Midcourse Report

Abstract
Molecular docking accelerates drug discovery by quickly screening many candidate molecules. Deep learning docking methods have the flexibility of learning the physics of protein-ligand interactions from existing data, making them poised to contribute to drug discovery. While DiffDock and HarmonicFlow represent advances over previous methods, the majority of their predictions still have RMSDs over 2 Å to the crystal structure poses. This work aims to inform further deep learning docking methods development through two approaches: analyzing factors influencing docking performance and investigating novel priors. First, better understanding drivers of deep learning docking success or failure may help improve performance. However, most chemical analyses of deep learning docking results focused more on noting infeasibility than on constructively identifying how to improve performance. There has been limited investigation of which chemical features correlate with RMSD and of how pose physical feasibility influences RMSD. We analyzed the correlation of DiffDock and HarmonicFlow RMSD with ligand, protein, and complex chemical features. For DiffDock, we found that more hydrophilic ligands are associated with higher RMSDs. While average binding site hydropathy is not meaningfully associated with RMSD, binding sites with worse RMSD performance have different amino acid composition than do binding sites with better RMSD performance. We offer some hypothesized chemical explanations. The aforementioned conclusions also hold for HarmonicFlow: however, HarmonicFlow RMSD results also display increased association with ligand size metrics. We hypothesize this is due to the HarmonicFlow prior's encoding only local structure, while DiffDock initializes with a guess that is more feasible in terms of global structure. This emphasizes how further prior development could improve performance. We also plan to conduct feasibility analyses of the docking output poses as well to see how well these methods are able to learn some of the fundamental restrictions on what can be physically designed. We want to see how well the HF method specifically does at understanding these fundamental rules governing interaction as well as see if there is a correlation between RMSD and the physical sensibility of the results. Second, improving (e.g., adding more chemical information to) the initial structure in docking may improve the final pose. HarmonicFlow uses a harmonic prior, placing bonded atoms near each other. However, longer-range information is not incorporated. Introducing long-range information into the initial guess structures could enable the method to focus more on pose refinement, possibly helping improve performance. We plan to analyze HarmonicFlow performance with several different priors incorporating longer-range information, as well as with a less informative Gaussian prior, in order to articulate features of a prior which provide a foundation for enhanced docking success. We found HarmonicFlow underestimates ligand radius of gyration, and the extent of underestimation is associated with docking performance, providing motivation for adding longer-range information to the prior and possibly helping prioritize different candidates for the initial guess structure.

Background
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. While an exciting recent report AlphaFold report [Google DeepMind AlphaFold Team et al., 2023] showed improved performance, as Jeremy discussed in lecture, much is not yet clear (e.g., whether there was a model improvement or a data improvement), and to the best of our knowledge, code is not yet available: while we are aware of this development, we therefore decided to focus on DiffDock and HarmonicFlow, which have code and methodological details available. 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).

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 and curate datasets to enable deep learning docking to have higher predictive power. Motivation for analyzing chemical drivers of deep learning docking success to improve model architecture is nicely illustrated by recent work [Corso et al., 2023] in which fine-tuning DiffDock for different protein domain types improved performance. Additionally, Hannes shared that improving generalization could help increase the methods' impact. 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. In addition to protein domain type, what other aspects of complexes could inform model and data curation advances?

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 significantly employed) would be to view the success or lack thereof from the perspective of particular ligands and pockets. Regarding ligands, 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. Regarding pockets, [Corso et al., 2023] found that different protein domain types had different docking performance both before and after fine-tuning, suggesting protein structure influences docking performance. However, this work did not investigate how other binding site features influence performance. Amino acids have distinct steric and electronic properties, prompting the question of whether binding site amino acid composition can influence performance.

Physical Feasibility and Docking

Papers that have been published recently like the paper by Harris et al analyzing the poses created by generative models to test whether or not they are generating results that actually make sense. It has been shown in this paper and others that sometimes the results that are given by generative models may have a good RMSD, but can create molecules that aren't physically sensible. These molecules actually have physical inconsistencies that can't be measured by the typical loss function. For this reason, additional analysis of the physical forces that are acting on the constituent molecules allows us to better know what the particles are doing.

Inspired by this paper we also set out to see how the generative models we come up with perform on these tests as well as other possible physical constraints on the poses that molecules can take. Specifically we are looking at strain energy and steric clashes in the predicted molecules. Strain energy describes the energy stored in the ligand from conforming to dock with the protein. This has entropy and enthalpy effects that need to be considered, but we can see that generally we want a lower value in this metric. Additionally, we want to look at steric clashes, which are situations in which the generative model predicts that the molecules are so close to one another that their atoms are actually overlapping.

We decided to use posebusters [Buttenschoen et al., 2023] to analyze the feasibility of the generated molecules that we come up with from our generative model. The paper focuses on the physical infeasibility of the molecules that are generated from physics based models, and diffusion based models. So we plan to test how it works on a flow based generative model that we will be looking at in the second part. In the paper, we see that the diffusion based models

performed very poorly and we are going to run the same analysis on DiffDock molecules and also look at how these same metrics look on the HF based molecules.

Priors

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. 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.

Results and Discussion

Chemical drivers

(A) Protein and ligand feature correlation with docking performance

The initial focus of the chemical drivers analysis was on DiffDock, due to Hannes Stärk's explanation that it represents a "polished docking pipeline". Hannes very kindly shared results from a DiffDock and HarmonicFlow run for PDBBind [Liu et al., 2017] structures. We will first focus on DiffDock results. For DiffDock, the highest confidence RMSD was obtained from each complex. From each complex identity, features describing the ligand, protein, and complex could be computed from the processed PDBBind data provided in the HarmonicFlow GitHub link. RDKit [Landrum, 2022] and MDAnalysis [Gowers et al., 2016, Michaud-Agrawal et al., 2011] were used to aid in feature generation. Correlation between the RMSD and chemical features of the protein, ligand, and protein-ligand complex could then be analyzed.

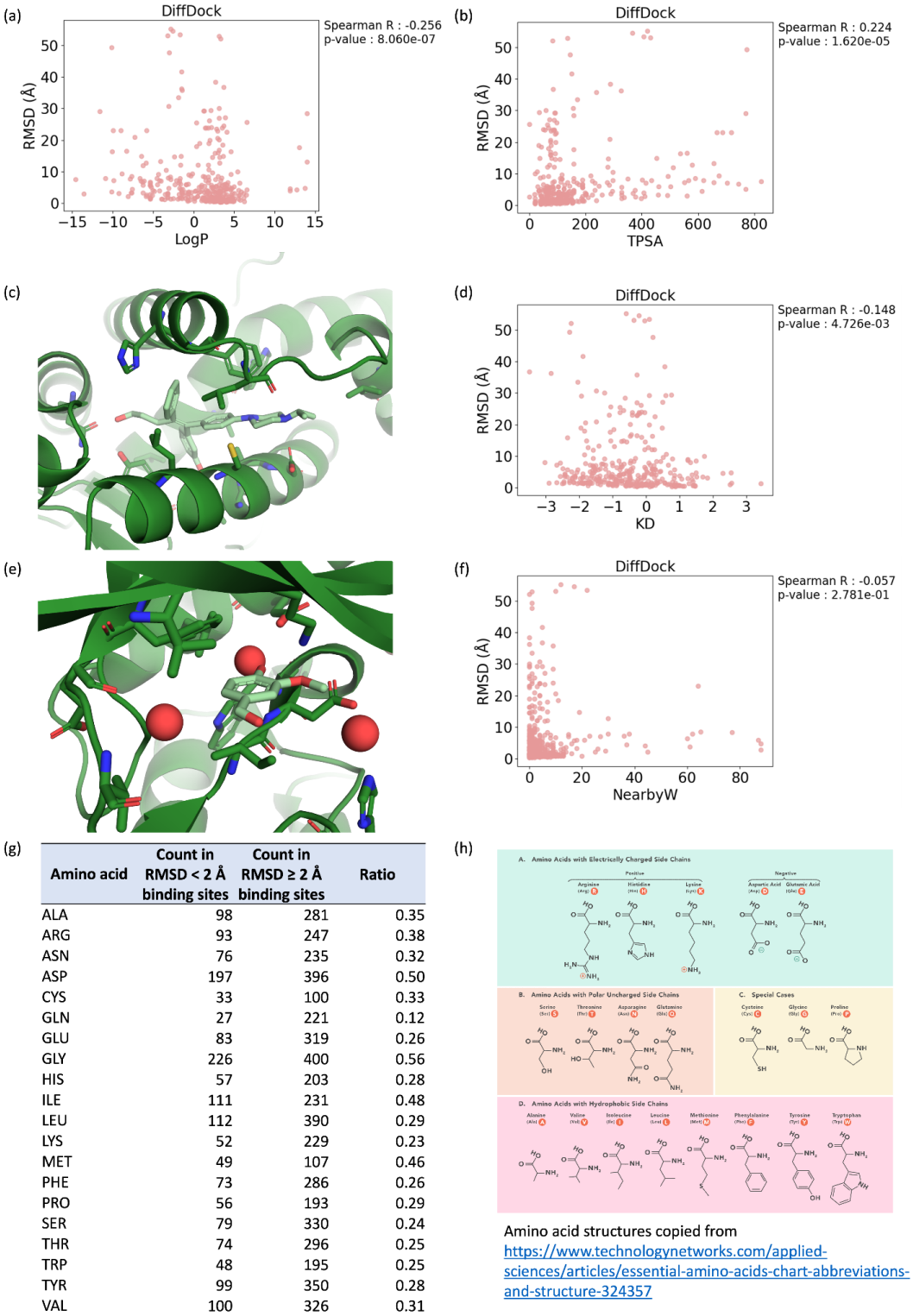
Metrics of ligand polarity, namely the logP (preference for polar vs nonpolar solvent [Poole et al., 2003]) and topological polar surface area (TPSA) [Ertl et al., 2000]) showed some correlation with RMSD (Figure 1a, b). More hydrophilic ligands (more negative/less positive logP and larger TPSA) have worse docking performance. This is in agreement with previous work [Harris et al., 2023] reporting that hydrogen bonds are not always well replicated by machine learning docking. Other ligand metrics considered have lower (absolute value) Spearman Rs and higher p-values in the RMSD correlation analysis (Figure 2).

Seeing as hydrophilicity may influence docking performance from a ligand perspective, we also investigated the hydrophilicity from the protein binding site and water perspective. For each complex, residues with a heavy atom-heavy atom distance to a ligand atom of 3.7 Å or less were identified (Figure 1c). The average Kyte-Doolittle hydropathy [Kyte et al., 1982] of each binding site was computed. This quantity was not very meaningfully correlated with RMSD performance, though the direction of the association is that more nonpolar binding sites (higher average Kyte-Doolittle hydropathy) have better performance (Figure 1d). The ligand

hydrophilicity finding also led to development of a hypothesis that bound waters may influence docking performance. Speculating, more polar ligands could possibly have more waters nearby them and the protein in the crystal structures, and this could lead to worse RMSD performance because the waters are removed before docking. To investigate this hypothesis, for each complex, the count of water molecules within 3.7 Å of both the protein and the ligand was determined (Figure 1e). (The implementation is preliminary: the crystal structure needed to be analyzed, and while the relevant chains could be selected from the PDBBind structure, we could not immediately determine how to only select the one ligand molecule in PDBBind. Thus, if there are multiple instances of the same ligand near the PDBBind chains analyzed, waters associated with either ligand as well as with the protein would be identified.) Binding site water count was not correlated with RMSD performance (Figure 1f). This highlights the remarkable power of deep learning docking to harvest insight from data. It implies that even when there are bound waters in the crystal structure- which do not appear in the protein structure used in docking- deep learning docking can infer that a water may be present and adjust the ligand pose. As Jeremy Wohlwend explained, the protein embeddings capture protein hydrophilicity properties, enabling the model to have an understanding of what constitutes a hydrophilic binding site where water may be present.

While the protein binding site average hydropathy did not drastically influence docking performance, this is a summary statistic and could obscure more nuanced influences of amino acid composition on RMSD results. To probe for possible such effects, we segmented the complexes by whether they had a docking RMSD above or below 2 Å. We then tabulated frequencies of each amino acid in the binding sites of the complexes with RMSDs below and 2 Å (Figure 1g). An initial chi square test attempt failed because counts of amino acids in the two RMSD groups were different. We thank Jeremy Wohlwend for advising on computing histograms for each RMSD group, which led to the test's succeeding. However, because the normalized frequencies were so low, this led to a large p-value, and scaling to higher frequencies deflated the p-value, in a way which made us hesitant to work with the transformed data given the arbitrariness in how to scale. Thus, as a preliminary solution, we use the Kolmogorov-Smirnov test and find a p-value of $1.33*10^{-6}$, suggesting binding sites with better RMSD performance have a different amino acid composition than binding sites with worse RMSD performance. We thus computed a ratio of each amino acid's frequency in binding sites with RMSD under 2 Å and in binding sites with RMSD over 2 Å. Although this ratio is not neatly correlated with amino acid Kyte-Doolittle hydropathy or side chain heavy atom count (Figure A1), there exist possible chemical interpretations of the results. One observation is that flexibility can affect relative amino acid presence in the different binding sites. Glycine has the largest ratio, indicating it is the amino acid most enriched in the binding sites with good RMSD performance. Because DiffDock coarse-grains the protein, glycine, without a heavy atom side chain (Figure 1h), should be relatively easy to model: because there are no flexible side chain heavy atoms to model, there is minimal risk of predicting a pose which has a clash or an interaction which implicitly assumes a side chain is in the wrong place. Glutamine has a lower ratio than asparagine, and glutamate has a lower ratio than aspartate. Glutamine and glutamate have an extra methylene compared to asparagine and aspartate (Figure 1h), leading to more flexible side chains whose interactions with ligands may not be predicted as well by a

coarse-grained model. However, noting how side chain heavy atom count is not meaningfully correlated with RMSD performance, it is important to note there are also additional factors at play. One such factor could be heterogeneity of heteroatoms. Glutamine has a lower ratio than glutamate, and asparagine has a lower ratio than aspartate. The carboxylates contain a side chain with two negatively charged chemically identical oxygens (Figure 1h). Their interactions may be easier to predict in a coarse-grained model than amino acids with an amide side chain, featuring both hydrogen bond donating and accepting atoms, possibly requiring more chemical detail in their depiction.



(g)

Amino acid	Count in RMSD < 2 Å binding sites	Count in RMSD ≥ 2 Å binding sites	Ratio
ALA	98	281	0.35
ARG	93	247	0.38
ASN	76	235	0.32
ASP	197	396	0.50
CYS	33	100	0.33
GLN	27	221	0.12
GLU	83	319	0.26
GLY	226	400	0.56
HIS	57	203	0.28
ILE	111	231	0.48
LEU	112	390	0.29
LYS	52	229	0.23
MET	49	107	0.46
PHE	73	286	0.26
PRO	56	193	0.29
SER	79	330	0.24
THR	74	296	0.25
TRP	48	195	0.25
TYR	99	350	0.28
VAL	100	326	0.31

(h)

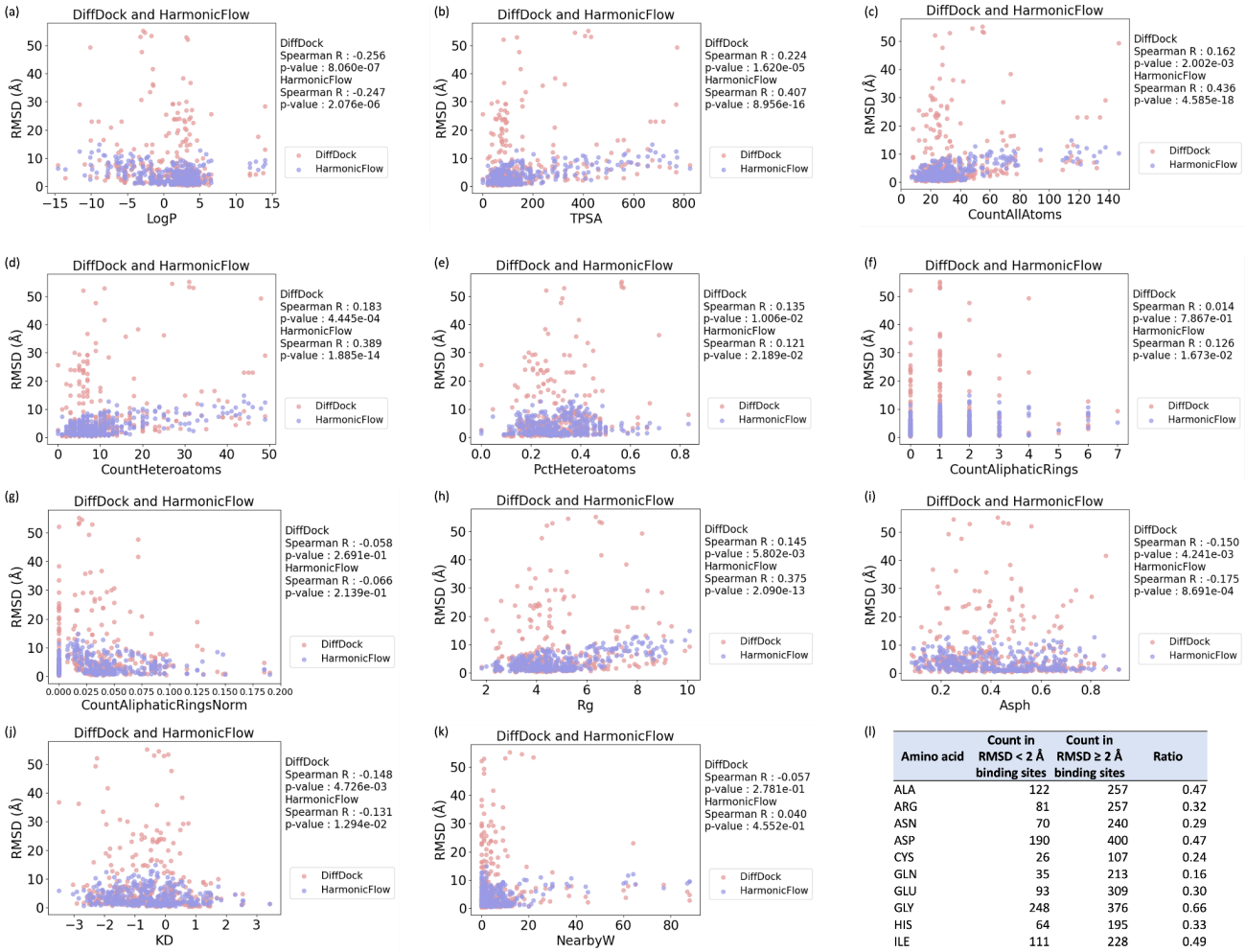
Amino acid structures copied from https://www.technologynetworks.com/applied-sciences/articles/essential-amino-acids-chart-abbreviations-and-structure-324357

Figure 1. DiffDock chemical feature analysis (a) Correlation of ligand logP with docking RMSD performance. (b) Correlation of ligand topological polar surface area with docking RMSD performance. (c) Schematic of selection of binding site residues for hydropathy analysis. PDB ID 6A6K [Kim et al., 2019], from processed PDBBind [Liu et al., 2017] linked on the FlowSite

GitHub. (d) Correlation of average binding site residue Kyte-Doolittle hydropathy with docking RMSD performance. (e) Schematic of identification of binding site waters, with waters shown in red spheres. PDB ID 6HOU [Cozza et al., 2020], from RCSB. (f) Correlation of binding site water count with docking RMSD performance. (g) Counts of amino acids in binding sites with better and worse RMSD performance, and the ratio of each amino acid's count in binding sites with better RMSD performance to the amino acid's count in binding sites with worse RMSD performance. (h) Amino acid structures, copied from a website.

We are grateful to Hannes for suggesting to focus on DiffDock due to its being a more "polished docking pipeline" (Slack exchange). Thus, we initially investigated DiffDock in order to comment on the state of chemical drivers of deep learning docking. However, then we looked into HarmonicFlow drivers as well, in order to compare. We used a HarmonicFlow inference output from Hannes, averaging over RMSDs to obtain one value for each complex. Similar conclusions hold of ligand polarity influencing results (Figure 2a and 2b), binding site polarity influencing results less than ligand polarity (Figure 2j), bound waters not influencing results (Figure 2k), binding site amino acid composition differing by RMSD performance (Figure 2l), and possible connections of these differences to chemical trends. However, the results differ in that ligand size influences results more for HarmonicFlow than for DiffDock. This can be seen when comparing correlations for the two methods of RMSD and ligand heavy atom count (Figure 2c) and ligand radius of gyration (Figure 2h).

One possible explanation can be found in the construction of the initial structure for the two methods. DiffDock begins with a RDKit-generated conformer [Corso et al., 2022]. Meanwhile, HarmonicFlow uses a harmonic prior [Stärk et al., 2023]. Thus, the DiffDock initial structure encodes both global and local ligand geometric information. Meanwhile, because HarmonicFlow focuses on individual bonds, it does not encode information beyond local bonding environments for each atom. One hypothesis is that the loss of this long-range information for the initial HarmonicFlow structures is more deleterious for larger ligands: because they are larger, information beyond the local scale could be more relevant.



(l)

Amino acid	Count in RMSD < 2 Å binding sites	Count in RMSD ≥ 2 Å binding sites	Ratio
ALA	122	257	0.47
ARG	81	257	0.32
ASN	70	240	0.29
ASP	190	400	0.47
CYS	26	107	0.24
GLN	35	213	0.16
GLU	93	309	0.30
GLY	248	376	0.66
HIS	64	195	0.33
ILE	111	228	0.49
LEU	147	353	0.42
LYS	65	213	0.31
MET	60	93	0.65
PHE	81	269	0.30
PRO	83	166	0.50
SER	84	321	0.26
THR	103	264	0.39
TRP	34	207	0.16
TYR	127	316	0.40
VAL	107	315	0.34

Figure 2. HarmonicFlow and DiffDock chemical feature analysis. (a-k) Correlations of RMSD with (a) ligand logP percent of ligand heavy atoms which are heteroatoms (b) ligand topological polar surface area ligand heteroatom count (c) ligand heavy atom count (analysis was also done in [Corso et al., 2022] for DiffDock) (d) ligand heteroatom count (e) percent of ligand heavy atoms which are heteroatoms (f) ligand aliphatic ring count (g) ligand aliphatic ring count to heavy atom count ratio (g) ligand logP (h) ligand radius of gyration (i) ligand asphericity [Todeschini et al., 2003] (j) average binding site residue Kyte-Doolittle hydropathy (k) binding site water count. (l) Counts of amino acids in binding sites with better and worse HarmonicFlow RMSD performance, and the ratio of each amino acid's count in binding sites with better RMSD performance to the amino acid's count in binding sites with worse RMSD performance.

(B) Physical feasibility and docking performance

We are still working on getting results from this analysis

Priors

The previous finding that HarmonicFlow results display an association with ligand size more than DiffDock results do highlights how encoding longer-range information into the HarmonicFlow prior may help improve performance.

RDKit is a well-known cheminformatics software. The first approach to constructing the prior is to directly infer structures from RDKit. We use the 2D chemical graph of the ligand to create the 3D structure using RDKit Chem.rdDistGeom.EmbedMolecule. We then use UFF to optimize a ligand's prior structure using rdForceFieldHelpers.UFFOptimizeMolecule. This structure serves as the initial prior for ligand poses. We would also perform the same steps as above using RDKit, and then add some noise to the prior structures.

Understanding HarmonicFlow 3D performance can also facilitate more rational prior selection. We compared- currently very preliminarily- HarmonicFlow inference output (from Hannes) ligand poses to PDBBind ligand poses on two 3D metrics: radius of gyration and asphericity (Figure 3). For each ligand, we found the difference of the PDBBind pose metric and the inference metric (PDBBind minus inference). Harmonic Flow is providing positive radius of gyration differences almost always, and positive differences show some correlation with worse docking performances. This means that HarmonicFlow is underestimating radius of gyration, creating ligand poses that are more folded on themselves and less extended than the corresponding PDBBind ligand poses.

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 motivates use of 3D RDKit conformers which should be less folded-in than the current HarmonicFlow starting structures. This finding could also help guide prior development because it suggests selecting a conformer with a larger radius of gyration for the initial HarmonicFlow prior could aid results because currently HarmonicFlow inference is underestimating the radius of gyration.

We would also use consensus distances to construct the prior structure. We first generate 10 conformers structures in RDKit. We then get the pairwise distance matrix for each conformers. We would take the average distances or the distances with least variation and reconstruct the 3D structure as prior for the mode. We may also try to use deep learning or diffusion methods for conformers generation.

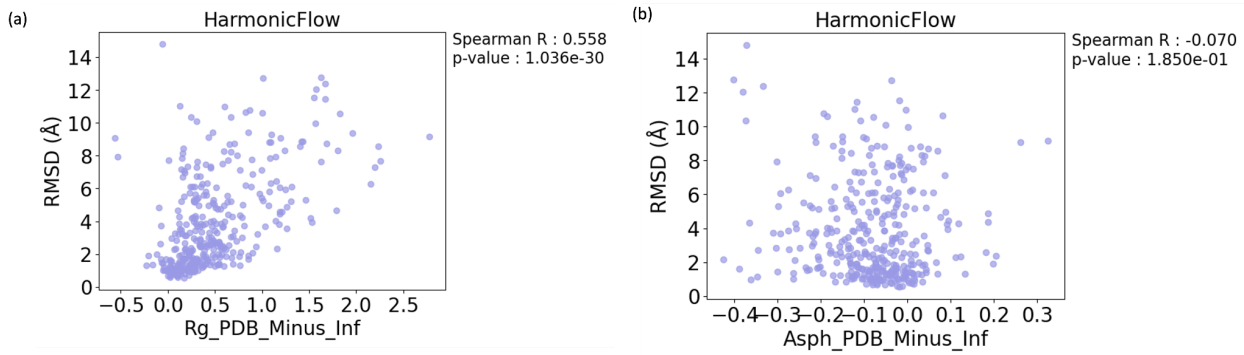


Figure 3. Very preliminary scatterplot of RMSD performance versus difference in 3D metrics between the PDBBind structure ligand and the inference pose (PDBBind minus inference) for (a) radius of gyration (b) asphericity [Todeschini et al., 2003].

Future Goals

To add in final report

Comparison with Original Proposal

To add in final report

Commentary on Experience

To add in final report

Division of Labor

To add in final report

Acknowledgements:

We are very grateful to Professor Regina Barzilay, Professor Manolis Kellis, Mr. Hannes Stärk, and Mr. Jeremy Wohlwend for helpful discussions which contributed to our understanding of the topic and our developing this report. We thank Hannes for being our mentor, sharing the docking output and Gaussian prior weights, and advising on project ideas and prior development and on how to run HarmonicFlow code. We thank Hannes for sharing the DiffDock and HarmonicFlow results which we analyze in this report. We thank Jeremy for very helpful office hours discussions and project advice, including using the Spearman R, communicating result significance, conformer generation approaches, and prior development ideas. We are very grateful to Dr. Abba Leffler for making us aware of the work of [Buttenschoen et al., 2023].

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

Corso, G., Deng, A., Polizzi, N., Barzilay, R., & Jaakkola, T. (2023) The Discovery of Binding Modes Requires Rethinking Docking Generalization NeurIPS 2023 Generative AI and Biology (GenBio) Workshop https://openreview.net/forum?id=FhFglOZbtZ

Cozza, G., Zonta, F., Dalle Vedove, A., Venerando, A., Dall'Acqua, S., Battistutta, R., Ruzzene, M., and Lolli, G. (2020). Biochemical and cellular mechanism of protein kinase CK2 inhibition by deceptive curcumin. FEBS J, 287: 1850-1864. https://doi.org/10.1111/febs.15111

Ertl, P., Rohde, B., & Selzer, P. (2000). Fast Calculation of Molecular Polar Surface Area as a Sum of Fragment-Based Contributions and Its Application to the Prediction of Drug Transport Properties Journal of Medicinal Chemistry, 43(20), 3714-3717. https://doi.org/10.1021/jm000942e

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

Google Deepmind AlphaFold Team & Isomorphic Labs Team. (2023). Performance and Structural Coverage of the Latest, In-Development AlphaFold Model, https://storage.googleapis.com/deepmind-media/DeepMind.com/Blog/a-glimpse-of-the-next-generation-of-alphafold/alphafold_latest_oct2023.pdf

Gowers, R. J., Linke, M., Barnoud, J., Reddy, T. J. E. Melo, M. N., Seyler, S. L., Dotson, D. L., Domanski, J., Buchoux, S., Kenney, I. M., & Beckstein, O. (2016). MDAnalysis: A Python package for the rapid analysis of molecular dynamics simulations. In S. Benthall and S. Rostrup, editors, Proceedings of the 15th Python in Science Conference, pages 98-105, Austin, TX, SciPy. https://doi.org/10.25080/Majora-629e541a-00e

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

Kim, J., Song, J., Ji,H. D., Yoo, E. K., Lee, J.-E., Lee, S. B., Oh, J. M., Lee, S., Hwang, J. S., Yoon, H., Kim, D.-S., Lee, S.-J., Jeong, M., Lee, S., Kim, K.-H., Choi, H.-S., Lee, S. W., Park, K.-G., Lee, I.-K., Kim, S. H., Hwang, H., Jeon, Y. H., Chin, J. & Cho, S. J. (2019). Discovery of Potent, Selective, and Orally Bioavailable Estrogen-Related Receptor-γ Inverse Agonists To Restore the Sodium Iodide Symporter Function in Anaplastic Thyroid Cancer. Journal of Medicinal Chemistry, 62(4), 1837-1858. https://doi.org/10.1021/acs.jmedchem.8b01296

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

Kyte, J. & Doolittle, R. F. (1982) A simple method for displaying the hydropathic character of a protein. J Mol Biol., 157(1), 105-132. https://doi.org/10.1016/0022-2836(82)90515-0

Landrum, G. 2022. "RDKit." https://www.rdkit.org/.10.5281/zenodo.6961488

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

Michaud-Agrawal, N., Denning, E. J., Woolf, T. B., & Beckstein, O. (2011). MDAnalysis: A Toolkit for the Analysis of Molecular Dynamics Simulations. J. Comput. Chem. 32, 2319–2327. https://doi.org/10.1002/jcc.21787

Poole, S. K., & Poole, C. F. (2003) Separation Methods for Estimating Octanol-Water Partition Coefficients. Journal of Chromatography B, 797(102), 3-19. https://doi.org/10.1016/j.jchromb.2003.08.032

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

Todeschini, R., & Consonni, V. (2003). Chapter VIII.2: Descriptors from Molecular Geometry in Handbook of Cheminformatics Descriptors for Chemical Compounds. Ed. J. Gasteiger. 1004-1033. https://doi.org/10.1002/9783527618279.ch37

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.

Appendix

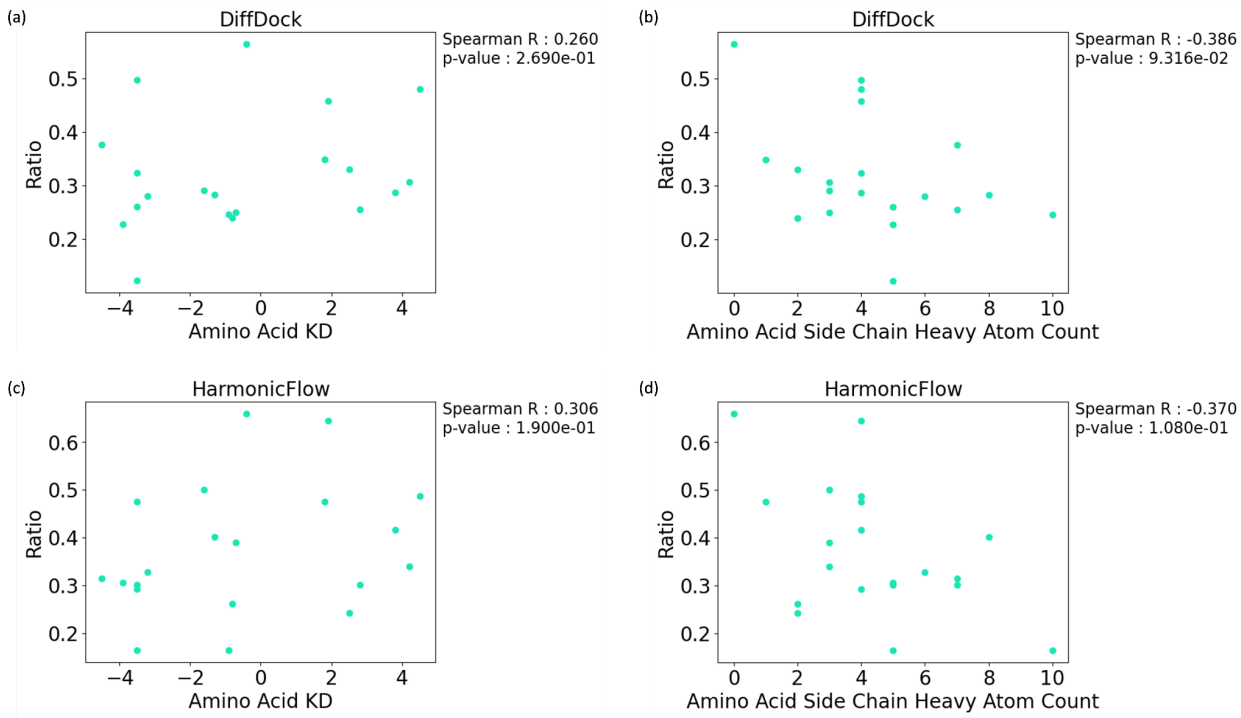


Figure A1. Scatterplot of ratio of each amino acid's instances in low RMSD complex binding sites to instances in high RMSD complex binding sites, plotted against (a) amino acid Kyte-Doolittle hydropathy for DiffDock (b) amino acid side chain heavy atom count for DiffDock (c) amino acid Kyte-Doolittle hydropathy for HarmonicFlow (d) amino acid side chain heavy atom count for HarmonicFlow