Conformer Generation for Structure Based Drug Design: How Many and How Good?

Andrew T. McNutt,[†] Fatimah Bisiriyu,[‡] Sophia Song,[¶] Ananya Vyas,[§] Geoffrey R. Hutchison,[*,‖,⊥] and David Ryan Koes[*,#]

[†]*Department of Computational and Systems Biology, University of Pittsburgh, Pittsburgh, PA*

[‡]*The Neighborhood Academy, Pittsburgh, PA, 15206*

[¶]*Upper St. Clair High School, Pittsburgh, PA, 15241*

[§]*Taylor Allderdice High School, Pittsburgh, PA, 15217*

[‖]*Department of Chemistry, University of Pittsburgh, Pittsburgh PA, 15213*

[⊥]*Department of Chemical and Petroleum Engineering, University of Pittsburgh, Pittsburgh, PA, 15213*

[#]*Department of Computational and Systems Biology, University of Pittsburgh, Pittsburgh, PA, 15213*

E-mail: geoffh@pitt.edu; dkoes@pitt.edu

Abstract

Conformer generation, the assignment of realistic 3D coordinates to a small molecule, is fundamental to structure based drug design. Conformational ensembles are required for rigid-body matching algorithms, such as shape-based or pharmacophore approaches, and even methods that treat the ligand flexibly, such as docking, are dependent on the quality of the provided conformations due to not sampling all degrees of freedom (e.g. only sampling torsions). Here we empirically elucidate some general principles about

the size, diversity and quality of conformational ensembles needed to get the best performance in common structure-based drug discovery tasks. In many cases our findings may parallel "common knowledge" well-known to practitioners of the field. Nonetheless, we feel it is valuable to quantify these conformational effects while reproducing and expanding upon previous studies. Specifically, we investigate the performance of a state-of-the-art generative deep learning approach versus a more classical geometry based approach, the effect of energy minimization as a post-processing step, the effect of ensemble size (maximum number of conformers), and construction (filtering by RMSD for diversity) and how these choices influence the ability to recapitulate bioactive conformations and perform pharmacophore screening and molecular docking.

Introduction

Generating a three-dimensional conformation of a molecule from its topological representation (e.g., a SMILES string) is a fundamental first step of most structure-based approaches to drug discovery.[1,2] Tasks such as 3D pharmacophore search,[3–5] molecular docking,[6–9] and 3D QSAR[10] all rely on the generation of a biochemically meaningful conformation. Traditionally, conformer generation algorithms have adopted either a systematic or stochastic approach. Systematic approaches attempt to enumerate all reasonable values for rotatable bonds, and thus often have difficulty scaling. Stochastic approaches use random sampling to make the search process more scalable. Distance geometry[11,12] uses bond length, angle and other, possibly knowledge-based,[13] constraints to constrain the stochastic search space. More recently, machine learning has been used to either generate conformations directly[14–16] or otherwise assist in the generation process (e.g., torsional sampling).[17–25] Although there are multiple ways to evaluate the quality of a conformer generator,[1] most relevant for structure-based drug discovery is the ability to produce a bioactive conformation. That is, a conformation close to the conformation found in a protein-ligand complex should be generated, even if it is not the lowest energy conformation. Both free[26] and commercial[27] conformer generators

were evaluated for this task and, provided that a sufficiently large ensemble is generated, most approaches succeed at identifying a low RMSD (< 2Å) conformation. In particular, the open source RDKit, which uses a stochastic distance geometry based approach combined with experimental torsional-angle and ring geometry preferences (ETKDG),[13] consistently performs as well as or better than other approaches such as Balloon, Confab, Frog2, Multiconf-DOCK, CREST, ConfGen, OMEGA, MOE and others (see Friedrich et al.[26], Friedrich et al.[27], and Folmsbee et al.[28]), hence we limit our evaluation to RDKit as a representative of a conventional conformer generator. We note the recently described Auto3D[29] uses RDKit conformers as a starting point for optimizing with a modified version of the ANI-2x deep learning molecular potential,[30] but this does not result in better performance than RDKit in the bioactive conformation identification task (see Figure S1).

The latest machine learning models have not been evaluated for their ability to generate bioactive conformations. Instead, they are mostly trained and evaluated on the GEOM[31] dataset, which contains 37 million conformers of more than 450,000 molecules with the goal of accurately representing, at the level of semi-empirical density functional theory,[32] the vacuum conformer-rotamer ensembles of these molecules using CREST.[33] Deep generative models significantly outperform RDKit at this particular task, but an extended sampling and clustering approach using RDKit achieves highly competitive performance.[34] It is not clear that it is a fair comparison to compare methods that utilize different amounts of sampling,[35] so here we evaluate RDKit and a deep generative model using identical sampling and ensemble formation criteria. As the Direct Molecular Conformation Generation (DMCG)[14] was found to perform best at the task of reconstituting the ensembles of the GEOM-Drugs subset of GEOM, we evaluate it here at the task of bioactive conformation recovery. DMCG is an end-to-end generative model with a variational encoder/decoder architecture that learns all network parameters from the training data distribution, GEOM-Drugs.[31] However, our main goal is not to extend previous evaluations[26,27,34] but to explore the impact of various choices made in the conformer generation process, such as the size of the ensemble, the criteria for

including conformers in the ensemble, and the use (or not) of energy minimization, has on the ultimate endpoint of the common structure-based tasks of pharmacophore search and molecular docking.

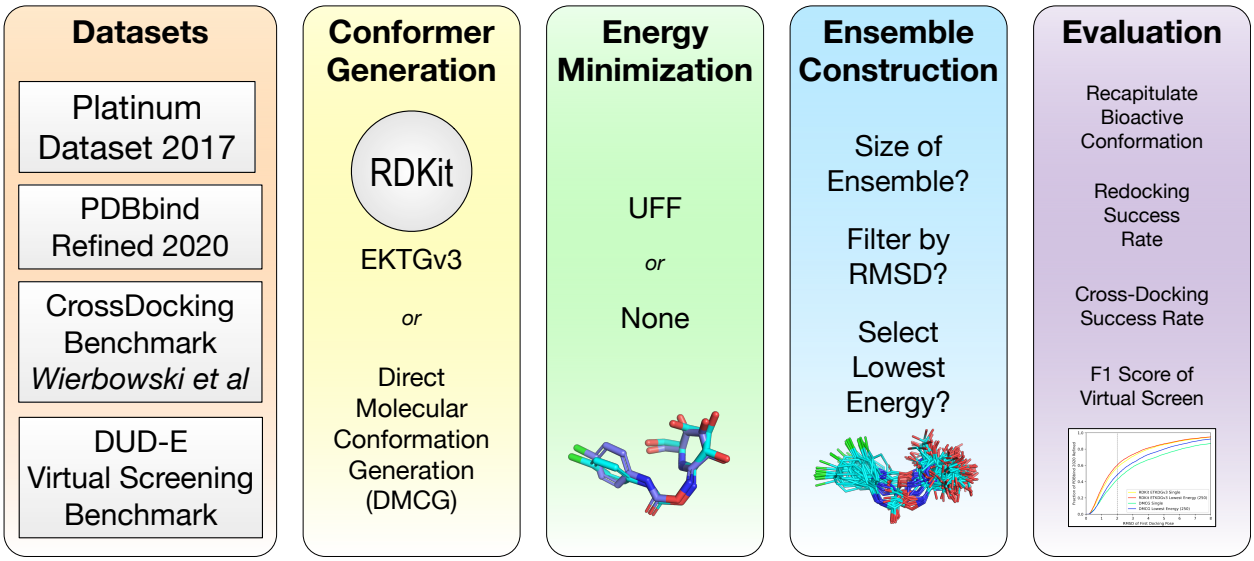

Figure 1: Overall workflow for conformer ensemble generation and evaluation in structure-based drug discovery tasks.

Methods

The overall workflow of our evaluation is shown in Figure 1. We evaluate a number of options for generating conformational ensembles from common datasets and evaluate them in two common structure-based tasks: pharmacophore search and molecular docking.

Datasets

In order to evaluate the ability of a conformer generator to produce a bioactive conformation, we use two datasets: Platinum 2017[27] and the refined subset of PDBBind 2020.[36] The PDBBind refined set curates high quality protein-ligand structures from the Protein Data Bank with known binding affinities. Of the 5316 ligand structures in this set, 5313 could be processed by RDKit and are considered here (the remaining three all had a molecular

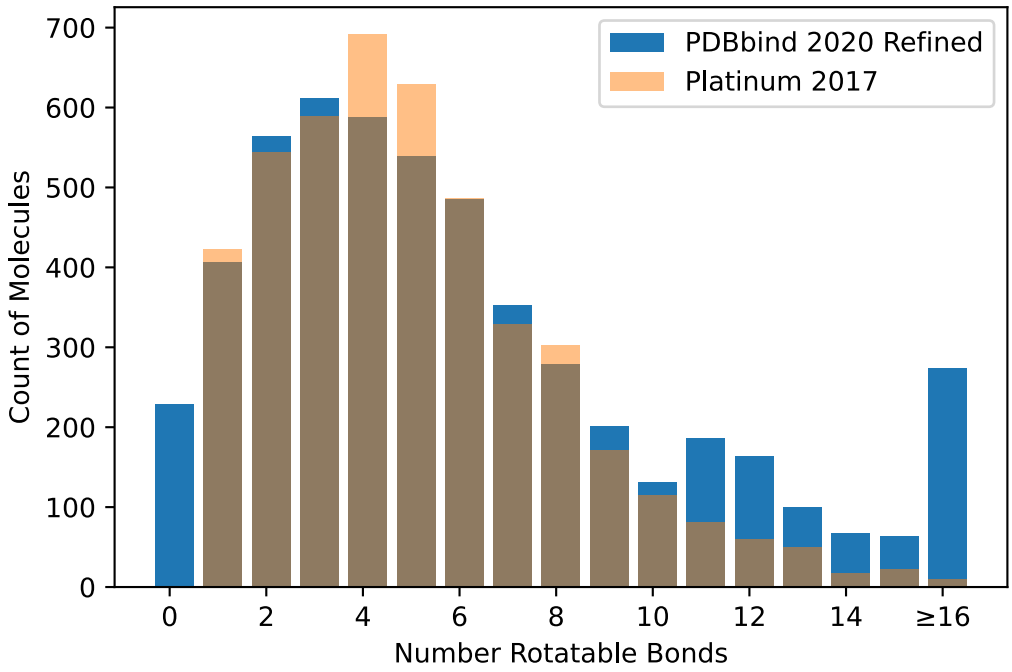

Figure 2: Histogram of rotatable bonds within our datasets.

weight of more than 900 Da). The Platinum dataset was designed for conformer generation evaluations and, in addition to considering the overall quality of a structure, evaluates the quality of the fit of a ligand structure to the electron density map, ensuring that the included conformations are accurate. It also imposes more stringent filtering, such as not considering molecules with more than 16 rotatable bonds or fewer than one (see Figure 2). While the Platinum dataset provides a high-quality ground truth, the PDBbind dataset contains more challenging (i.e., flexible) ligands.

For assessing pharmacophore virtual screening performance, we use the Database of Useful Decoys: Enhanced (DUDE),[37] which contains 102 protein targets each with their own set of experimentally confirmed actives and (possibly putative) decoys. We note that while there are significant issues in using DUDE for training and evaluating machine learning approaches[38,39] due to inherent biases in the dataset construction that can result in misleading evaluations of generalization performance, the structure-based pharmacophore search we evaluate does not have these drawbacks as it is not fitting to the data. We note that the

use of decoy molecules that are not experimentally validated in DUDE may result in false negatives, but this is not material for evaluating trends in virtual screening performance, which is our goal here. We also note that the actives in DUDE are experimentally validated against their target, which is essential when evaluating target-focused approaches and strongly preferable to using benchmarks, such as LIT-PCBA,[40] where actives have uncertain mechanisms of action due to being identified in phenotypic screens.

For assessing molecular docking performance, we use the refined set of PDBBind 2020 for re-docking evaluations and for cross-docking the dataset of Wierbowski et al.[41]. Re-docking evaluates docking a ligand to its cognate receptor while cross-docking docks a ligand to a similar, but not cognate, receptor. As the number of protein-ligand pairs in a cross-docking task grows combinatorially and we do not want to disproportionately weight targets with more structures, we randomly downsample Wierbowski et al.[41] to have at most 100 protein-ligand pairs for each of its 92 targets (63% of the targets require downsampling). Docking success is measured by calculating the root mean squared deviation (RMSD) between the top ranked docked pose and the crystal pose. In the case of cross-docking, the reference pose is determined by aligning the protein structure of the cognate receptor to the target receptor.

Conformer Ensemble Generation

We generate RDKit conformers using the ETKDG version 3[12,13] method of RDKit using default values and version 2022.03.1. This method combines distance geometry[11] sampling with knowledge based potentials to increase the efficiency of the algorithm without loss of accuracy (see Friedrich et al.[26] and Figure S2). To generate DMCG[14] conformers we use the pre-trained model `Large_Drugs/checkpoint_94.pt` with the recommended settings for drug-like molecules. We strip the input SMILES strings of stereochemistry information, as this information is often missing in virtual screening datasets and we want to evaluate the ability of conformer generators to sample appropriate geometries (Figure S3 shows the

relatively small contribution of stereochemistry in our evaluations). For both generators we evaluate further refining generated conformations using the UFF molecular force field[42] as implemented by RDKit and default convergence criteria.

Consistent with previous evaluations,[26,27] we generate ensembles with a maximum of 250 conformers. We consider different methods for sub-setting the full ensemble, including unbiased sampling, energy ranking, and energy ranking with RMSD filtering.

Pharmacophore Search

Pharmit[43] is used to perform pharmacophore search on the DUDE benchmark. Search databases are built from the relevant conformational ensembles of the active and decoy compounds of each DUDE target. Due to the large number of conformers required, only RDKit conformer generation was evaluated for this task. The provided reference crystal structure is used to elucidate all possible interacting features (hydrogen bonds, hydrophobic interactions, charge interactions, and aromatic interactions) between the ligand and receptor. Interactions are identified using the built-in heuristics of Pharmit. From this set of interactions, all possible pharmacophores with at least three features are enumerated. As our goal is to evaluate the effect of different conformer ensembles on virtual screening and not elucidating the best single pharmacophore query, we screen all the enumerated queries. We set a tolerance radius of 1.0Å and no other constraints (e.g. direction) on each feature. Since Pharmit uses the sub-linear time Pharmer[44] algorithm, despite the combinatorial number of queries and many thousands of compounds, this can be done efficiently. We emphasize that this algorithm finds matches between the specified query and rigid conformers and so the quality of the ensemble is essential. As only matching, not ranking, is performed, classification metrics are the most appropriate choice to evaluate virtual screening performance in this context (i.e., without a ranking it is not possible to calculate a meaningful AUC of a ROC or precision-recall curve). We use the F1 score, the harmonic mean of the precision and recall, and report the best F1 across all queries. Unlike an enrichment factor, the F1 score is a normalized quantity

(ranges from zero to one) and so can be sensibly compared across different screens, and it encapsulates the goal of virtual screening - to maximum the number of true positives (recall) while minimizing the number of false positives (higher precision).

Molecular Docking

GNINA[7] is used to perform molecular docking. GNINA is a fork of AutoDock Vina[45] that uses a convolutional neural network protein-ligand scoring function[46] to select and ranking poses. Independent evaluations[47,48] of GNINA have found it to have comparable performance to the commercial Glide software[49] while outperforming other open source docking programs such as smina.[50] Poses are sampled using a Monte Carlo Metropolis algorithm that perturbs the rigid body degrees of freedom (translation and rotation) and torsional degrees of freedom. The output docked poses therefore depend on the input conformation to determine bond lengths and angles. However, internal torsions are completely randomized at the start of each Monte Carlo chain, so the result does not depend on the input torsions.

To assess the impact of the input conformation on docking results, we consider single conformer and five conformer ensembles (larger ensembles were not considered due to the computational overhead of docking).

Results

Retrieval of Bioactive Conformers

To compare generated conformers to the experimental crystal structure we use `obrms` from the Open Babel toolkit,[51] which properly handles internal symmetries by reporting the lowest possible root mean squared deviation (RMSD) of any valid atom matching. In all cases, the minimized RMSD (`-m` option) is reported (i.e., the structures are optimally aligned before calculating the RMSD). For a variety of ensemble sizes (number of samples of the specified method) we evaluate the fraction of the dataset where a conformer exists within

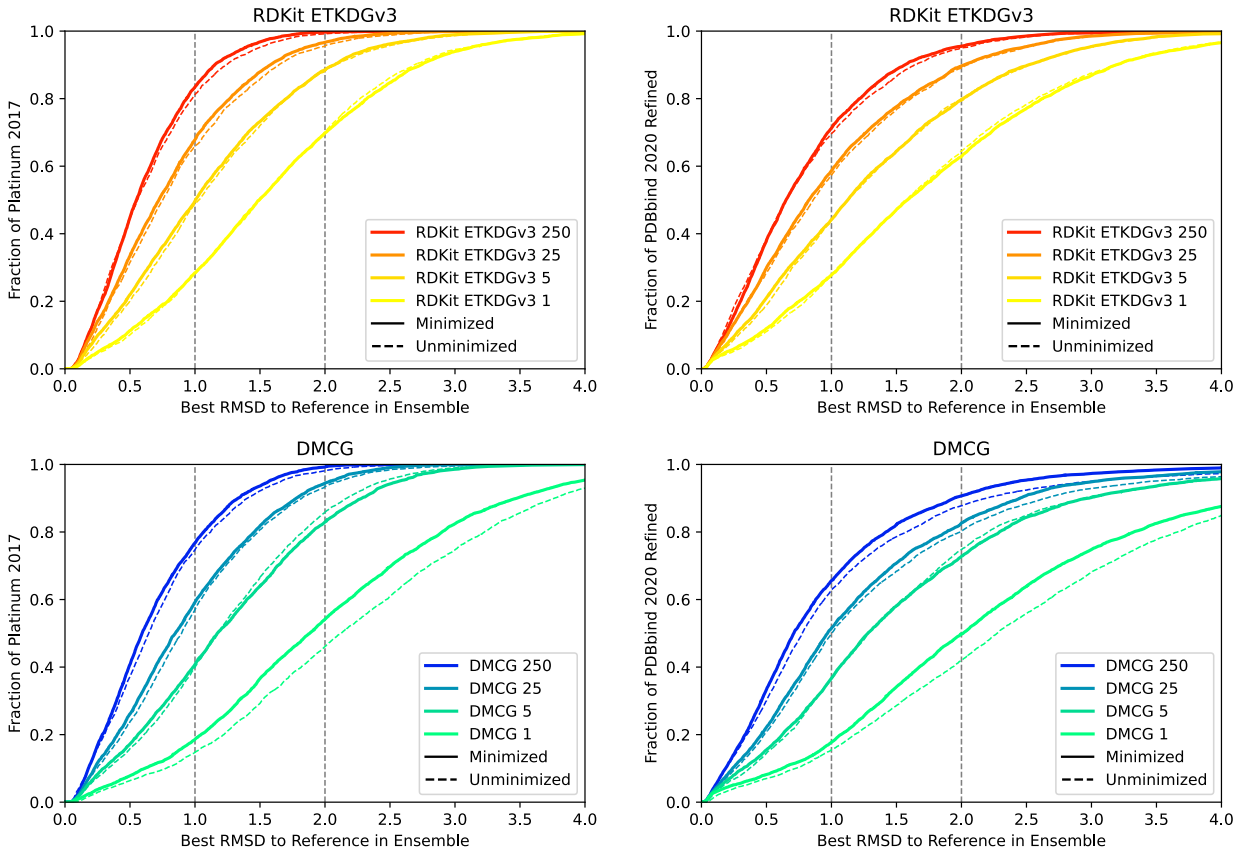

Figure 3: Fraction of the Platinum (left) and PDBbind2020 Refined (right) data sets where a conformer within a specified threshold (x-axis) of the experimental structure is retrieved for both RDKit (top) and DMCG (bottom) for various sized ensembles. Results for poses before (dashed line) and after (solid line) UFF minimization are shown.

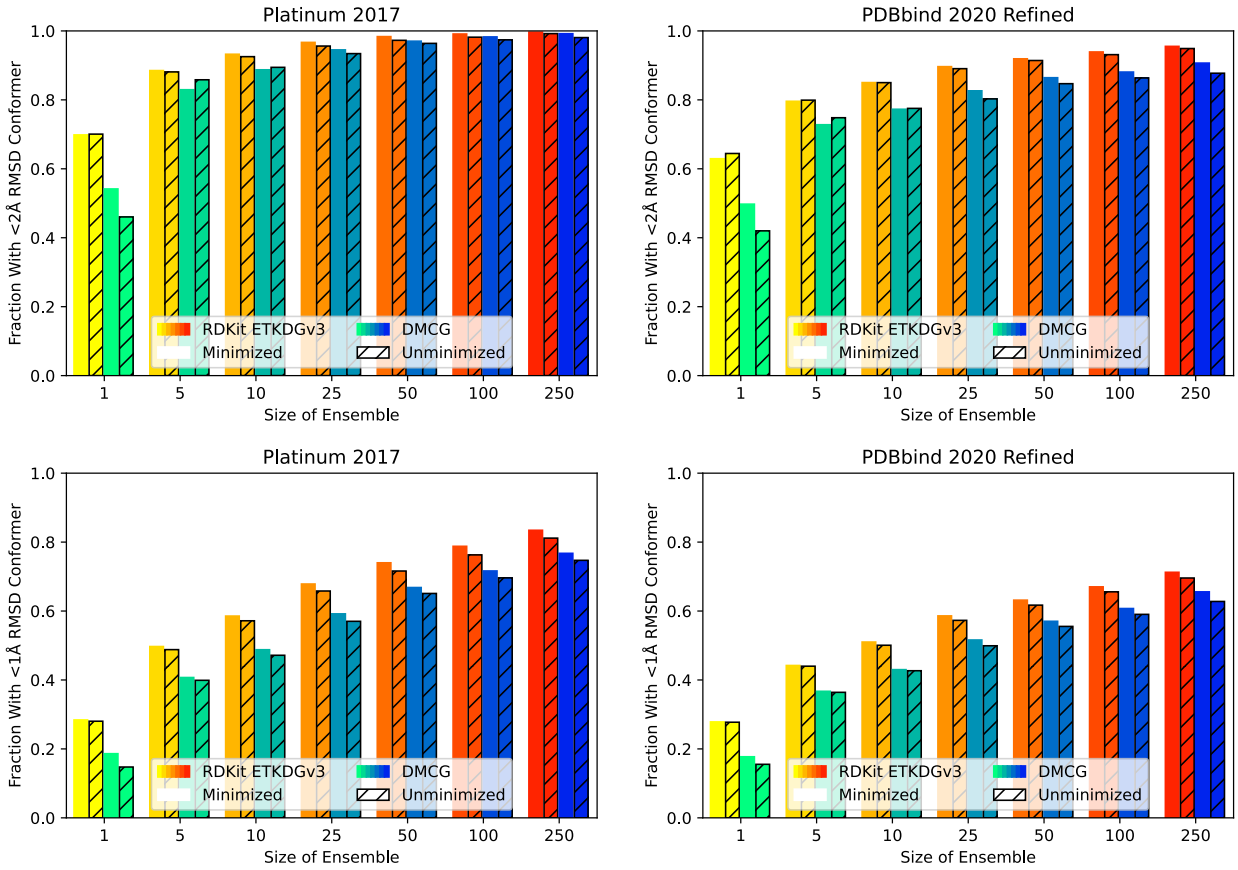

Figure 4: Fraction of the Platinum (left) and PDBbind2020 Refined (right) data sets where a conformer within 2.0Å (top) or 1.0Å (bottom) RMSD of the experimental structure is retrieved for both RDKit and DMCG for various sized ensembles. Results for poses before (hashed bars) and after (solid bars) UFF minimization are shown.

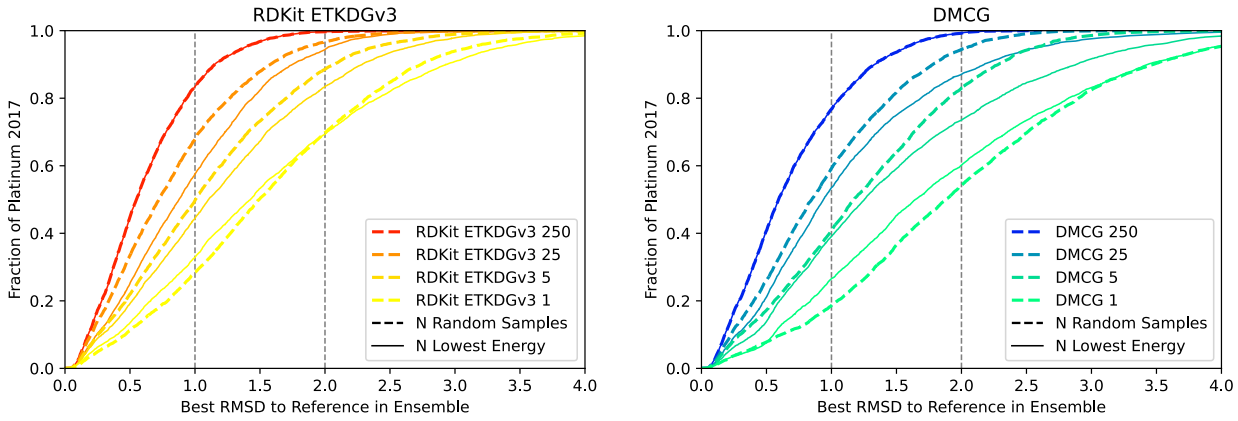

Figure 5: Comparison of RDKit ETKDGv3 (left) and DMCG (right) on the Platinum 2017 dataset when evaluated on different ensemble sizes where the ensemble is constructed by selecting the N lowest energy minimized conformers are selected from an ensemble of 250. Results for the PDBbind Refined dataset are shown in Figure S6 and exhibit a similar trend.

the ensemble for a specified RMSD threshold. That is, we consider the best possible RMSD across the ensemble. Results for a variable threshold are shown in Figure 3 with more ensemble sizes shown for two fixed thresholds, 1.0Å and 2.0Å in Figure 4. For reference, example structures at different RMSD values are shown in Figure S20. In general, we find that RDKit consistently matches or outperforms DMCG at conformer retrieval at every RMSD threshold (a more direct visual comparison is found in Figure S4). This advantage is greater on the PDBbind dataset as RDKit better handles larger, more flexible ligands (Figure S5). The larger, more flexible ligands in the PDBbind dataset result in consistently lower retrieval rates for both methods, but the trends between the two datasets are consistent. Energy minimization has a small, not always beneficial, effect that is more pronounced and generally beneficial for DMCG. Unless otherwise specified, we limit ourselves to evaluating minimized conformers for the remainder of our analysis.

Generating a larger ensemble will monotonically increase the likelihood of retrieving a bioactive conformation, but for efficiency reasons it is desirable to generate smaller ensembles. As shown in Figures 5 and S6, selecting the lowest energy poses from a larger ensemble does not improve the retrieval of bioactive conformations with the exception of reducing down to a single conformer. This is due to the lack of geometric diversity of subsets chosen using only

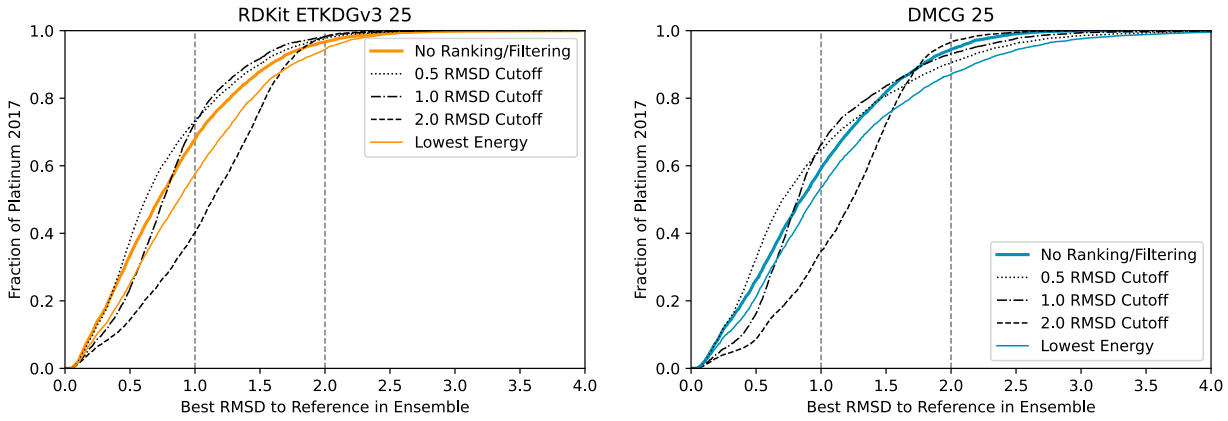

Figure 6: Evaluation of different methods of constructing an ensemble of 25 conformers selected from a 250 conformer ensemble generated using RDKit ETKGDv3 (left) or DMCG (right) on the Platinum 2017 dataset. PDBbind Refined results are shown in Figure S8.

energy as a criteria, as well as the energy evaluation reflecting an isolated ligand, neglecting non-bonded dispersion interactions with the surrounding protein. In Figures 6, S7, S8, and S9 we show the effect of imposing an RMSD cutoff when selecting conformers. In this case, the conformers of the full 250 conformer ensemble are sorted by increasing energy and we greedily add conformers to the selected subset only if their RMSD to every already selected conformer is greater than an RMSD threshold. This approach results in an improved retrieval rate relative to unbiased sampling or lowest energy selection when the RMSD threshold used to select conformers is similar to the RMSD cutoff used to classify a conformer as matching the experimental structure. For example, selecting 25 RDKit conformers for Platinum 2017 using an RMSD of 1.0 results in ensembles that contain a conformer within 1.0 RMSD of the true conformer 72.7% of the time, compared to 67.8% when unbiased sampling is used and 57.6% when the 25 lowest energy conformations are selected. However, if the RMSD criteria for determining what qualifies as a matching conformation deviates significantly from the RMSD threshold used to select the subset (either lower or higher), unbiased sampling can outperform the filtered subsets. Finally, we note that in our analysis the bioactive conformation depends on the structure of a receptor, which is hidden information from the conformer generator, but we observe similar trends in retrieval rates when we evaluate using

a curated subset[28] of the Crystallography Open Database[52] (see Figure S10) which contains a broader array of single molecule experimental structures.

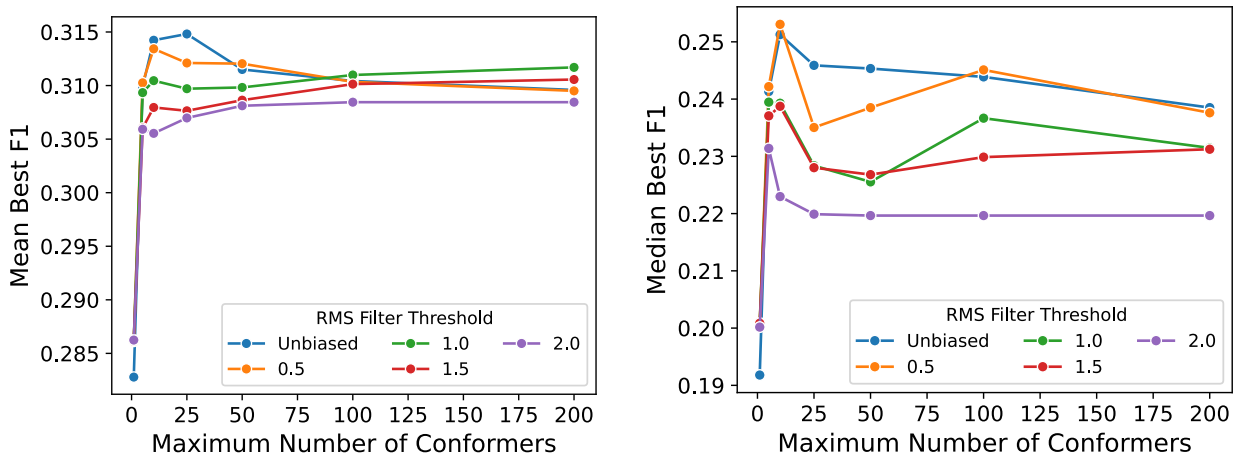

Figure 7: The average and median best F1 score achieved from pharmacophore search across the 102 DUDE targets as the maximum number of conformers allowed in the library is varied.

.

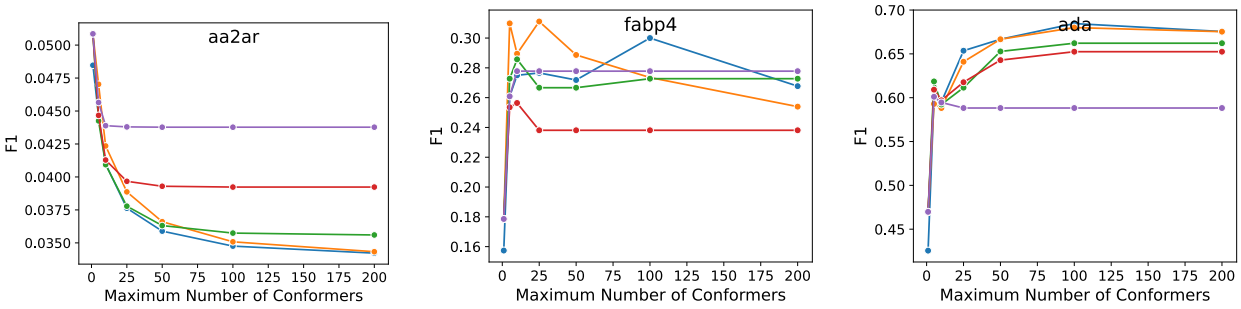

Figure 8: The best F1 score achieved from pharmacophore as the maximum number of conformers allowed in the library is varied for three distinctly different targets. Note the y-axis scales differ to better illustrate the trends. Individual F1 plots are shown for all targets in Figures S12 and S13.

Conformer Ensemble Effect on Pharmacophore Search

Although retrieval of bioactive conformations within a conformational ensemble is clearly desirable, it is not clear such an analysis is sufficient to determine the best approach for constructing conformational ensembles for structure-based tasks. To more directly address this question, we consider rigid pharmacophore matching against differently sized conformer

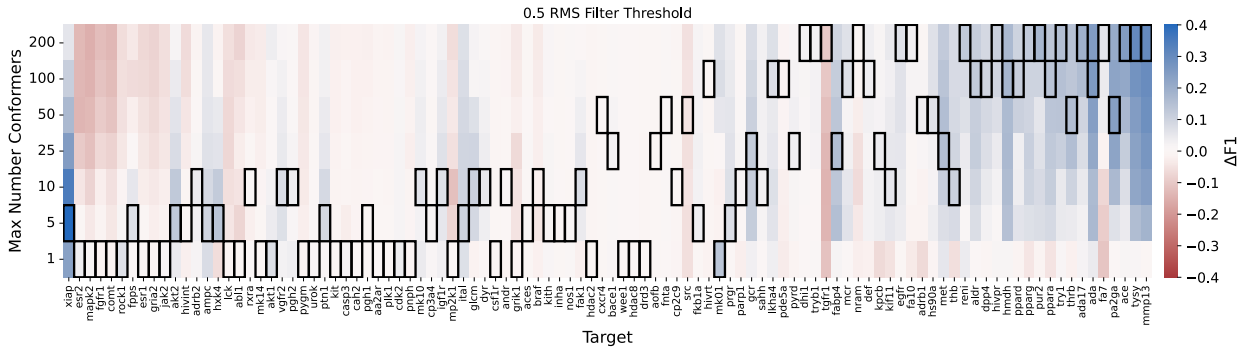

Figure 9: For each DUDE target, the difference in best achieved F1 score relative to the best F1 achieved by unbiased sampling of a single conformation is shown for different numbers of maximum allowed conformations. Conformers are selected by sorting by energy and then filtered by an RMS threshold of 0.5Å. Targets are sorted by the slope of the best fit line through the conformer/F1 data. Box outlines highlight the choice of maximum number of conformers that provides the highest F1 score.

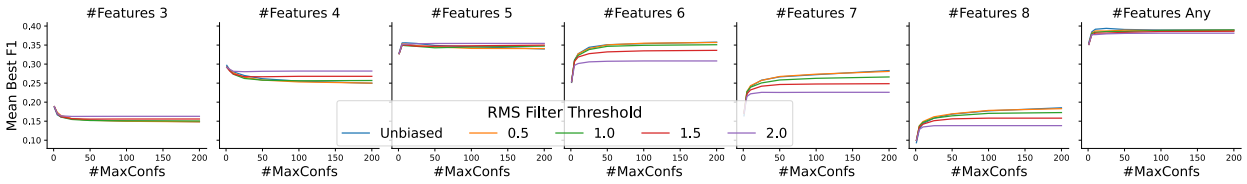

Figure 10: The best F1 score achieved from pharmacophores with a specified number of features for different choices of conformer ensembles. Only a subset of 63 targets is evaluated as the remaining targets have fewer than 8 interaction features to select from. For reference, the mean F1 when the number of features is not fixed ("Any") is shown.

ensembles. We consider generating conformers with RDKit and then unbiased sampling from energy minimized poses as well as sorting by energy and then filtering by a specified root mean squared deviation (RMSD) threshold, as this approach was found to be most effective at retrieving bioactive conformations. This threshold specifies the minimum distance between any two conformers in the ensemble.

The overall average effect on F1 score as the RMSD threshold and maximum number of conformers in a generated ensemble are varied is shown in Figure 7. For smaller ensembles (e.g. < 10 conformers), using lower energy poses filtered by RMSD provides the best average performance. For moderately sized ensembles, a larger RMS threshold reduces performance. Larger RMS thresholds result in significantly smaller ensembles (e.g., filtering at a 2Å threshold results in a reduction from 250 conformers to an average of only 6 conformers per a molecule - see Figure S11), hence larger amounts of filtering result in reduced performance and are relatively insensitive to increasing the maximum allowed number of conformers. While increasing the number of conformers can only increase the recall of known actives, it can also reduce the precision (i.e., increase the number of false positives due to more inactive compounds matching the pharmacophore). This leads to a reduction in average performance as the maximum size of the ensemble is increased with minimal filtering.

The average effect size shown in Figure 7 is small, however, as shown in Figure 8 and 9, the average trends hide a wide array of responses to changes in conformational ensemble size and the effect of ensemble size can be significant and varied. For the 0.5Å RMS filtered set there are 29 of the 102 targets where the best F1 score is achieved using a single conformer compared to 16 where the best F1 score is achieved using an ensemble of 200 conformers. For the majority of targets (68), the best F1 score requires 25 or fewer conformers and for cases where more conformers are preferred, the improvement over smaller ensembles is often minimal.

For the previous analysis we consider only the best performing (by F1 score) pharma-

cophore query, essentially assuming the pharmacophore query was designed by an omniscient oracle, in order to separate the issue of pharmacophore elucidation from the effect of the choice of conformer ensemble. However, it is instructive to consider the change in trends if the oracle is restricted to pharmacophore queries with a fixed number of features, as shown in Figure 10. As the number of features is increased, the specificity of the query increases and the number of matches decreases. For low specificity queries, small ensembles maximize the F1 as they counter-balance the lack of specificity while conversely high specificity queries benefit from large ensembles. In order to achieve the best F1 performance, a balance of both query specificity and ensemble size is needed.

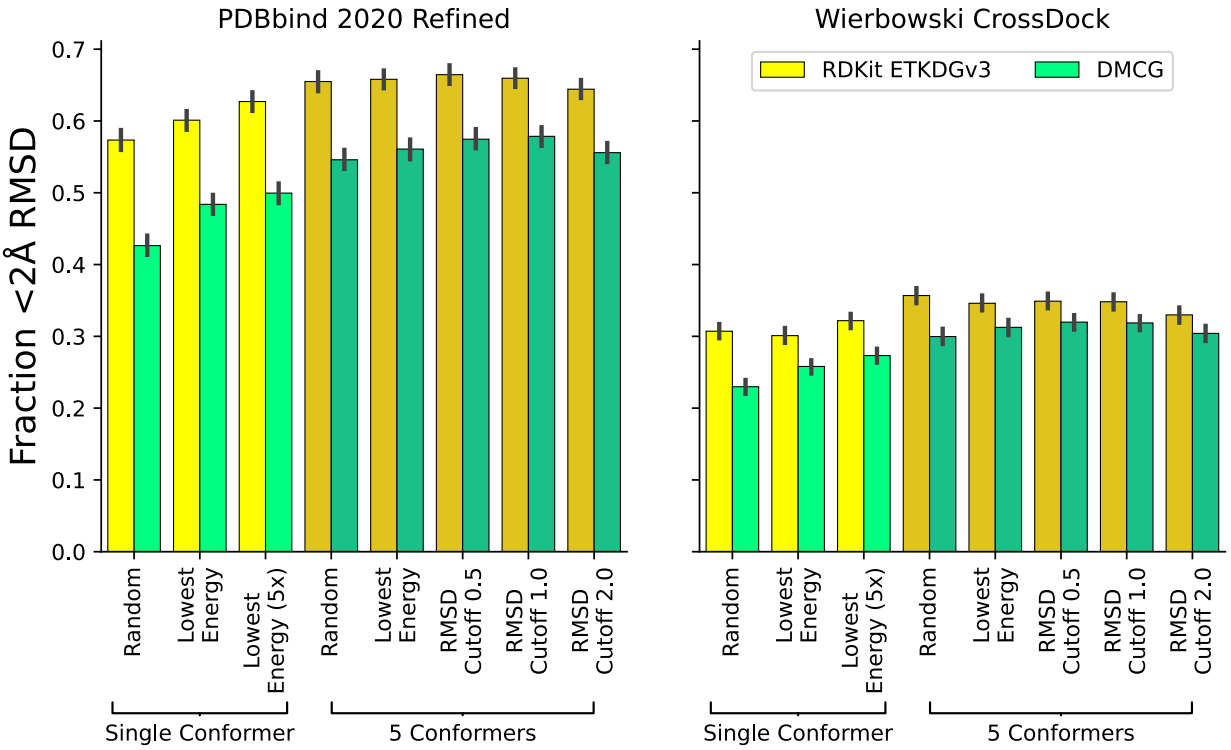

Figure 11: The effect of using different input conformer ensembles on docking performance as measured by the fraction of systems where a low (<2Å) RMSD pose is identified as the top ranked docked pose. Error bars indicate the 95% confidence interval determined from 1000 bootstraps. Figures S16, S17, S18, and S19 show similar trends for different choices of RMSD cutoff.

Conformer Ensemble Effect on Molecular Docking

While screening approaches that use rigid conformer ensembles are used for some structure-based screens, many approaches, such as molecular docking, treat input molecules as partially flexible and the rotatable bonds are explicitly optimized. While it might seem conformational ensembles are unnecessary for these approaches, exploring the non-torsional degrees of freedom may still have some value. For example, the popular Glide docking program explicitly considers non-torsional degrees of freedom by sampling alternative ring conformations and nitrogen inversions.[49]

We explore the impact of providing different conformer ensembles as input on docking performance in Figure 11. RDKit consistently outperforms DMCG at recapitulating low (<2Å) RMSD poses. We hightlight performance at a 2Å cutoff due to its frequent use in docking evaluations.[53,54] Similar trends are observed for different choices of cutoffs (Figures S16, S17, S18, and S19). Using the lowest energy sampled conformation (from an ensemble of 250 conformers) performs better than a randomly sampled conformation for both methods. DMCG in particular benefits from using an energy minimized conformation (Figure S16) as energy minimization fixes non-standard geometries. Interestingly, using an ensemble of five conformations outperforms a single conformation, even when the amount of Monte Carlo sampling during docking increased 5X to match the additional sampling performed using the ensemble input. This difference is statistically significant (p-value < 0.001), although there is not always statistically significant difference between a randomly selected ensemble and seemingly more principled methods (with the exception of imposing a 2.0Å RMSD cutoff which can reduce the ensemble size to less than five in some cases). These results point to the need to go beyond sampling only the torsional space when docking, but also indicate docking performance can be improved simply by providing conformational ensembles to dock; it may not be necessary to change the internal docking sampling algorithm.

Discussion

We have evaluated two conformer generators, the popular RDKit ETKDG method and the current state-of-the-art deep generative DMCG method, with a focus on exploring the effect of different choices in constructing conformer ensembles for structure-based drug discovery tasks. From the exercise we draw several conclusions.

Conventional methods remain preferable to deep generative models for practical applications. We find that the classical RDKit method is generally superior to DMCG at recovering bioactive conformations (Figures 3, 4, and 5) and providing conformations suitable for docking (Figure 11). Given the current rate of progress, it is likely that newer deep generative methods will be able to outperform RDKit, although this may require adapting the metrics these methods are trained for. We speculate that the performance gap between these two methods is due to DMCG being trained to maximize coverage of the GEOM-QM9 and GEOM-Drugs sets. That is, it prefers to sample uniformly from the space of reasonable conformers while RDKit may sample from something closer to a Boltzmann distribution. This speculation is supported by reports that resampling RDKit ensembles using clustering can substantially improve its coverage metric on GEOM-QM9 and GEOM-Drugs.[34]

Energy minimization is a valuable post-processing step. Energy minimizing generated conformers generally improves their ability to recapitulate bioactive conformers (Figures 3 and 4) and selecting the lowest energy conformer generally performs better than a random conformer (Figures 5, S6, 7, and 11). Energy minimization is particularly important for improving the quality of DMCG generated conformers, while the built-in geometric and knowledge-based constraints of RDKit's ETKDG v3 algorithm need less refinement.

Selecting only the lowest-energy conformers is not sufficient to achieve the best retrieval of bioactive conformations. While energy minimization improves poor-quality geometries, selecting the lowest energy poses from a larger ensemble does not improve re-

trieval of bioactive conformations when selecting more than a single conformer (Figure 5). The structure of a receptor, and the non-bonded interactions with the ligand or solvent are hidden information from the conformer generation, and thus the energy from an isolated molecule calculation, whether with UFF or a dispersion-corrected semiempirical method such as CREST-GFN2, is less useful than geometric diversity via RMSD clustering. Similarly, machine learning methods such as DMCG which train on CREST-generated ensembles may reflect a bias towards low-energy and not bioactive conformations.[28]

Larger ensembles are not always better. Larger ensembles will always have higher likelihood at sampling a bioactive conformation, but for reasonable RMSD thresholds, the point of diminishing returns is achieved relatively quickly. It is not necessary to generate many hundreds of conformations to achieve nearly perfect recall within 2Å RMSD (Figures 5 and S6). Furthermore, in screening tasks generating larger ensembles can decrease performance (Figures 7, 8, and 9) due to increasing the number of false positives. Although the best observed ensemble size for maximal pharmacophore screening performance varied dramatically (from one to the maximum of 200, Figure 9), the diminishing returns in screening performance and increased computational complexity incurred by increasing the ensemble size suggest a reduced conformational ensemble of less than 25 conformers is likely sufficient for most structure-based screening tasks that rely on conformational sampling.[1] We emphasize this recommendation is not primarily motivated by the reduced computational demands of generating and screening more conformers, but by the potential decrease in accuracy that arises when generating more conformers increases the false positive rate faster than the true positive rate.

Filtering for structural diversity can enhance the performance of a given ensemble size. Unsurprisingly, when constructing a smaller ensemble from a larger ensemble,

[1] ChatGPT 4.0, when asked the right number of conformers for such tasks, suggests 100-500 conformers per a molecule, suggesting that the data presented here runs counter to prevailing sentiment.

there is benefit to increasing diversity by filtering conformations by their respective RMSDs. When the goal is to recapitulate a bioactive conformer, the optimal choice of filtering threshold is strongly related to what RMSD value is used to determine a sufficiently close match (Figure 6). When performing pharmacophore search, more stringent thresholds are required and are especially important when smaller ensembles are used (Figure 7).

Conformer ensembles are useful even for tasks that sample torsional degrees of freedom. Finally, we note that conformational sampling is often focused on the sampling of torsions (indeed, some methods only sample torsions[22–25]). However, the non-torsional degrees of freedom also matter and can materially affect docking performance, as illustrated in Figure 11, which shows that providing an ensemble conformers with different non-torsional parameters improves docking performance over providing a single conformer.

Data and Software Availability

Instructions and scripts for reproducing all the described analyses can be found at `https://github.com/dkoes/conformer_analysis` under an open source Apache license.

Acknowledgement

The authors thank the UPMC Hillman Cancer Center Academy for supporting a summer research experience for high school students. The research reported in this publication was supported by the National Institute of General Medical Sciences of the National Institutes of Health under award number R35GM140753 and National Science Foundation under award CHE-2102474.

Supporting Information Available

Supporting Information includes Figures S1 – S20. This material is available free of charge via the Internet at https://pubs.acs.org/.

References

(1) Hawkins, P. C. Conformation generation: the state of the art. *J. Chem. Inf. Model.* **2017**, *57*, 1747–1756.

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

(3) Koes, D. R. Pharmacophore modeling: methods and applications. *Computer-Aided Drug Discovery* **2016**, 167–188.

(4) Schaller, D.; Šribar, D.; Noonan, T.; Deng, L.; Nguyen, T. N.; Pach, S.; Machalz, D.; Bermudez, M.; Wolber, G. Next generation 3D pharmacophore modeling. *Wiley Interdiscip. Rev.: Comput. Mol. Sci.* **2020**, *10*, e1468.

(5) Leach, A. R.; Gillet, V. J.; Lewis, R. A.; Taylor, R. Three-dimensional pharmacophore methods in drug discovery. *J. Med. Chem.* **2010**, *53*, 539–558.

(6) Kitchen, D. B.; Decornez, H.; Furr, J. R.; Bajorath, J. Docking and scoring in virtual screening for drug discovery: methods and applications. *Nat. Rev. Drug Discovery* **2004**, *3*, 935–949.

(7) McNutt, A. T.; Francoeur, P.; Aggarwal, R.; Masuda, T.; Meli, R.; Ragoza, M.; Sunseri, J.; Koes, D. R. GNINA 1.0: molecular docking with deep learning. *J. Cheminf.* **2021**, *13*, 1–20.

(8) Schaub, A. J.; Moreno, G. O.; Zhao, S.; Truong, H. V.; Luo, R.; Tsai, S.-C. In *Chemical and Synthetic Biology Approaches To Understand Cellular Functions – Part B*;

Shukla, A. K., Ed.; Methods in Enzymology; Academic Press, 2019; Vol. 622; pp 375–409.

(9) Zhao, S.; Ni, F.; Qiu, T.; Wolff, J. T.; Tsai, S.-C.; Luo, R. Molecular basis for polyketide ketoreductase–substrate interactions. *Int. J. Mol. Sci.* **2020**, *21*, 7562.

(10) Verma, J.; Khedkar, V. M.; Coutinho, E. C. 3D-QSAR in drug design-a review. *Curr. Top. Med. Chem.* **2010**, *10*, 95–115.

(11) Crippen, G. M. A novel approach to calculation of conformation: distance geometry. *J. Comput. Phys.* **1977**, *24*, 96–107.

(12) 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.

(13) 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**, *60*, 2044–2058.

(14) Zhu, J.; Xia, Y.; Liu, C.; Wu, L.; Xie, S.; Wang, Y.; Wang, T.; Qin, T.; Zhou, W.; Li, H., et al. Direct Molecular Conformation Generation. *Transactions on Machine Learning Research* **2022**, https://openreview.net/pdf?id=lCPOHiztuw.

(15) Mansimov, E.; Mahmood, O.; Kang, S.; Cho, K. Molecular geometry prediction using a deep generative graph neural network. *Sci. Rep.* **2019**, *9*, 20381.

(16) Xu, M.; Wang, W.; Luo, S.; Shi, C.; Bengio, Y.; Gomez-Bombarelli, R.; Tang, J. An end-to-end framework for molecular conformation generation via bilevel programming. International Conference on Machine Learning. 2021; pp 11537–11547.

(17) Simm, G. N.; Hernández-Lobato, J. M. A generative model for molecular distance geometry. Proceedings of the 37th International Conference on Machine Learning. 2020; pp 8949–8958.

(18) Xu, M.; Luo, S.; Bengio, Y.; Peng, J.; Tang, J. Learning Neural Generative Dynamics for Molecular Conformation Generation. International Conference on Learning Representations. 2021.

(19) Shi, C.; Luo, S.; Xu, M.; Tang, J. Learning gradient fields for molecular conformation generation. International Conference on Machine Learning. 2021; pp 9558–9568.

(20) Luo, S.; Shi, C.; Xu, M.; Tang, J. Predicting molecular conformation via dynamic graph score matching. *Advances in Neural Information Processing Systems* **2021**, *34*, 19784–19795.

(21) Xu, M.; Yu, L.; Song, Y.; Shi, C.; Ermon, S.; Tang, J. Geodiff: A geometric diffusion model for molecular conformation generation. *arXiv preprint arXiv:2203.02923* **2022**,

(22) Ganea, O.; Pattanaik, L.; Coley, C.; Barzilay, R.; Jensen, K.; Green, W.; Jaakkola, T. Geomol: Torsional geometric generation of molecular 3d conformer ensembles. *Advances in Neural Information Processing Systems* **2021**, *34*, 13757–13769.

(23) Chan, L.; Hutchison, G. R.; Morris, G. M. Bayesian optimization for conformer generation. *J. Cheminf.* **2019**, *11*.

(24) 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*, 5211–5219.

(25) Jing, B.; Corso, G.; Chang, J.; Barzilay, R.; Jaakkola, T. Torsional diffusion for molecular conformer generation. *arXiv preprint arXiv:2206.01729* **2022**,

(26) 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.

(27) 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.

(28) Folmsbee, D.; Koes, D.; Hutchison, G. Systematic Comparison of Experimental Crystallographic Geometries and Gas-Phase Computed Conformers for Torsion Preferences. *ChemRxiv* **2022**,

(29) Liu, Z.; Zubatiuk, T.; Roitberg, A.; Isayev, O. Auto3D: Automatic Generation of the Low-Energy 3D Structures with ANI Neural Network Potentials. *J. Chem. Inf. Model.* **2022**, *62*, 5373–5382.

(30) Devereux, C.; Smith, J. S.; Huddleston, K. K.; Barros, K.; Zubatyuk, R.; Isayev, O.; Roitberg, A. E. Extending the applicability of the ANI deep learning molecular potential to sulfur and halogens. *J. Chem. Theory Comput.* **2020**, *16*, 4192–4202.

(31) Axelrod, S.; Gomez-Bombarelli, R. GEOM, energy-annotated molecular conformations for property prediction and molecular generation. *Sci. Data* **2022**, *9*, 185.

(32) 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*, 1652–1671.

(33) Pracht, P.; Bohle, F.; Grimme, S. Automated exploration of the low-energy chemical space with fast quantum chemical methods. *Phys. Chem. Chem. Phys.* **2020**, *22*, 7169–7192.

(34) Zhou, G.; Gao, Z.; Wei, Z.; Zheng, H.; Ke, G. Do Deep Learning Methods Really Perform Better in Molecular Conformation Generation? *arXiv preprint arXiv:2302.07061* **2023**,

(35) Zhang, H.; Zhang, J.; Zhao, H.; Jiang, D.; Deng, Y. Infinite Physical Monkey: Do Deep Learning Methods Really Perform Better in Conformation Generation? *bioRxiv* **2023**, 2023–03.

(36) Wang, R.; Fang, X.; Lu, Y.; Wang, S. The PDBbind database: Collection of binding affinities for protein- ligand complexes with known three-dimensional structures. *J. Med. Chem.* **2004**, *47*, 2977–2980.

(37) Mysinger, M. M.; Carchia, M.; Irwin, J. J.; Shoichet, B. K. Directory of useful decoys, enhanced (DUD-E): better ligands and decoys for better benchmarking. *J. Med. Chem.* **2012**, *55*, 6582–6594.

(38) Chen, L.; Cruz, A.; Ramsey, S.; Dickson, C. J.; Duca, J. S.; Hornak, V.; Koes, D. R.; Kurtzman, T. Hidden bias in the DUD-E dataset leads to misleading performance of deep learning in structure-based virtual screening. *PloS One* **2019**, *14*, e0220113.

(39) Sieg, J.; Flachsenberg, F.; Rarey, M. In need of bias control: evaluating chemical data for machine learning in structure-based virtual screening. *J. Chem. Inf. Model.* **2019**, *59*, 947–961.

(40) Tran-Nguyen, V.-K.; Jacquemard, C.; Rognan, D. LIT-PCBA: an unbiased data set for machine learning and virtual screening. *J. Chem. Inf. Model.* **2020**, *60*, 4263–4273.

(41) Wierbowski, S. D.; Wingert, B. M.; Zheng, J.; Camacho, C. J. Cross-docking benchmark for automated pose and ranking prediction of ligand binding. *Protein Sci.* **2020**, *29*, 298–305.

(42) Rappé, A. K.; Casewit, C. J.; Colwell, K.; Goddard III, 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.

(43) Sunseri, J.; Koes, D. R. Pharmit: interactive exploration of chemical space. *Nucleic Acids Res.* **2016**, *44*, W442–W448.

(44) Koes, D. R.; Camacho, C. J. Pharmer: efficient and exact pharmacophore search. *J. Chem. Inf. Model.* **2011**, *51*, 1307–1314.

(45) Trott, O.; Olson, A. J. AutoDock Vina: improving the speed and accuracy of docking with a new scoring function, efficient optimization, and multithreading. *J. Comput. Chem.* **2010**, *31*, 455–461.

(46) Ragoza, M.; Hochuli, J.; Idrobo, E.; Sunseri, J.; Koes, D. R. Protein–ligand scoring with convolutional neural networks. *J. Chem. Inf. Model.* **2017**, *57*, 942–957.

(47) Stärk, H.; Ganea, O.; Pattanaik, L.; Barzilay, R.; Jaakkola, T. Equibind: Geometric deep learning for drug binding structure prediction. International Conference on Machine Learning. 2022; pp 20503–20521.

(48) Corso, G.; Stärk, H.; Jing, B.; Barzilay, R.; Jaakkola, T. Diffdock: Diffusion steps, twists, and turns for molecular docking. *arXiv preprint arXiv:2210.01776* **2022**,

(49) Friesner, R. A.; Banks, J. L.; Murphy, R. B.; Halgren, T. A.; Klicic, J. J.; Mainz, D. T.; Repasky, M. P.; Knoll, E. H.; Shelley, M.; Perry, J. K., et al. Glide: a new approach for rapid, accurate docking and scoring. 1. Method and assessment of docking accuracy. *J. Med. Chem.* **2004**, *47*, 1739–1749.

(50) Koes, D. R.; Baumgartner, M. P.; Camacho, C. J. Lessons learned in empirical scoring with smina from the CSAR 2011 benchmarking exercise. *J. Chem. Inf. Model.* **2013**, *53*, 1893–1904.

(51) 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*, 1–14.

(52) 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*, D420–D427.

(53) Scarpino, A.; Ferenczy, G. G.; Keserű, G. M. Comparative evaluation of covalent docking tools. *J. Chem. Inf. Model.* **2018**, *58*, 1441–1458.

(54) Tuccinardi, T.; Poli, G.; Romboli, V.; Giordano, A.; Martinelli, A. Extensive consensus docking evaluation for ligand pose prediction and virtual screening studies. *J. Chem. Inf. Model.* **2014**, *54*, 2980–2986.

TOC Graphic