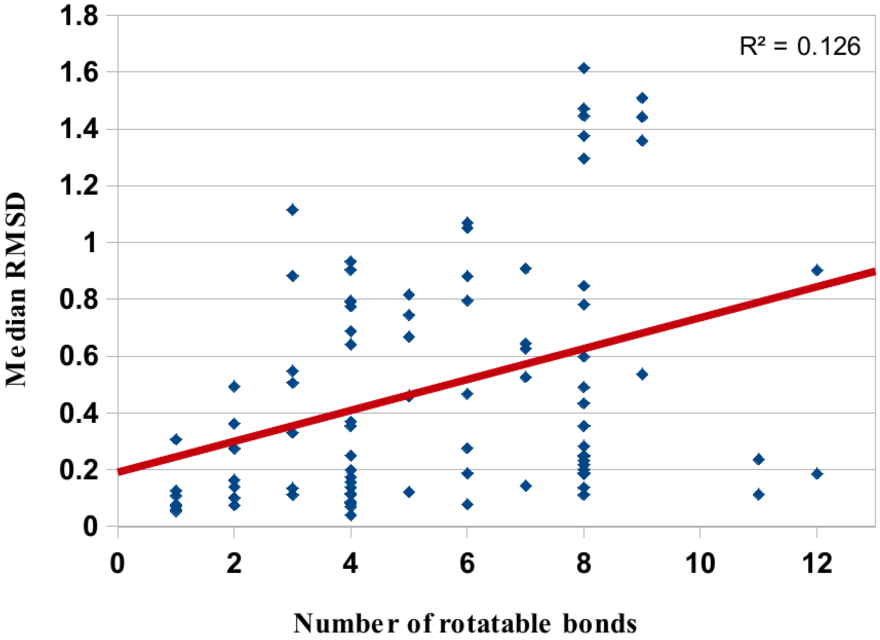


Figure S1. Minimum median RMSD values plotted against the number of rotatable bonds.

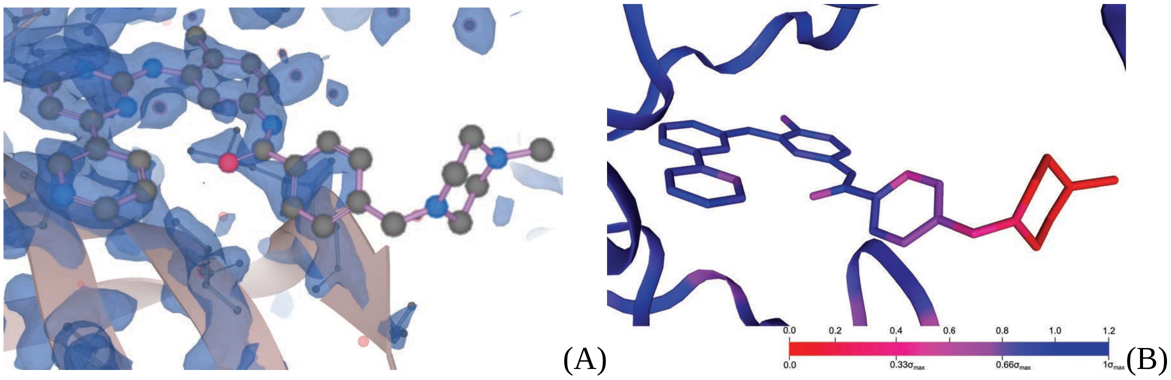


Figure S2. (A) Weak electron density support for parts of imatinib bound to human SYK modeled in an uncommon orientation (1XBB; $EDIA_m$ = 0.21). (B) The EDIA score is indicated by a color gradient, ranging from dark red (no or poor electron density support) via magenta to blue (good electron density support). For single atoms, EDIA values above 0.8 mark well-supported atoms, values in the range of 0.4 to 0.8 atoms with medium support and values below 0.4 poorly-supported atoms. For the EDIAm, the developers of the method concluded that only structures with a score higher than 0.8 should be considered as well supported by the electron density.

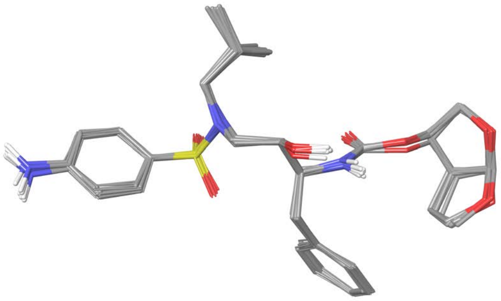


Figure S3. Ligand-based alignment of all 14 conformers of darunavir present in the Sperrylite Dataset.

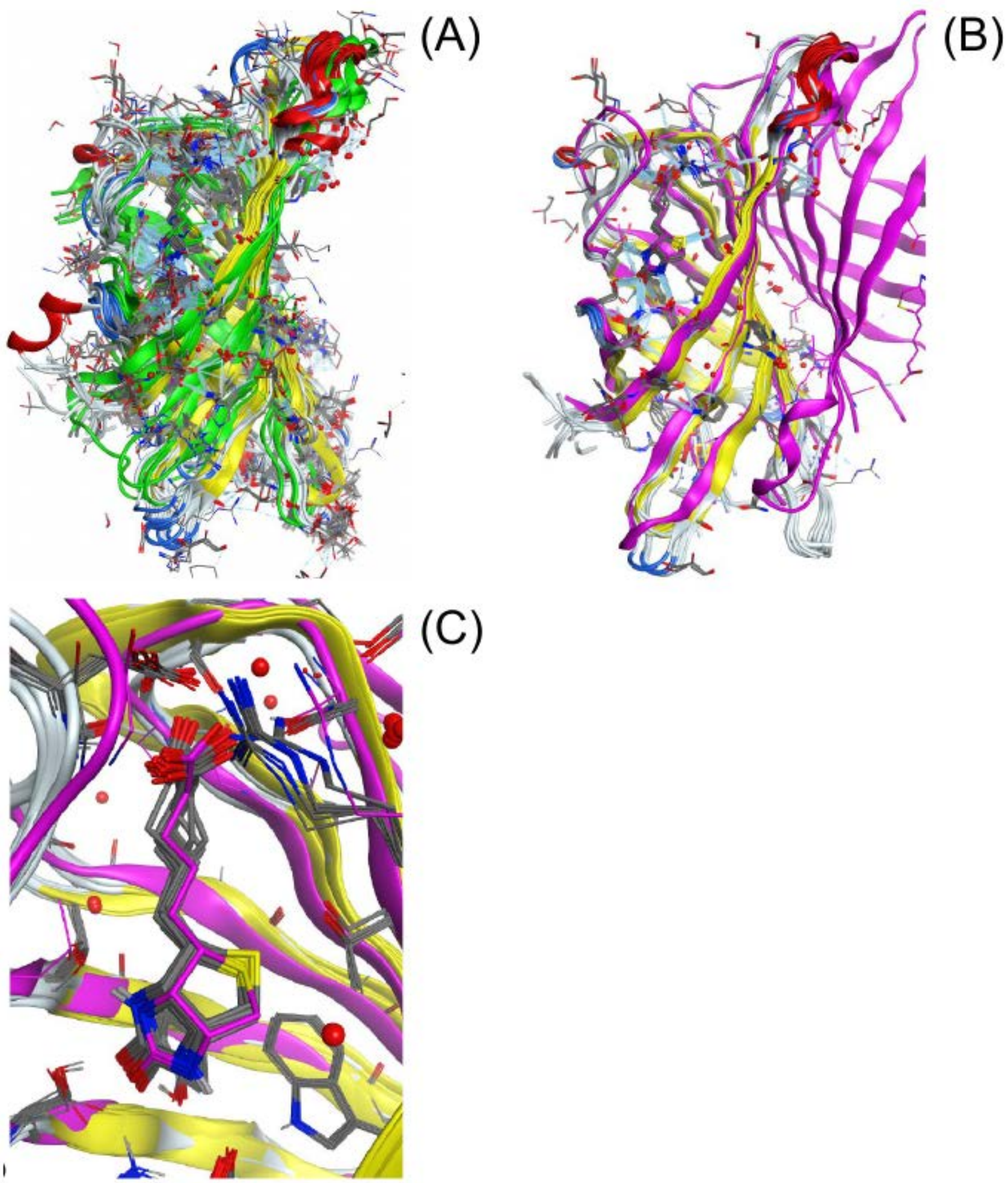


Figure S4. (A) Superposition of all 34 high-quality core streptavidin structures (including tetramers) bound to biotin. The four high-quality structures of the streptavidin N49/G48 mutant from *Streptomyces avidinii* (4GD9) in green show the structural consequences of cutting a binding loop. (B) Ten high-quality core streptavidin structures superposed with the loosely-related crystal structure of engineered dual chain avidin (2C4I), in violet, (C) with very similar binding modes. In dual chain avidin, two circularly permuted chicken avidin monomers are fused into one polypeptide chain.

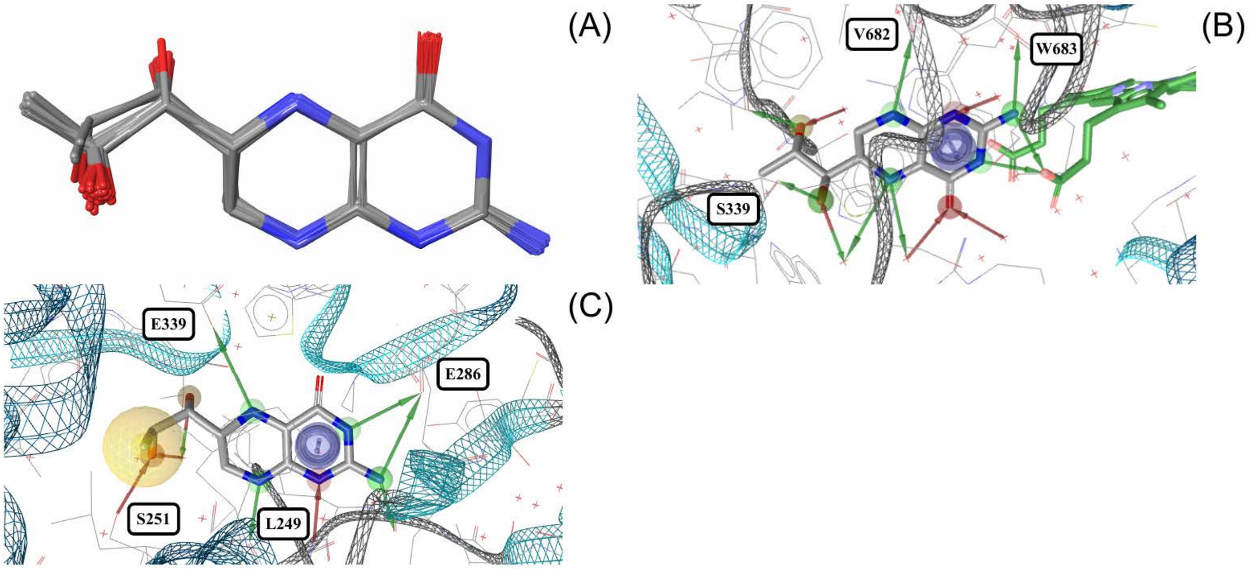


Figure S5. (A) Superposition of 188 high-quality structures of sapropterin. Sapropterin in the binding pocket of (B) human nitric oxide synthase (4D1N, hemoglobin in green) is missing a hydrophobic interaction present in (C) human phenylalanine hydroxylase (1MMK).

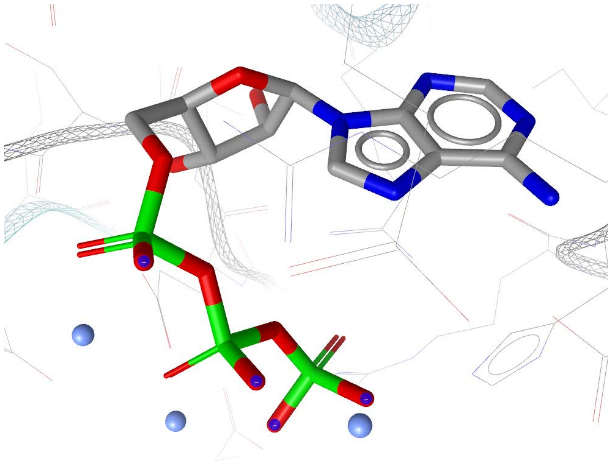
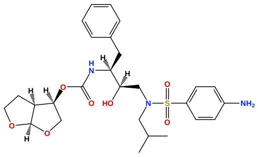
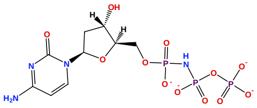
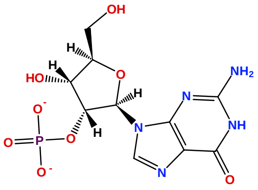
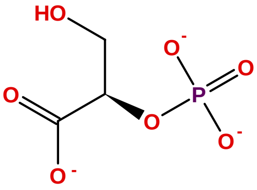
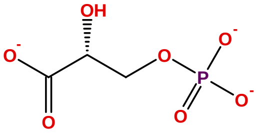
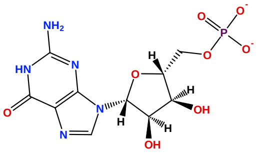
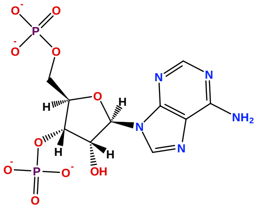
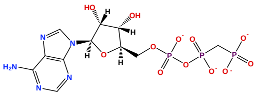
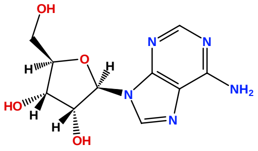
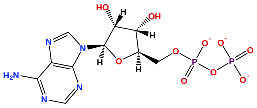
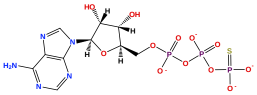
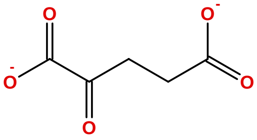
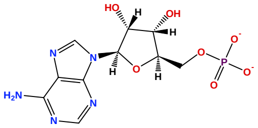
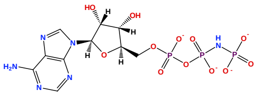
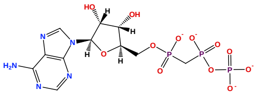
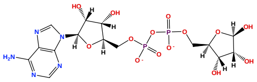
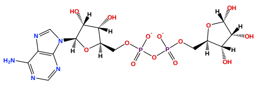
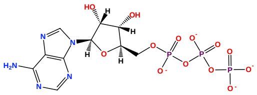
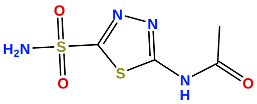
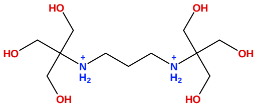
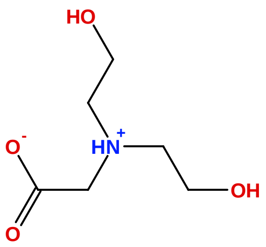
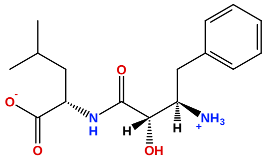
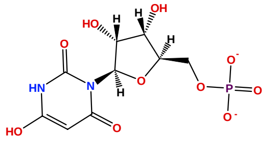
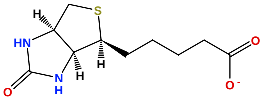
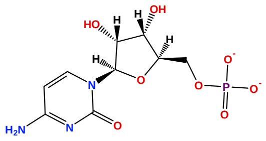
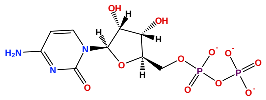
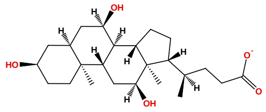
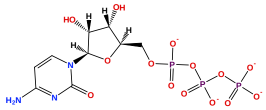
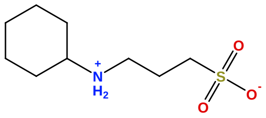
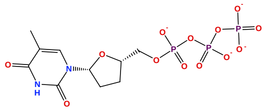
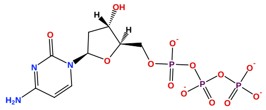
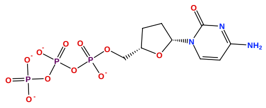
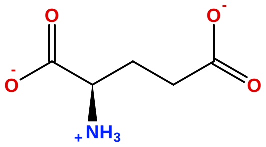
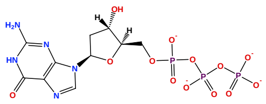
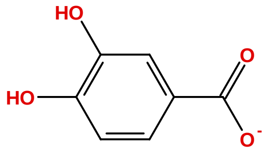
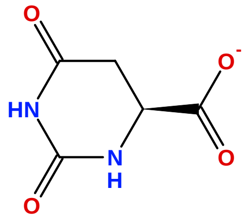
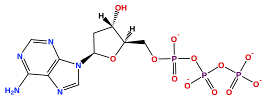
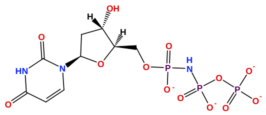
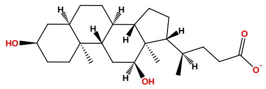
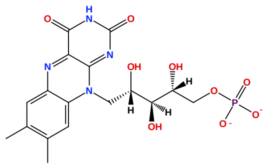
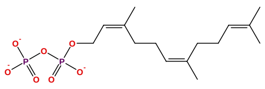
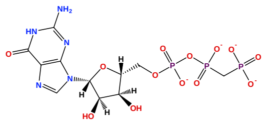
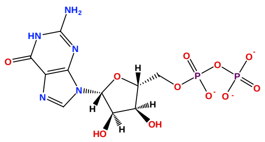
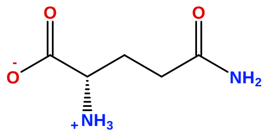
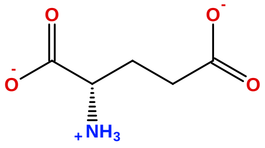
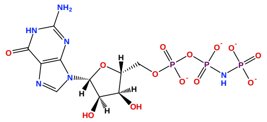
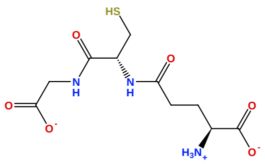
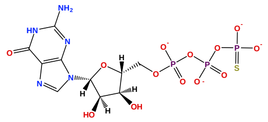
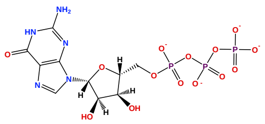
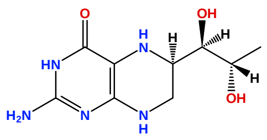
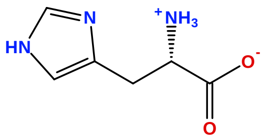
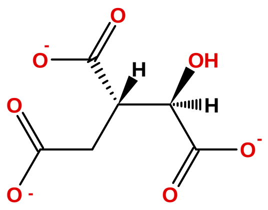
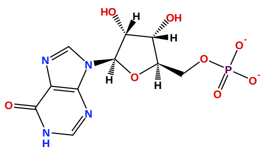
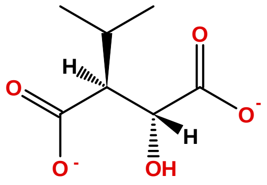
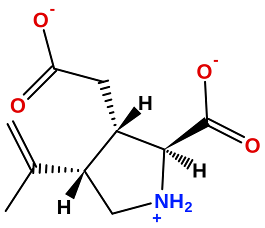
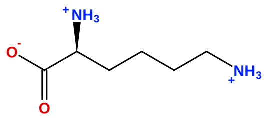
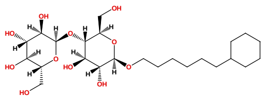
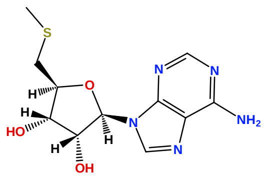
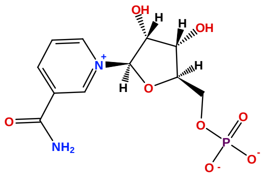
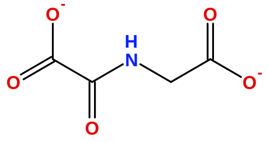


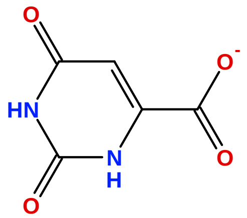
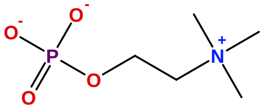
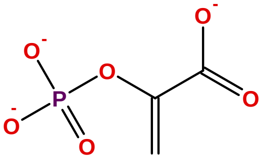
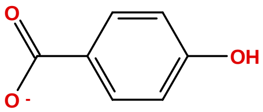
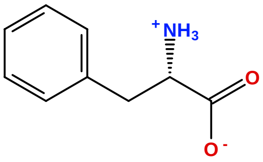
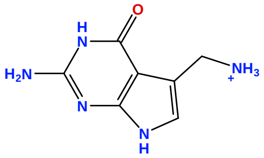
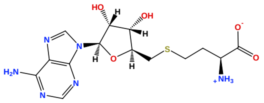
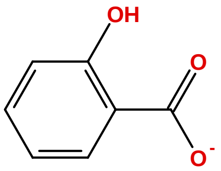
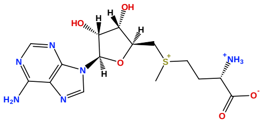
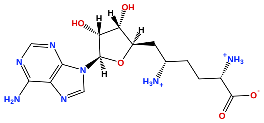
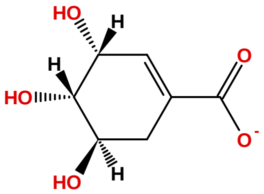
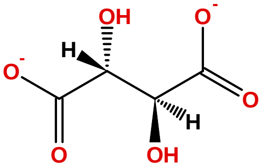
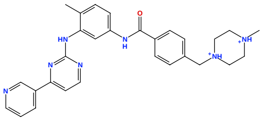
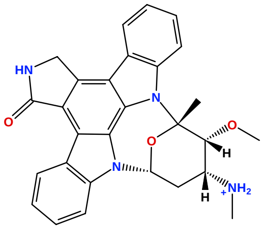
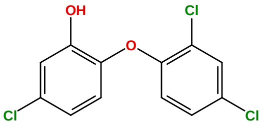
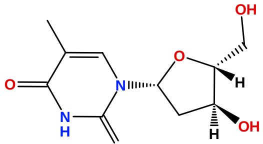
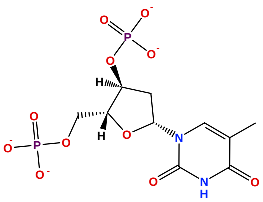
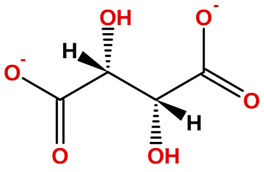
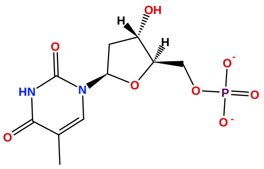
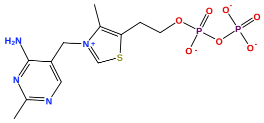
Figure S6. Unusually bent conformer of ATP in the binding pocket of aspartyl-tRNA synthetase from *Pyrococcus kodakaraensis* (1B8A) interacting with three manganese atoms (light blue).

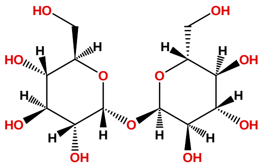
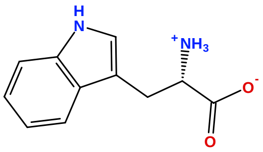
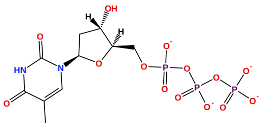
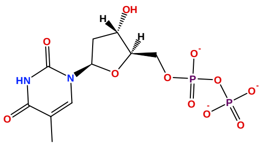
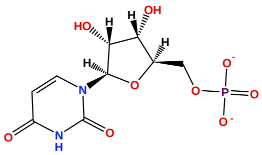
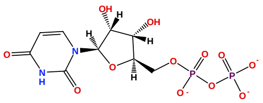
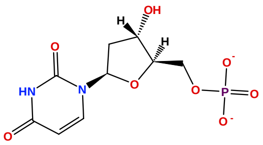
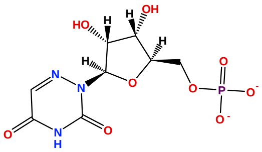
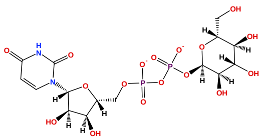
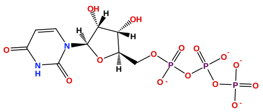
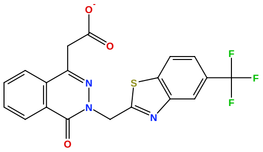
Scheme S1. Overview of all (91) small molecules represented by at least ten conformers in the Sperrylite Dataset.[1]

 017 14	 0KX 11	 2GP 14	 2PG 10
 3PG 23	 5GP 31	 A3P 29	 ACP 29
 ADN 57	 ADP 462	 AGS 17	 AKG 77
 AMP 171	 ANP 140	 APC 27	 APR 13
 AR6 15	 ATP 212	 AZM 10	 B3P 22

 BCN 13	 BES 11	 BMP 29	 BTN 43
 C5P 30	 CDP 17	 CHD 13	 CTP 19
 CXS 18	 D3T 10	 DCP 39	 DCT 11
 DGL 14	 DGT 28	 DHB 10	 DOR 15
 DTP 33	 DUP 26	 DXC 18	 FMN 370

 FPP 15	 GCP 14	 GDP 261	 GLN 16
 GLU 113	 GNP 139	 GSH 74	 GSP 20
 GTP 96	 H4B 188	 HIS 28	 ICT 10
 IMP 26	 IPM 14	 KAI 10	 LYS 18
 MA4 18	 MTA 24	 NMN 13	 OGA 44

 ORO 41	 PC 10	 PEP 41	 PHB 12
 PHE 14	 PRF 10	 SAH 311	 SAL 18
 SAM 119	 SFG 30	 SKM 20	 SRT 17
 STI 11	 STU 18	 TCL 12	 THM 18
 THP 127	 TLA 122	 TMP 20	 TPP 30

 TRE 27	 TRP 22	 TYP 30	 TYD 14
 U5P 21	 UDP 102	 UMP 41	 UP6 10
 UPG 20	 UTP 14	 ZST 11	

[1] For each ligand (specified by the PDB three-letter ligand identifier) the number of conformers in the Sperrylite Dataset is reported.

Supporting Information

Conformator: A Novel Method for the Generation of Conformer Ensembles

Nils-Ole Friedrich,[1] *Florian Flachsenberg,*[1] *Agnes Meyder,*[1] *Kai Sommer,*[1] *Johannes Kirchmair,*[1,2,3] *Matthias Rarey*[1]*

[1] Universität Hamburg, Center for Bioinformatics, Bundesstr. 43, Hamburg, 20146, Germany

[2] Department of Chemistry, University of Bergen, N-5020 Bergen, Norway

[3] Computational Biology Unit (CBU), University of Bergen, N-5020 Bergen, Norway

*E-mail: rarey@zbh.uni-hamburg.de. Tel.: +49 40 42838-7351.

TABLE OF CONTENTS

Table S1: Empirically Determined Weights for the MCOS ….. 3

Table S2: Mann-Whitney U Test Results (250)…... 4

Table S3: Mann-Whitney U Test Results (50)…... 5

Table S4: Complete Sets of Conformers …... 6

Algorithm S1: RMSD-Clustering of Conformers 7

Figure S1: Visualization of Conformator's Clustering Algorithm 8

Figure S2: Visualization of Macrocycle Conformer Generation …... 9

Figure S3: MCOS function for the overlay score 11

Figure S4: MCOS function for the bond angle score 11

Figure S5: MCOS penalty function for limiting bond angles 12

Figure S6: MCOS bond length term 12

Figure S7: MCOS distance factor 13

Figure S8 MCOS bond angle factor 14

Figure S9: MCOS clash score 15

Figure S10: Median pairwise RMSD 16

Figure S11: Minimum pairwise RMSD 16

Table S1. Empirically Determined Weights for the MCOS.

Contribution	Weight
$w_{overlay}$	1.0
w_{bond}	1.0
w_{angle}	1.0
w_{limit}	500.0
$w_{torsion}$	0.1
$w_{torsion,conjugated}$	1.0
w_{clash}	1.0

Table S2. Mann-Whitney U Test Results of RMSD Values from Conformer Ensemble Generation for Platinum Diverse Dataset with a Maximum of 250 Conformers.[a]

Conformer ensemble generator		RDKit DG (UFF and clustering)	OMEGA (default)	Conformator Fast	Conformator Best
OMEGA (default)	**p**	< 0.001	-		
	Z	-5.05	-		
	U	3747852	-		
Conformator Fast	**p**	**0.03**	< 0.001	-	
	Z	-1.85	-7.25	-	
	U	3959194	3614564	-	
Conformator Best	**p**	< 0.001	**0.07**	< 0.001	-
	Z	-6.03	-1.48	-8.09	-
	U	3698854	3973330	3574312	-
CONFECT	**p**	< 0.001	< 0.001	< 0.001	< 0.001
	Z	-7.15	-11.62	-5.69	-12.47
	U	3387977	3116302	3478176	3076036

[a] The Mann-Whitney U test was used to test for statistical significance at $\alpha = 0.05$ and $\alpha = 0.01$, adjusted with the Holm−Bonferroni method to control the familywise error rate. Differences between Conformator Best and OMEGA, as well as Conformator Fast and the RDKit DG algorithm are not statistically significant (bold p values).

Table S3. Mann-Whitney U Test Results of RMSD Values from Conformer Ensemble Generation for Platinum Diverse Dataset with a Maximum of 50 Conformers.[a]

Conformer ensemble generator		**RDKit DG (UFF and clustering)**	**OMEGA (default)**	**Conformator Fast**	**Conformator Best**
OMEGA (default)	p	< 0.001	-		
	Z	-10.12	-		
	U	3399855	-		
Conformator Fast	p	**0.03**	< 0.001	-	
	Z	-1.94	-8.50	-	
	U	3917040	3537062	-	
Conformator Best	p	< 0.001	**0.02**	< 0.001	-
	Z	-7.89	-2.06	-6.18	-
	U	3549130	3937297	3693374	-
CONFECT	p	< 0.001	< 0.001	< 0.001	< 0.001
	Z	-3.78	-12.25	-5.47	-10.74
	U	3550478	3075448	3487404	3174824

[a] The Mann−Whitney U test was used to test for statistical significance at $\alpha = 0.05$ and $\alpha = 0.01$, adjusted with the Holm−Bonferroni method to control the familywise error rate. Differences between Conformator Best and OMEGA, as well as Conformator Fast and the RDKit DG algorithm are not statistically significant (bold p values).

Table S4. Percentage of Structures Successfully Reproduced within a Specified RMSD Threshold by Complete Sets of Conformers[a]

Setting	**RMSD threshold [Å]**		
	0.5	**1.0**	**1.5**
complete[b]	92	99	100
default	84	98	99

[a] On a subset of the Platinum Diverse Dataset of 987 molecules (with a maximum of 6 rotatable bonds)

[b] Conformator Best, no clustering, no maximum ensemble size, maximum runtime of 72 h per molecule

Algorithm S1 RMSD-Clustering of Conformers[a]

```
Input: List of conformers (Y)              //candidate conformers, partially presorted
Input: quality_level                 //1 = Fast, 2 = Best (default 2)
Input: max_ensemble_size             //maximum ensemble size (default 250)
Output: List of cluster centers (Z)  //output conformer ensemble

rmsd_threshold ← 0.1                 //RMSD starting threshold in Å for Best (default)
rmsd_increase ← 0.05                 //RMSD threshold enlargement per round
if (quality_level == 1)
        rmsd_threshold ← 0.5         //RMSD starting threshold in Å for Fast
        rmsd_increase ← 0.5
while (Z.size() > max_ensemble_size)       //starting new clustering
        Z.clear                            //empty list of cluster centers
        candidate_conformer ← Y.begin()    //first conformation is the first cluster center
        while (candidate_conformer != Y.end())           //starting new clustering round
                for (cluster_center = Z.end() to cluster_center Z.begin()     //in reverse
                        rmsd = calculate_rmsd(candidate_conformer, cluster_center)
                        if (rmsd < rmsd_threshold)
                                tooclose ← true
                                break      //no further comparisons
                end for
                if (tooclose)
                        Y.erase(candidate_conformer)
                        //remove candidate conformer permanently
                else
                        Z.push_back(candidate_conformer)
                        //add candidate as cluster center
                        candidate_conformer = Y.next()
                if (Z.size() > max_ensemble_size)    //too many cluster centers
                        break                        //start new round (inner while loop)
        end while
        rmsd_threshold ← rmsd_threshold + rmsd_increase
end while
return Z                //output list of cluster centers as the conformer ensemble
```

[a] Note that the representation with two separate lists (Y and Z) was chosen for didactic reasons. The algorithm should be implemented with a single array of conformers running in place with indices marking the current end of the cluster center set and the beginning of the unprocessed conformer list.

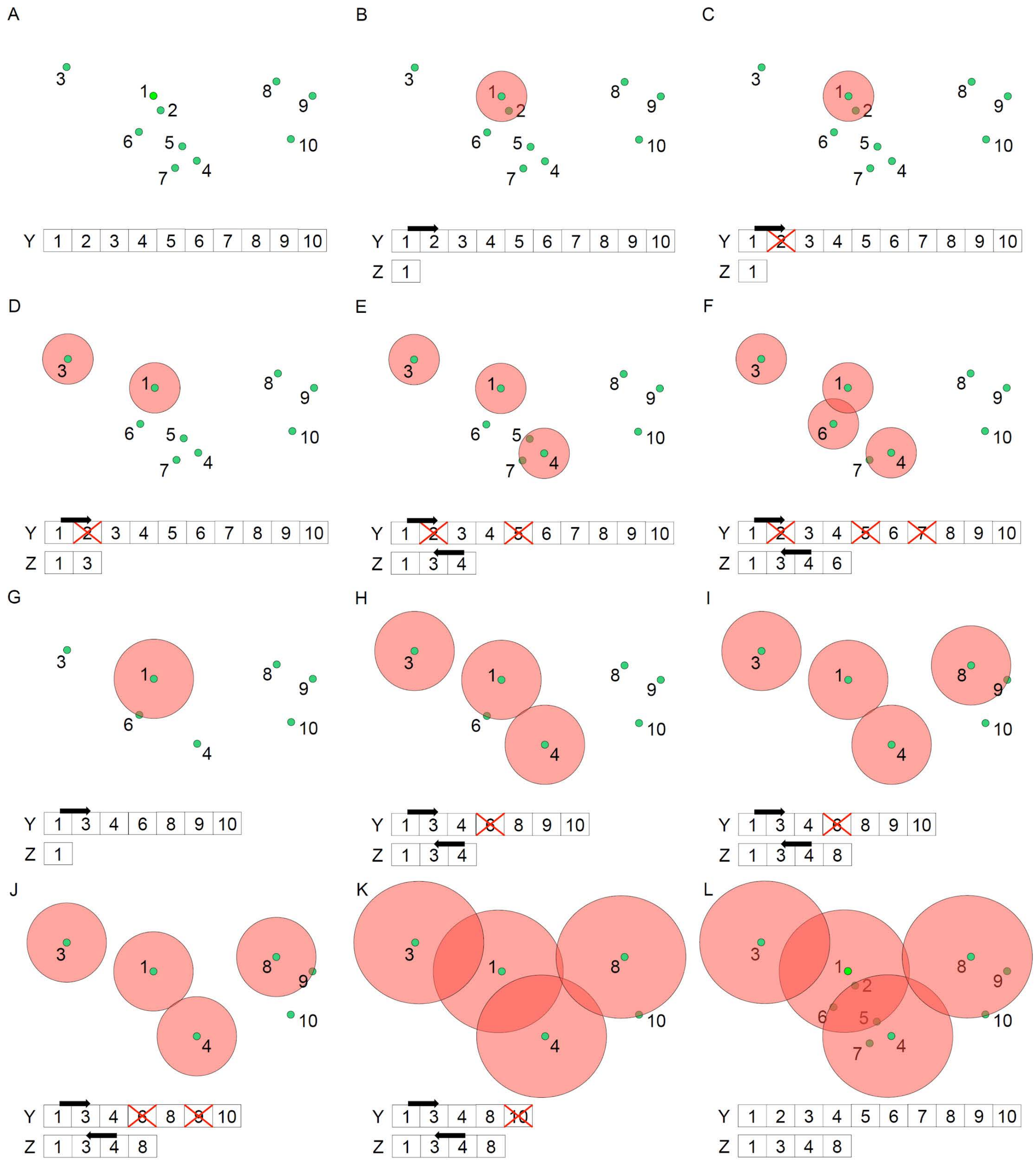


Figure S1. Visualization of Conformator's clustering algorithm by the example of the generation of an ensemble of four representative conformers starting from a set of ten candidate conformers. The green dots represent the candidate conformations. Their distances in 2D space is indicative of their RMSD. The increasing RMSD thresholds are illustrated by the red spheres. Arrows indicate the directions in which the lists of all remaining candidate conformers (Y, top list) and cluster centroids (Z, bottom list) are accessed. Crossed-out numbers indicate conformers that have been removed by the clustering algorithm from the list of candidate conformers. (a) Clustering starts from a list of ten candidate conformations generated with Conformator.

Importantly, these lists are partially presorted, meaning that sequentially generated conformers are likely similar. (b) The first conformer (usually based on very likely torsion angles; see Conformer Generation Algorithm) in the list of candidate conformers is always the first cluster center. (c) The candidate conformers are compared to any of the existing cluster centers. If they are within the RMSD radius (like it is the case for conformer 2) they are removed from the list of candidate conformers. (d) Outliers such as conformer 3 become cluster centers. This behavior is desired as it assures that a sufficiently large part of the relevant conformational space is covered. (e) To take advantage of the fact that conformers generated sequentially with Conformator are likely similar, the list of cluster centers is reversed when comparing candidate conformers to existing cluster centers. While this has no effect on conformer 4 (it is compared against all cluster centers, is dissimilar to all of them and thus becomes a new cluster center), most candidate conformers can be excluded from extensive pairwise comparison, such as conformer 5, which is only compared to conformer 4 before it is removed. (f) Conformer 6 is defined as a new cluster center and conformer 7 is removed from the list of candidate conformers because it is too similar to conformer 4. Conformer 8 is sufficiently distant to any of the existing cluster centers and hence would become a new cluster center. However, this would exceed the maximum ensemble size (which is 4 in this example), for which reason (g) the clustering is repeated with larger RMSD threshold, an empty list of cluster centers and the list of remaining candidates (in other words, previously removed conformers are not considered again). Over several iterations this process determines an appropriate RMSD threshold for each individual molecule. The final threshold depends on the maximum ensemble size and quality level, as well as the size and flexibility of the molecule. (h) Conformers 1, 3 and 4 are again defined as cluster centers but the former cluster center “conformer 6” is removed since it is closer to conformer 1 than the increased distance value allows. (i) Conformer 8 is another cluster center and conformer 9 is removed from the list of candidate conformers. Conformer 10 would become the next cluster center but this would exceed the maximum ensemble size. (j) Once more the clustering process is restarted with a larger RMSD threshold, an empty list of cluster centers and the list of remaining candidate conformers. Conformers 1, 3, 4 and 8 are still far enough apart to become cluster centers but conformer 10 is now too similar to conformer 8 and removed. (k) Now all conformers have been successfully assigned to a cluster center and the ensemble size is equal to (or below) the maximum ensemble size. The final list of cluster centers is then reported as the conformer ensemble.

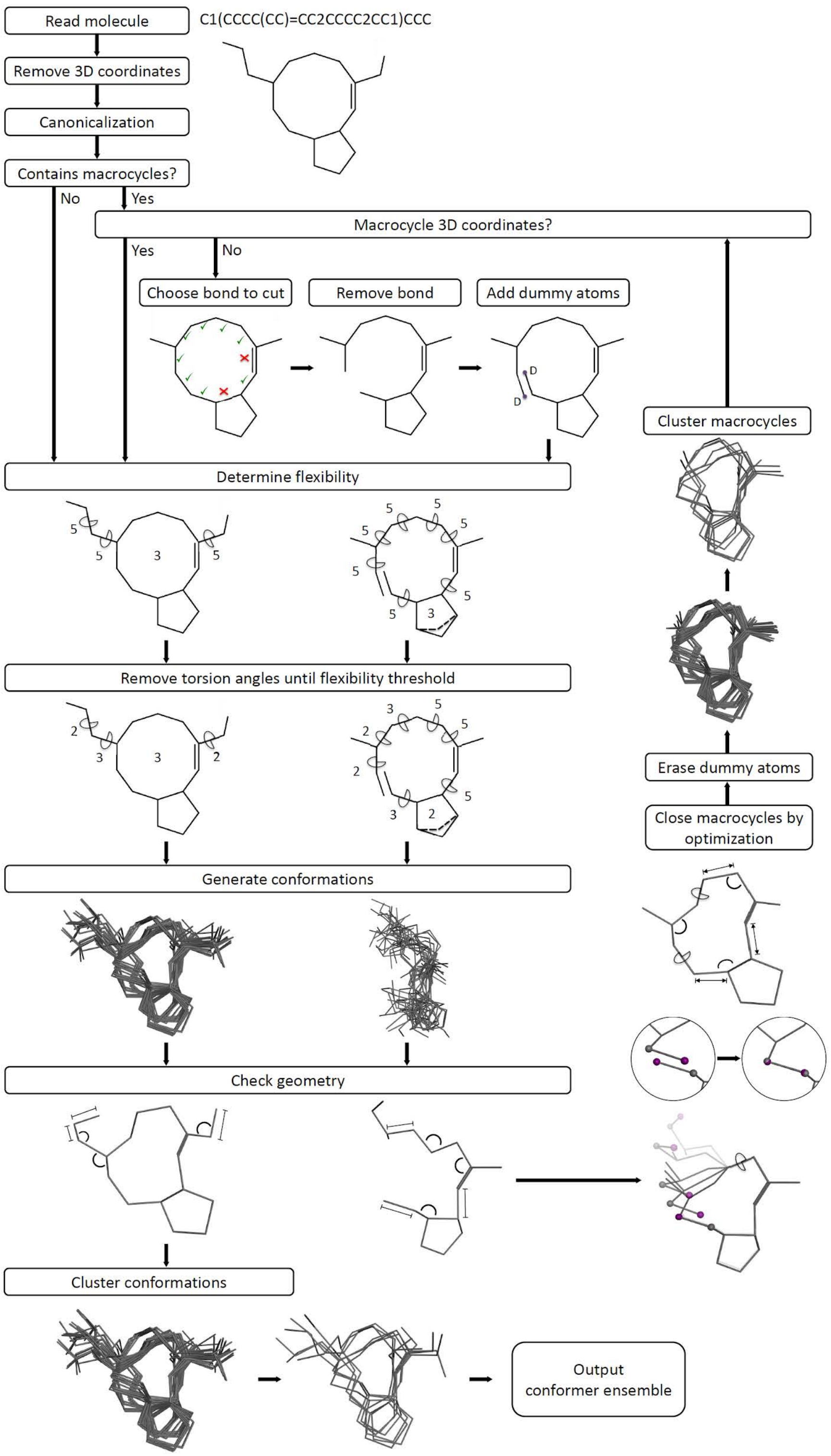
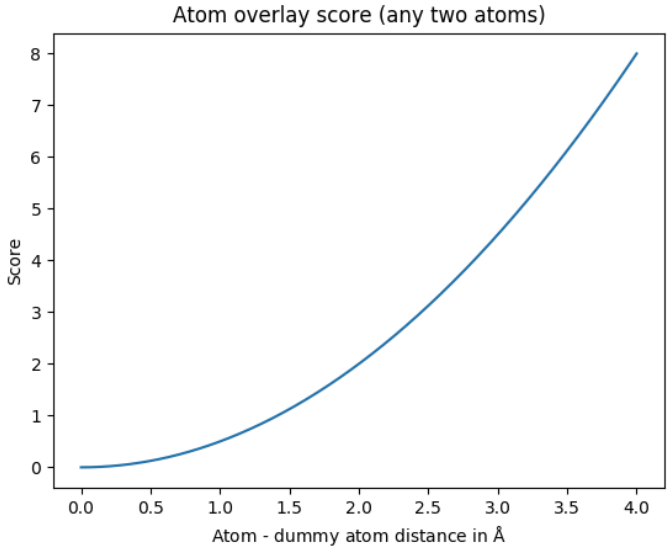
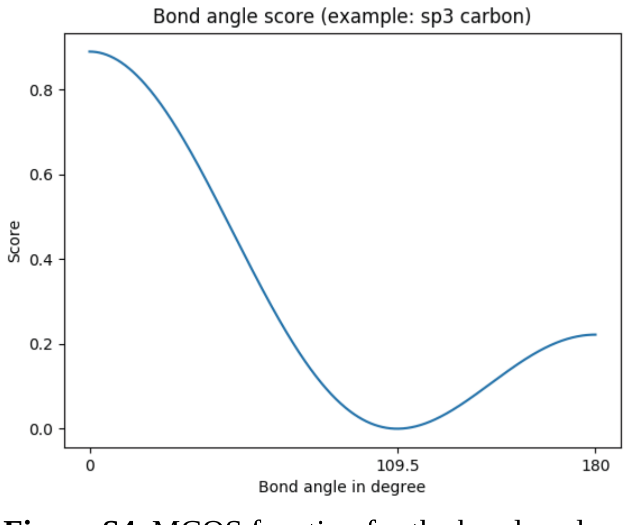


Figure S2. Schematic representation of Conformator's macrocycle conformer generation algorithm (for a detailed description see "Conformer Generation for Macrocycles" in the main text).



$$s_{overlay}(d) = \frac{1}{2}d^2$$

Figure S3. MCOS function for the overlay score for the distance d between the dummy atoms and the atoms in the original macrocycle they replaced. Ideally, this distance should be close to 0. It ensures that the bond angle and bond length across the cut bond will be restored during local optimization and also supports the preservation of local stereochemistry.



$$s_{angle}(\theta, \theta_0) = \frac{1}{2}(\cos\theta - \cos\theta_0)^2$$

Figure S4. MCOS function for the bond angle score. It uses a harmonic potential that is calculated on the cosine of the bond angle θ, to account for deviations from the ideal values θ_0. The bond angle score is calculated only for bond angles directly altered during optimization (i.e. angles that are optimization parameters) and the angles involving the cut bonds.

$$s_{limit}(\theta) = \begin{cases} \frac{1}{2}(\cos\theta - \cos 30°)^2 & \cos\theta > \cos 30° \\ \frac{1}{2}(\cos\theta - \cos 150°)^2 & \cos\theta < \cos 150° \\ 0 & \text{otherwise} \end{cases}$$

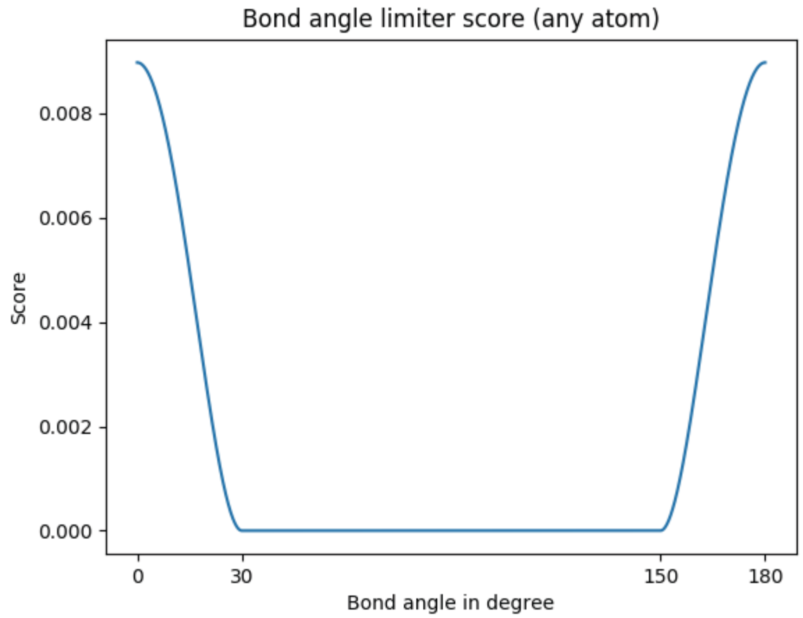
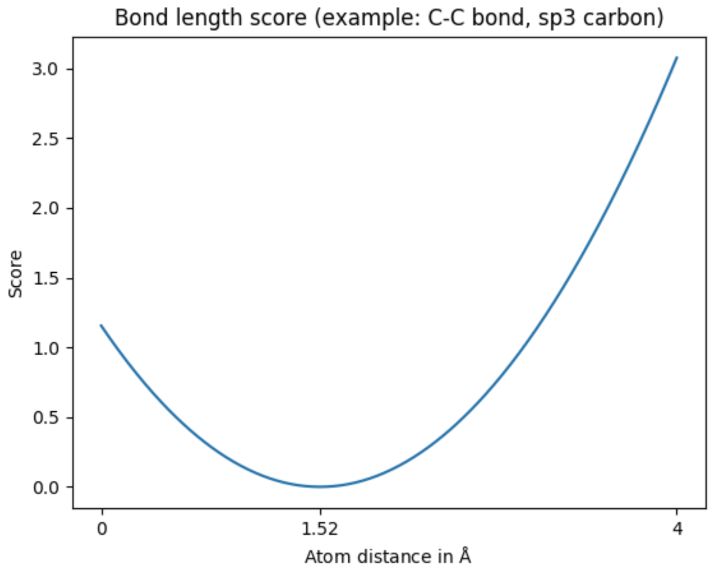


Figure S5. MCOS penalty function for limiting bond angles θ to guide the optimization of bond angles in macrocycles away from 0 and 180 degrees (if the atom does not have linear VSEPR geometry). It leads to a preference of bond angles between 30 and 150 degrees.



$$s_{bond}(d, d_0) = \frac{1}{2}(d - d_0)^2$$

Figure S6. The MCOS bond length term uses a harmonic potential to account for deviations of the bond length d from ideal values d_0. Only the bond lengths of the cut bonds are scored.

$$s_{distance\_factor}(d) = \begin{cases} 1 - g(d) & d \leq 0.5\text{Å} \\ 1 & \text{otherwise} \end{cases}$$

$$g(x) = \begin{cases} a_0 + a_1 \cdot x + a_2 \cdot x^2 + a_3 \cdot x^3 & 0 \leq x < 0.25 \\ b_0 + b_1 \cdot x + b_2 \cdot x^2 + b_3 \cdot x^3 & 0.25 \leq x < 0.375 \\ c_0 + c_1 \cdot x + c_2 \cdot x^2 + c_3 \cdot x^3 + c_4 \cdot x^4 & 0.375 \leq x < 0.5 \\ d_0 & 0.5 \leq x < \infty \end{cases}$$

	0	1	2	3	4
a	1	0	0	−6.5641	0
b	2.23077	−14.7692	59.0769	−85.3333	0
c	−100.923	960	−3337.85	5060.92	−2835.69
d	0	0	0	0	0

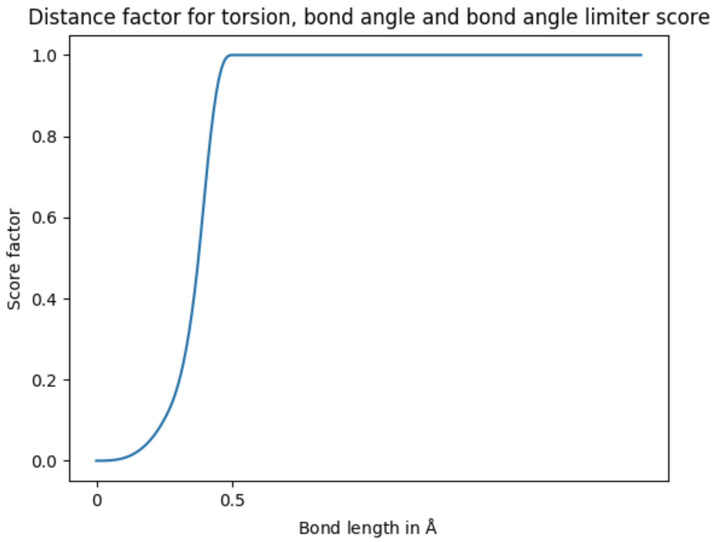


Figure S7. The MCOS distance factor by which the torsion angle potential, the bond angle potential and the bond angle limiter score are multiplied to reduce the respective score to 0 in cases where any bond length is close to 0 Å. This is necessary to ensure the continuity of the score contributions that depend on torsion angles or bond angles. It is a piecewise polynomial approximation to a plateau function that is twice continuously differentiable. The function $g(x)$ was modeled by fixing function and derivative values at defined points and solving the system of equations for the coefficients of the polynomials (the coefficients are shown in the table).

$$s_{angle\_factor}(\theta) = \begin{cases} 1 - g(1-\cos\theta) & \theta \leq 20° \\ 1 & 20° < \theta < 160° \\ 1 - g(\cos\theta + 1) & \theta \geq 160° \end{cases}$$

$$g(x) = \begin{cases} a_0 + a_1 \cdot x + a_2 \cdot x^2 + a_3 \cdot x^3 & 0 \leq x < 0.0151922 \\ b_0 + b_1 \cdot x + b_2 \cdot x^2 + b_3 \cdot x^3 & 0.0151922 \leq x < 0.0377498 \\ c_0 + c_1 \cdot x + c_2 \cdot x^2 + c_3 \cdot x^3 + c_4 \cdot x^4 & 0.0377498 \leq x < 0.0603074 \\ d_0 & 0.0603074 \leq x < \infty \end{cases}$$

	0	1	2	3	4
a	1	0	0	-10046.7	0
b	0.987639	2.44088	-160.666	-6521.48	0
c	-11.3122	1115.49	-36828.1	507524	$-2.52014 \cdot 10^6$
d	0	0	0	0	0

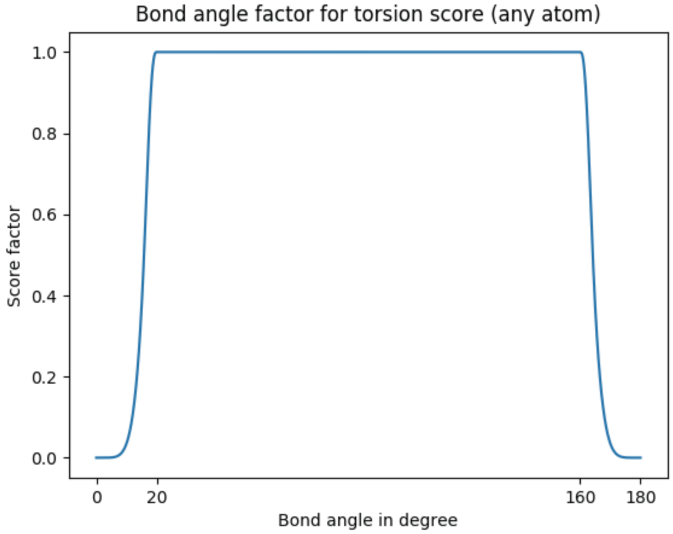


Figure S8. The MCOS bond angle factor by which the torsion angle potential is multiplied to reduce the torsion angle score to 0 in cases where any bond angle θ along that torsion bond is either close to 0 or 180 degrees. This is necessary because the torsion angle, as a function of the four atom coordinates, has a discontinuity when three consecutive atoms are collinear. It is a piecewise polynomial approximation to a plateau function that is twice continuously differentiable. The function $g(x)$ was modeled by fixing function and derivative values at defined points and solving the system of equations for the coefficients of the polynomials (the coefficients are shown in the table).

$$s_{clash}(d, d_{vdw}) = \begin{cases} \frac{1}{2}(d - 0.7 \cdot d_{vdw})^2 & d < 0.7 \cdot d_{vdw} \\ 0 & \text{otherwise} \end{cases}$$

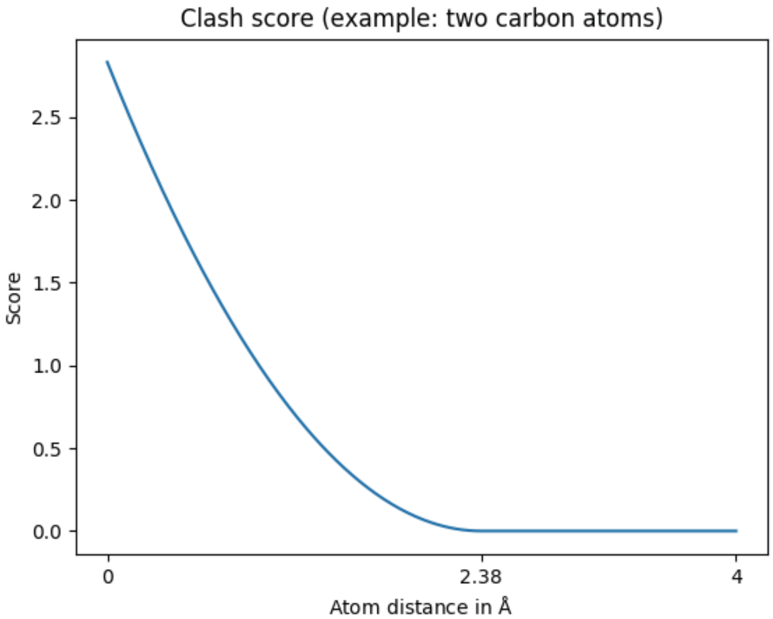


Figure S9. The MCOS clash score prevents intramolecular clashes. It is a quadratic function depending of the atomic distance d and the sum of the van der Waals radii of the atoms d_{vdw} and penalizes van der Waals overlaps between 1-4-connected (or further away) heavy atoms that exceed the threshold level of 30%.

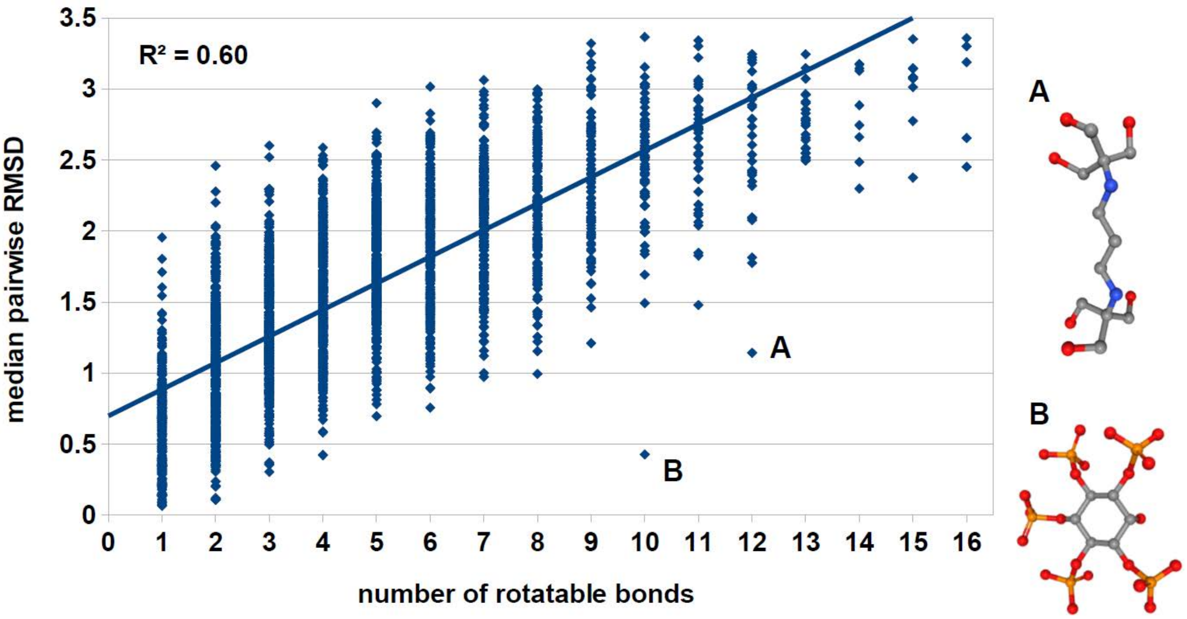


Figure S10. Median pairwise RMSD of all-against-all comparisons for each conformer ensemble generated for the Platinum Diverse Dataset with Conformator (default settings) plotted versus the number of rotatable bonds. The two labeled outliers are the highly symmetrical ligands *B3P* (A) and *5MY* (B). The R^2 for the correlation between median pairwise RMSD of all conformers and the number of rotatable bonds was 0.60.

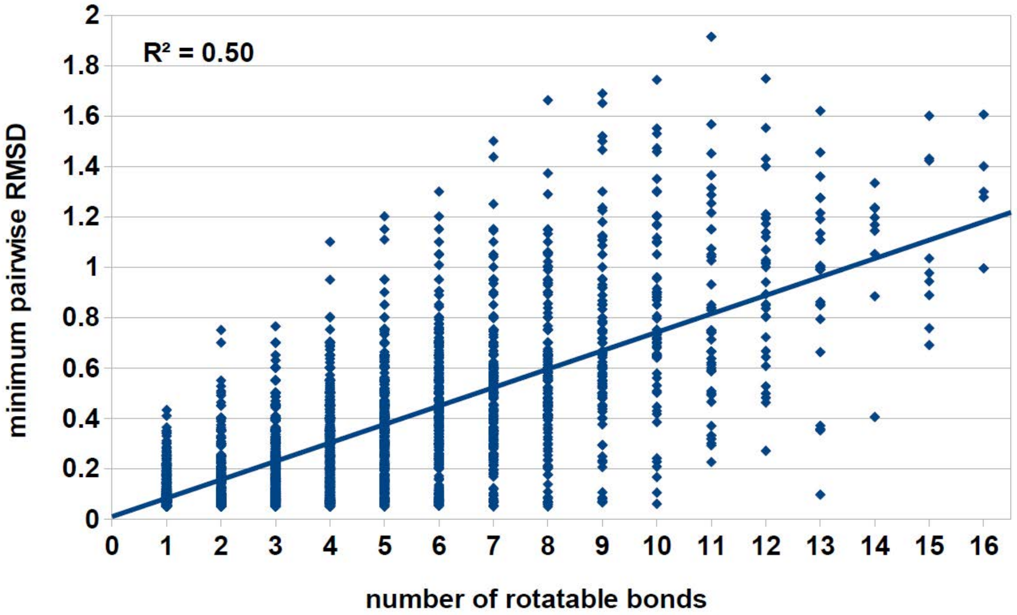


Figure S11. Minimum pairwise RMSD of all-against-all comparisons for each ensemble generated for the Platinum Diverse Dataset with Conformator (default settings) plotted versus the number of rotatable bonds. The R^2 for the correlation between median pairwise RMSD of all conformers and the number of rotatable bonds was 0.50.

Eidesstattliche Versicherung

Hiermit erkläre ich an Eides statt, dass ich die vorliegende Dissertationsschrift selbst verfasst und keine anderen als die angegebenen Quellen und Hilfsmittel benutzt habe.

Neumünster, den 21.06.2020

Nils-Ole Friedrich