HOW TO GO WITH THE FLOW: AN ANALYSIS OF FLOW MATCHING MOLECULAR DOCKING PERFORMANCE WITH PRIORS OF VARYING INFORMATION CONTENT

Anonymous authors
Paper under double-blind review

ABSTRACT

Molecules are both architectural marvels, exhibiting well-defined structural features, and proficient gymnasts, flexibly adjusting their structure to bind to protein pockets. Predicting molecular docking poses with flow matching algorithms represents both a promising opportunity and a challenging task. Recently, a flow matching algorithm, HarmonicFlow, has been reported to yield encouraging molecular docking results. The method employs a harmonic prior. In light of the importance of long-range information for molecular structure, we sought to understand the consequences of the harmonic prior for docking results. We found that the method often provides compressed poses, and there is some correlation between this compression and docking performance. We retrained the method to use a prior incorporating information from a molecular conformation, to determine whether a prior with long-range information would provide better performance. Performance did not improve with this new prior, whether the exact long-range information was used or whether noise was added. This finding suggests that further prior development is unlikely to improve performance, implying perhaps advances in the neural network could be another avenue to consider. Therefore, we present and discuss some possible ways to leverage local and long-range structural information in the neural network. Bond distances could be incorporated, and also information could be harvested from conformer ensembles regarding longer-range distances. By understanding chemical features associated with flow matching docking performance, investigating results with a more chemically-informed prior, and suggesting possible neural network advances, this work advances the molecular machine learning community's understanding of the repertoire of opportunities available to improve docking performance.

1 INTRODUCTION

In a rewarding but challenging drug discovery journey,(Kola & Landis, 2004) structural information can help illuminate the path forward.(Blundell, 1996) Viewing how a protein interacts with a target ligand can catalyze progress in drug design: by seeing the structure, a medicinal chemist can generate ideas for new molecules with enhanced interactions.(Greer et al., 1994) Experimentally obtaining structures can be resource-intensive. Molecular docking employs *in silico* techniques to generate a structure of a protein-ligand complex, providing critical structural information with lower cost.(Shoichet et al., 2002) Traditionally, docking methods were physics-based, employing first principles governing intermolecular interactions in order to predict a ligand's pose.(Friesner et al., 2004) While helpful, these methods simplify the underlying physics, which can limit output quality.

Machine learning methods offer a promising alternative. Instead of hard-coding in particular physics, they learn from the available data, analyzing input protein-ligand complexes in order to grasp the principles governing protein-ligand interactions. Recently, generative models have demonstrated strong docking performance.(Corso et al., 2022; Stärk et al., 2023) HarmonicFlow (HF) is a flow-matching generative model, learning a vector field to find the ligand pose.(Stärk et al., 2023) This method, although representing an improvement over others, does not provide RMSDs under 2 Å of the correct pose in over half of cases even when the pocket is defined (Stärk et al., 2023), offering opportunities for further improvement.

Previous diffusion-based docking models (Corso et al., 2022) sample from a Gaussian, then denoise it to find the docking pose. Flow matching algorithms learn the flow field between the initial and the true data distribution. This allows the model to use an informative prior as input instead of a Gaussian. An informative prior will provide positive inductive bias for the flow-matching process.

The current flow-matching algorithm for ligand docking employs a harmonic prior, sampling bonded atoms near each other.(Jing et al., 2023) We noted that this prior would therefore neglect longer-range information. This observation motivated two lines of inquiry discussed in this paper. First, we sought to understand the consequences of the loss of long-range information for poses and for docking performance. Second, we modified the HF prior to incorporate long-range information and compared performance across priors. We present both of these investigations and then share some possible future avenues based on the results and further structural analysis.

2 METHODS

The training, validation, and test sets were from PDBbind (Liu et al., 2017), using a time split. Distance-Pocket was used to define the binding pocket. We experiment with an additional informative prior that uses RDKit(rdk), an open-source toolkit for cheminformatics, to calculate the atom coordinates in the ligand. Specifically, we use RDKit to sample a different conformer each time, and then use RDKit to generate the coordinates using distance geometry. We also experimented with a prior that adds Gaussian noise to the coordinates generated by RDKit using distance geometry. We retrained HarmonicFlow using Gaussian, harmonic, RDKit, and RDKit with Gaussian noise priors. Each model is trained with a batch size of 8 and 200 epochs. Root-mean-square deviation (RMSD) of the docked ligand pose compared to PDBbind was used to analyze docking performance. There were 268 ligands in the test set used after preprocessing.

RMSD was analyzed in two different ways for the analysis. In 3.1 and 3.3, RMSD for each complex was computed for one output pose out of the the samples, so that analysis could be connected to a single structure. Meanwhile, in 3.2, from the HF output, a summary of test set RMSD performance was found directly from the HarmonicFlow output. In both cases, RMSD was relative to the PDBbind ligand, and no symmetry adjustment was applied. From each complex identity, features describing the ligand could be computed from the processed PDBbind data provided in the HF GitHub Zenodo link. RDKit(rdk) was used to calculate ligand features.

Conformers for the consensus analysis in 3.3 were generated with RDKit ETKDGv2,(Riniker & Landrum, 2015) following a procedure adapted from earlier work.(Stärk et al., 2022) Several modifications were made, including making conformer generation deterministic to aid analysis. Conformers were generated via a different procedure for the retraining in 3.2.

3 RESULTS

3.1 DRIVERS OF HARMONICFLOW WITH HARMONIC PRIOR DOCKING PERFORMANCE

In light of the harmonic prior's only incorporating local information, we sought to investigate: do ligands with more long-range information perform less well in docking? We found that there is a slight association between ligand heavy atom count and HF performance (Figure 6a). (HarmonicFlow discussed in this section refers to HarmonicFlow with a harmonic prior: we focused on this prior because of its extensive use in earlier work.(Stärk et al., 2023)) Intuitively, larger ligand size should correspond to higher count of non-local contacts (i.e., larger ligands have larger counts of atom pairs not directly bonded to each other). Another relevant type of long-range information is count of rotatable bonds, measuring flexibility of the ligand. We found there is not much association between ligand rotatable bond count and HF performance (Figure 6b), though the association is in the direction of higher rotatable bond count with worse RMSD performance. Thus, loss of long-range information for the initial HarmonicFlow structures appears slightly more deleterious for larger ligands: because they are larger, non-local information could be more relevant.

To further understand possible structural underpinnings of RMSD performance, we considered the consequences of the lack of long-range information for the output predicted structure poses. The initial harmonic prior produces a quite compressed structure.(Stärk et al., 2023) To what extent does

the final pose retain some folded-in character? We compared the radius of gyration of the PDBbind ligand structure with the HF-predicted ligand structure. Radius of gyration is a geometric parameter which can be interpreted as measuring the extent of dispersion of a set of points, such as atoms of a molecule. (Lobanov et al., 2008) We found percent error of the radius of gyration for the HF-predicted ligand structure, compared to the PDBbind ligand structure. Percent error helps avoid possible confounding issues due to size effects. We found that many of the HF-predicted poses have a negative radius of gyration percent error, meaning that HF is predicting poses which are more compact, or folded-in, compared to the actual crystal structure poses (Figure 1b).

Furthermore, we found that higher over-estimation of compactness is associated with worse RMSD performance (Figure 1c), but not with heavy atom count(Figure 8c). The explanation could be that HarmonicFlow begins with quite folded-in structures, with steric repulsions. While the inference process should help unfold the structure, there seems to still be some folded-in quality retained. This finding further motivated our existing interest in incorporating long-range information into the HarmonicFlow prior: could starting with more extended structures improve performance?

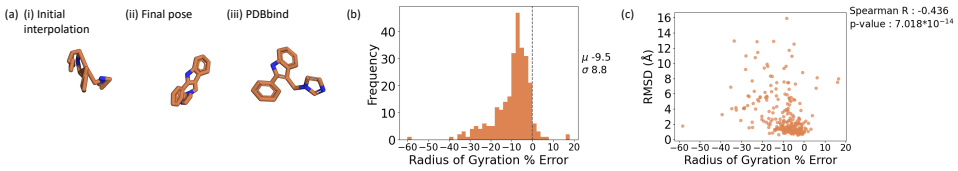

Figure 1: (a) 6PYA(Round et al., 2020) (i) Initial interpolation structure (ii) Final pose (iii) PDBbind pose. This structure was selected to be illustrative of compression, not to be representative. (b) Distribution of percent errors in radii of gyration from inference (percent error relative to PDBbind structure). (c) Scatterplot of RMSD with radius of gyration percent error.

3.2 Results for HarmonicFlow with a RDKit Prior

We calculated the RMSD on the test set for each model with different priors. RMSD < 2Å and Median RMSD are reported in Table 1. The results reveal that the harmonic prior still has the best percentage of RMSD < 2Å and lowest median RMSD. RDKit has the worst performance in both metrics, indicating that using an RDKit prior might provide negative inductive bias to the model. However, the RDKit with noise prior outperforms the Gaussian prior in both metrics, implying introducing additional randomness to the RDKit prior could provide positive inductive bias. All of the priors underestimate radius of gyration (Figure 2).

	RMSD < 2Å	Median RMSD
Gaussian	49.5	2.02
Harmonic	51.8	1.92
RDKit	45.3	2.35
RDKit w/ Noise	50.2	1.98

Table 1: **Comparison of different priors.** RMSD < 2Å represents the percentage of predictions that have an RMSD to the ground truth within 2Å. Median RMSD represents the median RMSD to the ground truth.

3.3 Analysis of Structures Suggests Possible Future Directions

This work considers priors with a range of amounts of information: while a harmonic prior provides only information on bonded atoms, a RDKit prior provides an entire feasible three dimensional structure. The similarity of performance for these priors with vastly different amounts of information suggests that further developing a prior based on a single structure is unlikely to yield a meaningful performance improvement. Furthermore, the finding that priors with long-range structures underestimate radius of gyration (Figure 2) suggests that long-range prior information is not sufficient for avoiding compressed poses. Therefore, another possible route to improve performance

could be to incorporate additional information outside of the prior, such as into the neural network architecture. This section describes initial preliminary analyses to investigate: how can chemical understanding provide relevant information for the neural network?

In light of the graph representation of the molecule employed, pairwise distance data represents a particularly promising resource. Because the atoms are already represented as nodes, and a framework exists to encode information regarding pairs of atoms as edges, pairwise information is particularly compatible with the model architecture.Stärk et al. (2023) While the current graph representation does incorporate distance information in edges, the underestimated radii of gyration raise the question of whether richer distance information could be helpful. We present some thoughts first on bond distances, then on longer-range distances. We emphasize the goal is to share general observations to guide future method development, and that we have not yet tested possible advances suggested in this forward-looking brainstorming section. In this section, HarmonicFlow refers to HarmonicFlow with a harmonic prior, and the test set ligands were analyzed.

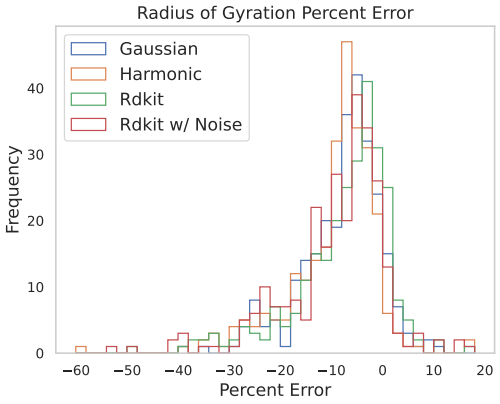

Figure 2: **Comparison of Radius of Gyration.** Percentage error of radius of gyration is calculated using the PDBbind structure as a reference. The radius of gyration percentage error is calculated for each different prior.

We found that HarmonicFlow underestimates bond lengths, relative to the PDBbind structures (Figure 3a). There are very often negative bond length percent errors. We found the average bond length percent error for each structure, to summarize the extent of this underestimation. It is correlated with both radius of gyration percent error and RMSD (Figure 3b). Such a correlation raises the question: if only bonded information was in the prior, as is the case in the current harmonic prior, but bond length information was provided outside the prior, would RMSD performance improve? One further direction could be to include information in the neural network input graph bonded atom edges about bond distances based on chemical knowledge of bond lengths from element and hybridization information.

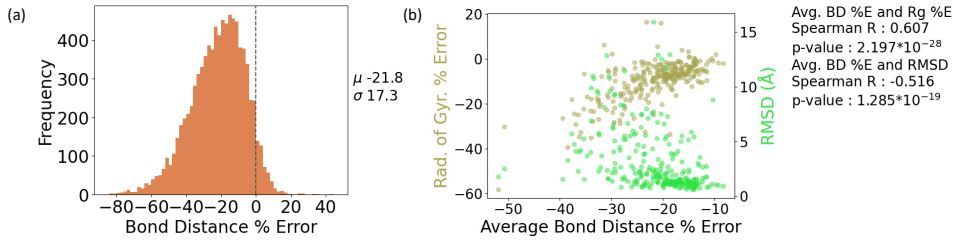

Figure 3: (a) Percent error of HarmonicFlow individual bond distances relative to PDBbind values. (b) Scatterplot of each complex's average HarmonicFlow bond distance percent error with RMSD and radius of gyration percent error.

We investigated whether an RDKit-generated conformer ensemble could provide longer-range distance information relevant to the final ligand pose. For each ligand, we generated 10 RDKit conformers. Because bond distances and angles should be expected to vary little (by chemical intuition), we focused on atoms separated by a torsion or more (over 2 bonds away from each other). For each such pair of atoms, we found the pairwise distance in each conformer (Figure 4). We computed the standard deviation of each such pairwise distance across conformers. There exists a wide spread of these standard deviations(Figure 10). Pairwise distance uncertainty may be relevant to neural network architecture. One possibility is high certainty (low standard deviation across conformers) information could be incorporated, so that distance information that is inferred from the conformers

to be non-controversial helps guide the model. A second possibility is low certainty (high standard deviation) information could also help the model understand opportunities for flexibility in the molecule. One further step could be to correct for the number of bonds between atoms, so that the standard deviation is contextualized more explicitly in the bonding framework. For this paper, we focus on high certainty pairwise distances, though there may also be insight to be gained from other pairwise distances.

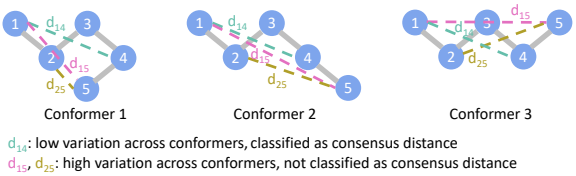

Figure 4: Conformer consensus distance analysis schematic

Those pairwise distances with low variation (standard deviation under 0.10 Å) across conformers represented consensus distances (Figure 4): conformers are in approximate agreement on their value. The existence of consensus distances is quite common, with 94% of structures having at least 5 consensus distances (Figure 5a). We found that these consensus distances have low percent error when comparing their average values (across conformers) and the corresponding value for the distance between the same atom pair in the PDBbind ligand pose. Furthermore, in the HarmonicFlow pose, these same distances are often significantly underestimated (Figure 5b). A further direction could be to incorporate into the neural network input graph edges information corresponding to these consensus distances, thus including insights from the agreement between conformers which are both longer-range and likely to be accurate. This initial preliminary analysis focused on the output of conformer generation, and we also note that the distance geometry field(Havel, 1998; Blaney & Dixon, 1994), including the conceptual foundation of ETKDG approach(Riniker & Landrum, 2015), could be an area poised to contribute to future work. Further review of this literature could be fruitful, and it is possible distance geometry information could be directly relevant to the neural network edges, without conformer generation acting as an intermediary of sorts. Also, it is likely (by chemical intuition) a number of these consensus distances correspond to aromatic rings: if this is the case, adding aromatic ring geometric information to the neural network may be another approach.

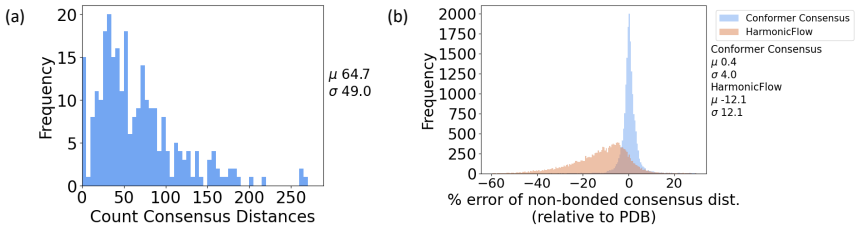

Figure 5: (a) Count of consensus distances in each structure. (b) Distribution of consensus distance percent error, with PDBbind as the reference, for the RDKit conformer ensemble's average distance value and for HarmonicFlow.

4 Conclusion, Limitation and Future Work

We have found that the existing HarmonicFlow method with a harmonic prior produces compressed poses. Including longer range information does not necessarily improve flow matching results. Further investigation is needed to understand why the harmonic prior performs best even without longer range information. Incorporating distance information from structural analysis into the neural network architecture represents a possible future direction. By investigating the effect of priors on flow matching performance and suggesting neural network advances, this work expands the molecular machine learning community's understanding of how its toolbox can be leveraged for docking.

REFERENCES

Rdkit: Open-source cheminformatics. https://www.rdkit.org.

scipy.stats.spearmanr. https://docs.scipy.org/doc/scipy/reference/generated/scipy.stats.spearmanr.html.

Jeffrey M Blaney and J Scott Dixon. Distance geometry in molecular modeling. *Reviews in computational chemistry*, pp. 299–335, 1994.

Tom L Blundell. Structure-based drug design. *Nature*, 384(6604):23, 1996.

Gabriele Corso, Hannes Stärk, Bowen Jing, Regina Barzilay, and Tommi Jaakkola. Diffdock: Diffusion steps, twists, and turns for molecular docking. *arXiv preprint arXiv:2210.01776*, 2022.

Richard A Friesner, Jay L Banks, Robert B Murphy, Thomas A Halgren, Jasna J Klicic, Daniel T Mainz, Matthew P Repasky, Eric H Knoll, Mee Shelley, Jason K Perry, et al. Glide: a new approach for rapid, accurate docking and scoring. 1. method and assessment of docking accuracy. *Journal of medicinal chemistry*, 47(7):1739–1749, 2004.

Jonathan Greer, John W Erickson, John J Baldwin, and Michael D Varney. Application of the three-dimensional structures of protein target molecules in structure-based drug design. *Journal of medicinal chemistry*, 37(8):1035–1054, 1994.

Charles Harris, Kieran Didi, Arian R Jamasb, Chaitanya K Joshi, Simon V Mathis, Pietro Lio, and Tom Blundell. Benchmarking generated poses: How rational is structure-based drug design with generative models? *arXiv preprint arXiv:2308.07413*, 2023.

Timothy F Havel. Distance geometry: Theory, algorithms, and chemical applications. *Encyclopedia of Computational Chemistry*, 120:723–742, 1998.

Bowen Jing, Ezra Erives, Peter Pao-Huang, Gabriele Corso, Bonnie Berger, and Tommi Jaakkola. Eigenfold: Generative protein structure prediction with diffusion models. *arXiv preprint arXiv:2304.02198*, 2023.

Ismail Kola and John Landis. Can the pharmaceutical industry reduce attrition rates? *Nature reviews Drug discovery*, 3(8):711–716, 2004.

Zhihai Liu, Minyi Su, Li Han, Jie Liu, Qifan Yang, Yan Li, and Renxiao Wang. Forging the basis for developing protein–ligand interaction scoring functions. *Accounts of chemical research*, 50 (2):302–309, 2017.

M Yu Lobanov, NS Bogatyreva, and OV Galzitskaya. Radius of gyration as an indicator of protein structure compactness. *Molecular Biology*, 42:623–628, 2008.

Sereina Riniker and Gregory A Landrum. Better informed distance geometry: using what we know to improve conformation generation. *Journal of chemical information and modeling*, 55(12): 2562–2574, 2015.

Phillip Round, Samir Das, Tsung-Sheng Wu, Kristiina Wähälä, Filip Van Petegem, and Geoffrey L Hammond. Molecular interactions between sex hormone–binding globulin and nonsteroidal ligands that enhance androgen activity. *Journal of Biological Chemistry*, 295(5):1202–1211, 2020.

Brian K Shoichet, Susan L McGovern, Binqing Wei, and John J Irwin. Lead discovery using molecular docking. *Current opinion in chemical biology*, 6(4):439–446, 2002.

Hannes Stärk, Octavian Ganea, Lagnajit Pattanaik, Regina Barzilay, and Tommi Jaakkola. Equibind: Geometric deep learning for drug binding structure prediction. In *International conference on machine learning*, pp. 20503–20521. PMLR, 2022.

Hannes Stärk, Bowen Jing, Regina Barzilay, and Tommi Jaakkola. Harmonic self-conditioned flow matching for multi-ligand docking and binding site design. *arXiv preprint arXiv:2310.05764*, 2023.

A Appendix

A.1 Code Availability

Code can be accessed at: Anonymous GitHub Link.

A.2 Additional Methodological Details

A.2.1 Ligand Processing

All ligand files were checked to ensure that they only contain one molecule, and ligand files with more than one molecule were excluded. The PDBbind ligands were sanitized in RDKit, although the final poses were not due to RDKit errors encountered. While sanitizing in general changes some properties, our initial not-yet-comprehensive checks thus far have not indicated changes on sanitizing in properties which we compared between PDBbind ligands and final poses. For radius of gyration, we note that, because the HF poses do not contain hydrogens, we compared the output of HF to the PDBbind ligand with hydrogens removed: although RDKit did retain several hydrogens in some PDBbind poses after removal was requested, we doubt that this will drastically change results, although further cleaning may be helpful.

As mentioned above in the discussion of RMSD, symmetry and chemical equivalence across structures may be an issue for conformer pairwise distance standard deviations as well, and this is an area for possible future work. If a further symmetry adjustment does turn out to be necessary, that should only lower the standard deviations, so our impression is that the consensus distance analysis would reveal more consensus distances in this scenario.

A.2.2 Correlation Analysis and Plotting

To check correlations, we used the Spearman rank correlation coefficient. We note that the documentation(sci) mentioned p-values are only valid for more data points than we had. We retain the p-values for qualitative general interpretation purposes, but we emphasize to the reader that the p-value is not entirely suitable here. The permutation analysis noted in the documentation is a possible area of future interest.(sci) While we made an effort to always include all data points in plots, we did sometimes truncate bounds when extreme values were present, to aid interpretation.

A.3 Possible Further Extensions

While this work is ligand-focused, considering protein-ligand interactions could be another possible future direction. We have carried out an initial preliminary PoseCheck(Harris et al., 2023) study of results from another HarmonicFlow run and compared to DiffDock PoseCheck results. While differences between the methods' settings complicate a direct comparison, we noted that the HarmonicFlow run did have fewer steric clashes as determined by PoseCheck. These fewer clashes may be due to the aforementioned pose compression. Comparing protein-ligand interactions in HarmonicFlow output versus PDBbind structures, and further physical feasibility analysis, may yield suggestions of additional neural network architecture updates for components relating to protein-ligand interactions.

A.4 Additional Figures

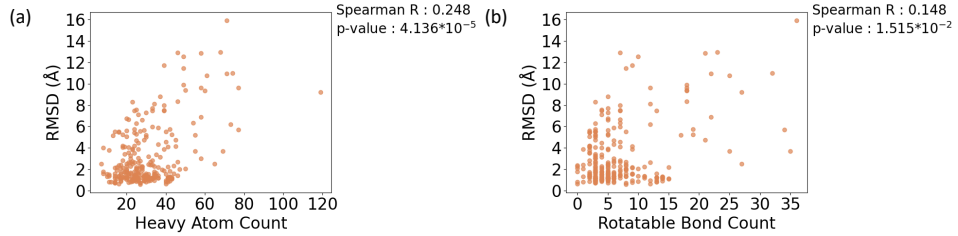

Figure 6: Scatterplot of RMSD and (a) ligand heavy atom count (b) ligand rotatable bond count

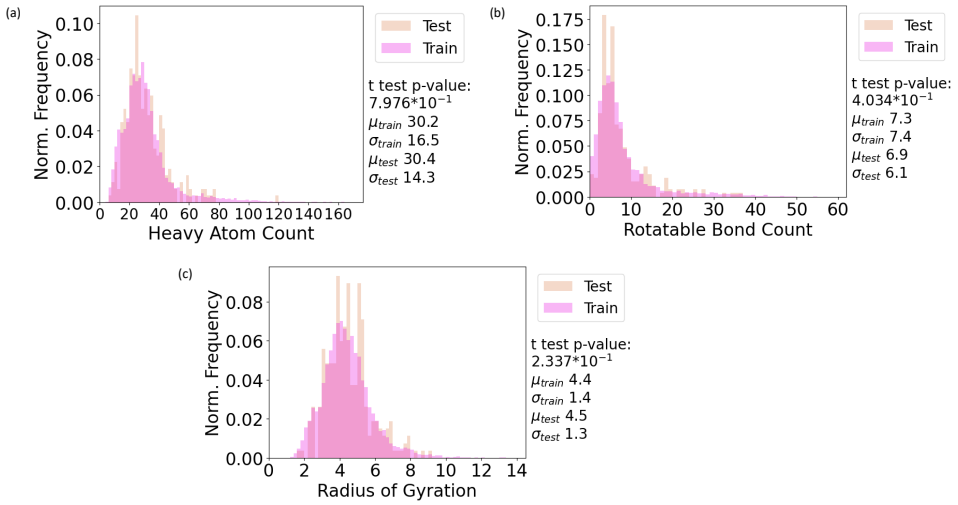

Figure 7: Distributions in the test and training sets of (a) heavy atom count, (b) rotatable bond count, and (c) radius of gyration

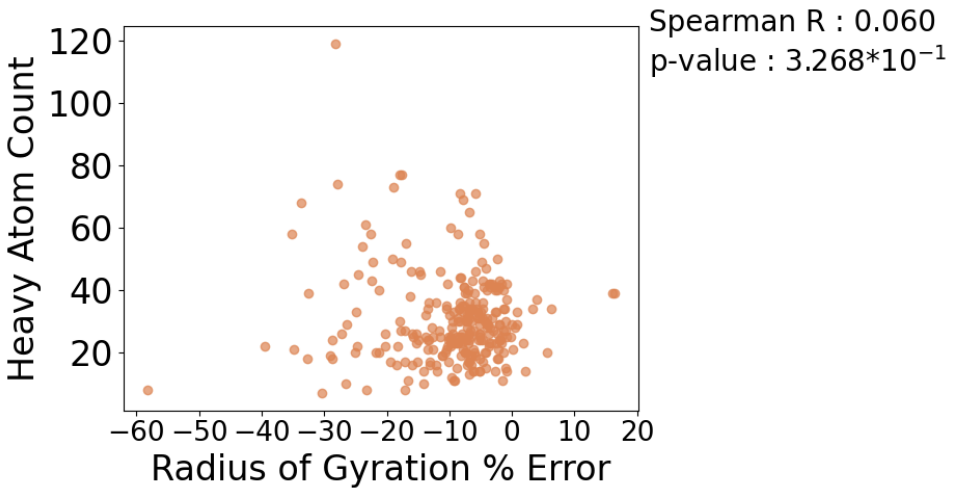

Figure 8: Scatterplot of radius of gyration and heavy atom count

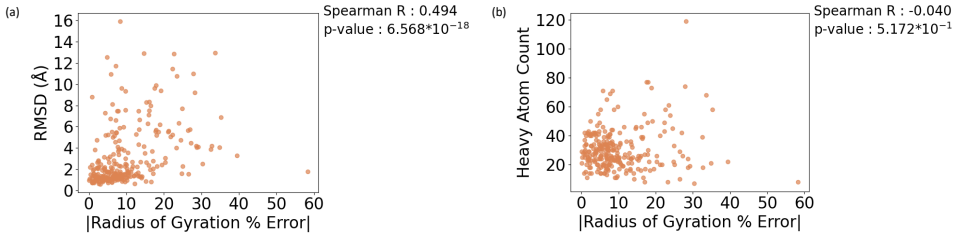

Figure 9: (a) Scatterplot of RMSD with absolute value of radius of gyration percent error. (b) Scatterplot of heavy atom count with absolute value of radius of gyration percent error.

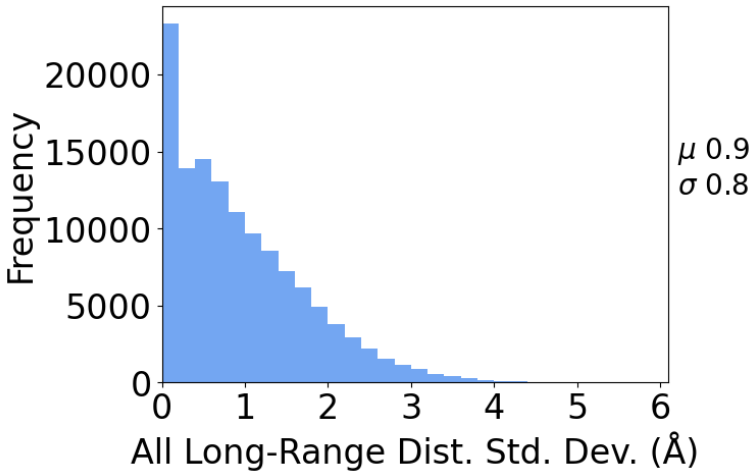

Figure 10: Distribution of all conformer ensemble standard deviations of pairwise distances of atoms separated by 3 or more bonds

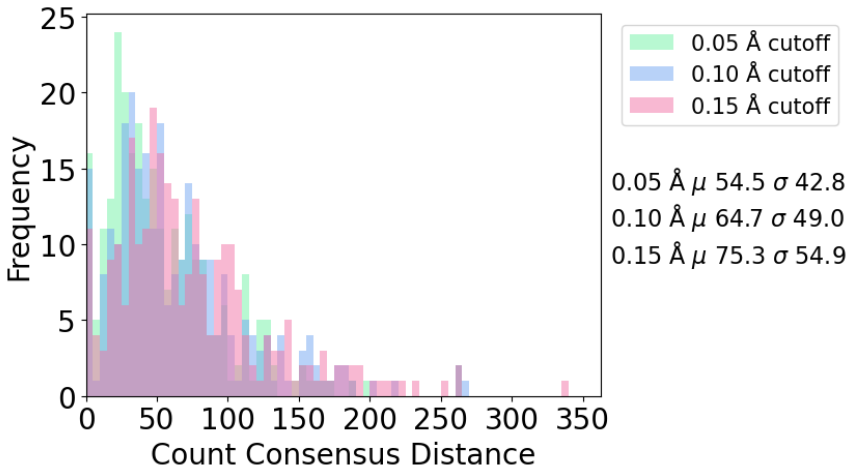

Figure 11: Consensus distance count distribution, for different consensus distance definition threshold values

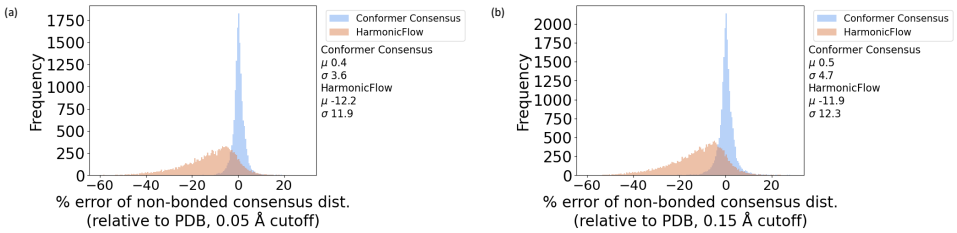

Figure 12: Consensus distance percent error (relative to PDBbind) comparison in the conformer ensemble's average distance value versus in HarmonicFlow, for different consensus distance definition threshold values