How to Go with the HarmonicFlow: Understanding Drivers of HarmonicFlow Performance and Improving Priors

Sam Huang, Sammy Mustafa, Mo Oyewole, Dina Sharon

Background/Introduction:

Discovering a new drug can be a long, arduous journey [Kola et al., 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 binding pose in a pocket [Friesner et al, 2004]. While sometimes helpful, these methods are limited by the physics which they use. For instance, many methods model atoms as point charges, neglecting features such as polarizability. A method which simplifies the underlying physics is inherently limited in the quality of output it can produce.

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. Many previous docking methods are regression-based [Stärk et al., 2022]: this is an issue because they may select an average of two good solutions, which is not in itself a good solution. Recently, generative models have demonstrated improved docking performance. DiffDock is a diffusion generative model, denoising ligand torsions and position relative to the protein [Corso et al., 2023]. HarmonicFlow (HF) is a flow-matching generative model, learning a vector field to find the ligand pose [Stärk et al., 2023]. While these high-quality methods represent significant advances, they still do 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. In addition, deep learning methods can produce poses which do not entirely align with physical intuition [Buttenschoen et al., 2023]. So despite the fact that on paper these models may generate great dockings, many times those dockings cannot happen in real life due to physical constraints. Avenues for further development include understanding chemical drivers of deep learning docking success or lack thereof and improving priors of the flow matching model (including using the drivers discovered to inform prior improvement).

Innovation (we do not expect you to make significant innovations but want to know what you feel is new about the work you are proposing):

Understanding Chemical Drivers: Unfortunately, previous work on evaluating deep learning for docking has focused on identifying failure modes [Buttenschoen et al., 2023; Yu et al. 2023] or pinpointing interactions which are and are not well-represented [Harris et al., 2023]. While these provide some information on issues of which to be aware, a much more constructive approach (which to the best of our knowledge has not yet been employed) would be to view the success or lack thereof from the perspective of particular ligands and pockets, and to connect physical feasibility more directly to extent of docking success. While RMSD correlation analysis with Tanimoto similarity to the training set, atom count, and rotatable bond count has been conducted for DiffDock (Figures 8 and 9 of [Corso et al, 2023].), there exists a wealth of additional ligand features which could be investigated. Furthermore, exploration of possible protein drivers of deep learning docking success has not yet been carried out, to the best of our knowledge.

Improving prior for HarmonicFlow: A prior is used to incorporate existing knowledge or beliefs into the flow matching model. A useful prior will improve the model efficiency and improve the quality of docking. It will add constraints to the flow matching model such that highly implausible structures are less likely to be reached.

Significance:

Understanding Chemical Drivers: The ability of machine learning to comprehend and apply the rules of physics is inspiring. Better understanding which protein-ligand complex structures are more and less difficult to predict can inform on opportunities to develop models with higher predictive power. Hannes shared that improving generalization could help increase the methods' impact. Also, Hannes shared that building a better dataset may help improve performance, because there can be issues when a ligand or pocket does not resemble training data. Articulating properties of less successful predictions could pave the way to rational curation of the additional training data which would be the most impactful for improving performance.

Prior for HarmonicFlow: Different from previous work DiffDock, HarmonicFlow does not restrict ligand flexibility to torsions. A realistic and informative prior might be more important to bias models to generate viable dockings. Also, most molecule generation and optimization tasks focus on the process of diffusion or flow models, but fewer are done on the priors. Developing an informative prior will not only improve Harmonic Flow but also may be applied to other diffusion/flow models of different tasks.

Specific Aims (including approaches):

We strive to explore this new frontier from a multifaceted perspective and leverage our interdisciplinary expertise. Thus, we propose to both better comprehend where deep learning docking does and does not succeed, and to rationally employ this knowledge and other insight obtained from our learning and discussions to develop improved HarmonicFlow priors, and possibly novel evaluation metrics also.

1. Understanding chemical drivers of existing deep learning docking results: Initial inspection of docked poses will enable generation of hypotheses regarding cases where deep learning docking is highly accurate, and cases where deep learning docking is less accurate. (We will initially be focusing on HF, but possibly also include DiffDock.) A workflow will then be developed to test these hypotheses in a high-throughput fashion. Correlations of RMSD with different ligand or protein-specific metrics will be developed. Three possible failure modes are: (a) difficulty with docking ligands/proteins with certain chemical features (i.e., providing ligand structures which are physically feasible but have high RMSD), (b) lack of predicted structures' physical feasibility, and (c) difficulty with generalization. All three failure modes will be investigated initially, with preliminary findings guiding the mode(s) on which we will focus.

(a) Chemical features: We will investigate whether ligand or protein chemical features are correlated with docking success. For instance, one possible hypothesis is that deep learning docking is more successful for more nonpolar ligands and binding sites, because atomistic information (not in the protein model) may be more important for polar directional interactions. Evidence for this hypothesis can be found in [Harris et al., 2023] which showed hydrogen bonds are more difficult for machine learning to capture than nonpolar interactions are. We are also grateful to Hannes for suggesting by Slack that protein pocket shape could be another metric to consider: indeed, visual inspection of Figure 7 in [Stärk et al., 2023] shows that the bottom right structure has a docked pose quite different from the experimental pose, and the binding site appears mostly on the surface, not buried. A possible hypothesis to test is whether deep learning docking is more successful for buried pockets, with presumably more protein-ligand interactions.

(b) Physical feasibility: If a predicted ligand-protein complex structure is not physically feasible, then a high RMSD may arise as well. For instance, an incorrect double bond stereochemistry (*E* vs *Z*) in a center of a molecule can dramatically displace almost half of a molecule's atoms in a pose. Comparison of ligand structure and protein-ligand clashes in predicted poses and ground truth PDB files will help in understanding how well deep learning docking is providing physically feasible structures. In contrast to [Buttenschoen et al., 2023], which (based on our understanding) focused on different ways low-RMSD poses are not physically feasible, we will take a more constructive and pragmatic approach, seeing if this known lack of perfect physical feasibility is a driver of the very impressive RMSD success rate's not being even higher.

(c) Generalization: Hannes shared that deep learning docking may not be as successful when a ligand or protein is very different from the training set. Hannes' comment led to an idea that a possible hypothesis to evaluate is whether binding pocket composition similarity (of a test set protein pocket to training set pockets) can predict RMSD. Regarding ligand similarity, while one Tanimoto similarity analysis has been done for DiffDock (Figure 9 of [Corso et al., 2023]), further HF Tanimoto similarity studies may also be conducted.

2. *Improving prior for HarmonicFlow*: Current HF models use a harmonic prior (from EigenFold [Jing et al., 2023]) that samples atoms to be close to each other if they are connected by a bond. This prior created an inductive bias where atoms of different molecules should already be spatially separated. However, many other priors can be used in the flow matching framework. Hannes suggested some preliminary ideas for improving priors, writing "using structures from RDKit, structures from RDKit with some noise added to them, and a mixture of priors such as several structures from RDKit to condition the model and the transformed coordinates…from a Gaussian."

Unification: The overall goal is to leverage the chemical drivers of docking success identified in (1) to make suggestions to improve HF docking performance, by directly informing prior generation in (2) and possibly also developing new metrics to evaluate HF performance with different priors. Regarding informing prior generation, (1) can reveal what types of physics are not currently fully captured- and perhaps could be incorporated into the prior. For instance, if ligand poses are not physically feasible, we then want to try changing the prior that is used in HF to see if another type of prior can lead to a more physically feasible molecule. Another example is: if polar ligands have high RMSDs, perhaps additional geometric information for polar atoms (e.g., bond angles) can be added into the prior.

In addition to developing new priors, the analysis in (1) can lead to creating new metrics that can be used to measure the performance of the protein docking model and then investigating how well HF performs with different priors from (2). For instance, we may want to evaluate docking pose feasibility. This is a very interesting problem because various recently published papers [Buttenschoen et al., 2023, Harris et al. 2023] have shown that generative AI methods tend to struggle to make physically sensible molecules that have unprecedented chemical properties. We want to contribute to this space. Thus, we want to use chemical knowledge to create metrics of how physically feasible a molecule is based on its structure and the properties of its components. While [Buttenschoen et al., 2023, Harris et al. 2023] seemed to only use binary or discrete metrics (based on our understanding), we may want to develop continuous ones (or apply existing ones in other sub-fields which have not previously been used for deep learning docking to the best of our knowledge), such as a feasibility score combining results of different checks. This will give us a good way to estimate the physical sensibility of the molecules generated by HF. We can use these metrics to compare how well different priors perform, supplementing insight the RMSD analysis provides.

Resources:

Data availability: The PDBBind dataset contains protein-ligand complexes which are highly informative for training and testing [Liu et al., 2017]. These were used in DiffDock and HarmonicFlow [Corso et al., 2023, Stärk et al., 2023]. Furthermore, the HarmonicFlow GitHub page contains a link to a (we believe somewhat preprocessed) PDBBind dataset, so data cleaning should be minimal if it is necessary at all.

References:

Blundell TL. Structure-based drug design. Nature. 1996 Nov 7;384(6604 Suppl):23-6. doi: 10.1038/384023a0. PMID: 8895597.

Buttenschoen, M., Morris, G. M., & Deane, C. M. (2023). PoseBusters: AI-based docking methods fail to generate physically valid poses or generalise to novel sequences. Zenodo (CERN European Organization for Nuclear Research). https://doi.org/10.48550/arxiv.2308.05777

Corso, G., Hannes Stärk, Jing, B., Barzilay, R., & Jaakkola, T. S. (2022). DiffDock: Diffusion Steps, Twists, and Turns for Molecular Docking. https://doi.org/10.48550/arxiv.2210.01776

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., Shaw, D. E., Francis, P., & Shenkin, P. S. (2004). 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. https://doi.org/10.1021/jm0306430

Greer, J., Erickson, J. W., Baldwin, J. J., & Varney, M. D. (1994). Application of the Three-Dimensional Structures of Protein Target Molecules in Structure-Based Drug Design. Journal of Medicinal Chemistry, 37(8), 1035–1054. https://doi.org/10.1021/jm00034a001

Harris, C. B., Kieran Didi, Jamasb, A. R., Joshi, C. K., Mathis, S. V., Píetro Lió, & Blundell, T. L. (2023). Benchmarking Generated Poses: How Rational is Structure-based Drug Design with Generative Models? ArXiv (Cornell University). https://doi.org/10.48550/arxiv.2308.07413

Jing, B., Erives, E., Pao-Huang, P., Corso, G., Berger, B. & Jaakkola, T. (2023). EigenFold: Generative Protein Structure Prediction with Diffusion Models. In *arXiv [q-bio.BM]*. arXiv. http://arxiv.org/abs/2304.02198

Kola, I., & Landis, J. (2004). Can the pharmaceutical industry reduce attrition rates? Nature Reviews Drug Discovery, 3(8), 711–716. https://doi.org/10.1038/nrd1470

Liu, Z., Su, M., Han, L., Liu, J., Yang, Q., Li, Y., & Wang, R. (2017). Forging the Basis for Developing Protein–Ligand Interaction Scoring Functions. Accounts of Chemical Research, 50(2), 302–309. https://doi.org/10.1021/acs.accounts.6b00491

Shoichet, B. K., McGovern, S. L., Wei, B., & Irwin, J. J. (2002). Lead discovery using molecular docking. Current Opinion in Chemical Biology, 6(4), 439–446. https://doi.org/10.1016/s1367-5931(02)00339-3

Stärk, H., Octavian-Eugen Ganea, Lagnajit Pattanaik, Barzilay, R., & Jaakkola, T. S. (2022). EquiBind: Geometric Deep Learning for Drug Binding Structure Prediction. ArXiv (Cornell University). https://doi.org/10.48550/arxiv.2202.05146

Stärk, H., Jing, B., Barzilay, R., & Jaakkola, T. S. (2023). Harmonic Self-Conditioned Flow Matching for Multi-Ligand Docking and Binding Site Design. ArXiv (Cornell University). https://doi.org/10.48550/arxiv.2310.05764

Yu, Y., Lu, S., Gao, Z., Zheng, H., & Ke, G. (2023). Do Deep Learning Models Really Outperform Traditional Approaches in Molecular Docking? *arXiv* 2023. DOI arXiv:2302.07134v3.