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

Sam Huang, Sammy Mustafe, Mo Oyewole, Dina Sharon

**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 and larger and more flexible ligands are very slightly to somewhat (depending on the metric) associated with higher RMSDs. 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 with more global information. This emphasizes how further prior development could improve performance. We conducted 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 DiffDock 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. We did this by looking at the Strain energy and steric clashes of the protein ligand poses .We also did the same analysis to Harmonic flow to see how it performed in comparison . 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 using RDKit: inferring the 3D structure using the 2D chemical graph, leveraging multiple conformers with a universal pairwise distance STDEV threshold to create consensus distances with consistent long-range distances for the prior, 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*

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. 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 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, a 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*

Strain energy describes the energy stored in the ligand and the protein from conforming to dock with one another. Even though, there are a variety of physical effects that can come from various binding energy levels, we can generally say that a lower energy level is better, since it can affect binding afinity. A steric clash is when the two neutral atoms are so close together that their van der Waals radii overlap. This is an issue since the van der Waals radius can be thought of an an imaginary hard sphere around each atom that acts as the volume that it takes up. No other molecule can enter into this space so a pose in which the van der Waals radii of two neutral atoms overlapping is very energetically unfavorable.

We used the Posecheckers [Harris et al., 2023] tool from the paper to analyze the feasibility of the generated molecules that we come up with from our generative model.

*Priors*
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. A useful prior will improve the model efficiency and improve the quality of docking. Developing an informative prior will not only improve Harmonic Flow performance but also may be applied to other diffusion/flow models of different tasks.

**Results and Discussion**

*Chemical drivers*
*(A) Datasets and data cleaning and processing*

Hannes Stärk very kindly shared results from a DiffDock and HarmonicFlow run. Both were for docking to PDBBind [Liu et al., 2017] structures. For the HarmonicFlow run, the last pose was extracted from the …xt.pdb file, representing the end of the flow matching trajectory. For DiffDock, the highest confidence RMSD was obtained for each complex. For HarmonicFlow, RMSD was found by averaging over RMSDs to obtain one value for 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 Zenodo link. RDKit [Landrum, 2022] and MDAnalysis [Gowers et al., 2016, Michaud-Agrawal et al., 2011] were used to aid in feature generation.

Data cleaning was conducted differently for ligand, protein-ligand interaction, bound water, and clash analyses, in order to obtain all relevant data points for an analysis without including irrelevant data points. For all analyses, complexes were removed if they were only in the test set for one of the two docking methods, if they had two or more ligand residue names (because this could mean the ligand is broken up physically), or if the Protein Data Bank (PDB) entry no longer existed. This was the extent of cleaning for the ligand-only analysis and clash analysis. For the protein-ligand interaction analyses, complexes were additionally removed if there were no protein-ligand interactions in the PDBBind structure within the cutoff distance. For the bound water analysis, complexes were additionally (to the all analyses removals) removed if the PDBBind ligand residue name was a protein residue (since this would complicate finding waters near the ligand residue in the structure), if there were two or more identical ligand molecules in the reference PDB crystal structure (not the cleaned PDBBind file, having multiple molecules would complicate finding count of waters near one molecule and the protein), or if there were not any residues named HOH in the PDB file (because it was not immediately clear whether there were no crystallographic waters or there were crystallographic waters by another residue name, though this could be resolved in the future by manual visualization or advanced parsing). For the protein-ligand analysis, if the PDBBind ligand residue name was a protein residue name, the ligand name was changed to “LIG” (since this would complicate finding protein residues near the ligand residue). While the clash analysis underwent the checks for all analyses and also checks for whether any protein-ligand interactions were present (now in the docked poses), no complexes were removed due to these checks.

*(B) 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”. We will first focus on DiffDock results. 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 a bit of correlation with RMSD (Figure 1a, b- logP was only slightly correlated). 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. We also found atom count, rotatable bond count, and radius of gyration are a bit correlated with RMSD (Figure 1c,d,e). Larger ligands - in more atoms and more “spread” (bigger radius of gyration) perform

worse. More flexible ligands perform worse. This agrees with [Corso et al., 2022]. We checked the distribution of logP, TPSA, all atom count, rotatable bond count, and radius of gyration in the test and training sets and found each metric's distribution is similar in the two sets (Figure A2). Therefore as discussed with Jeremy (paraphrasing his helpful guidance), these properties not only are associated with performance, but also this association is not solely due to differences in the testing and training set. Additional metrics' RMSD correlations are shown in Figure A1.

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 1f). The average Kyte-Doolittle hydropathy [Kyte et al., 1982] of each binding site was computed. This quantity was only slightly 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 1g). 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 1h). (The implementation is preliminary, with much data excluded now) Binding site water count was not very correlated with RMSD performance (Figure 1i). 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 1j). While statistical analysis (complicated by different counts in the different RMSD groups) is required to be certain, inspection suggests 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. This ratio is not neatly correlated with amino acid Kyte-Doolittle hydropathy and is a bit correlated with side chain heavy atom count (Figure A3). 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 one of the largest ratios, indicating it is an amino acid enriched in the binding sites with good RMSD performance. Because DiffDock coarse-grains the protein, glycine, without a heavy atom side chain (Figure 1k), 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 1k), 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 1k). 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.

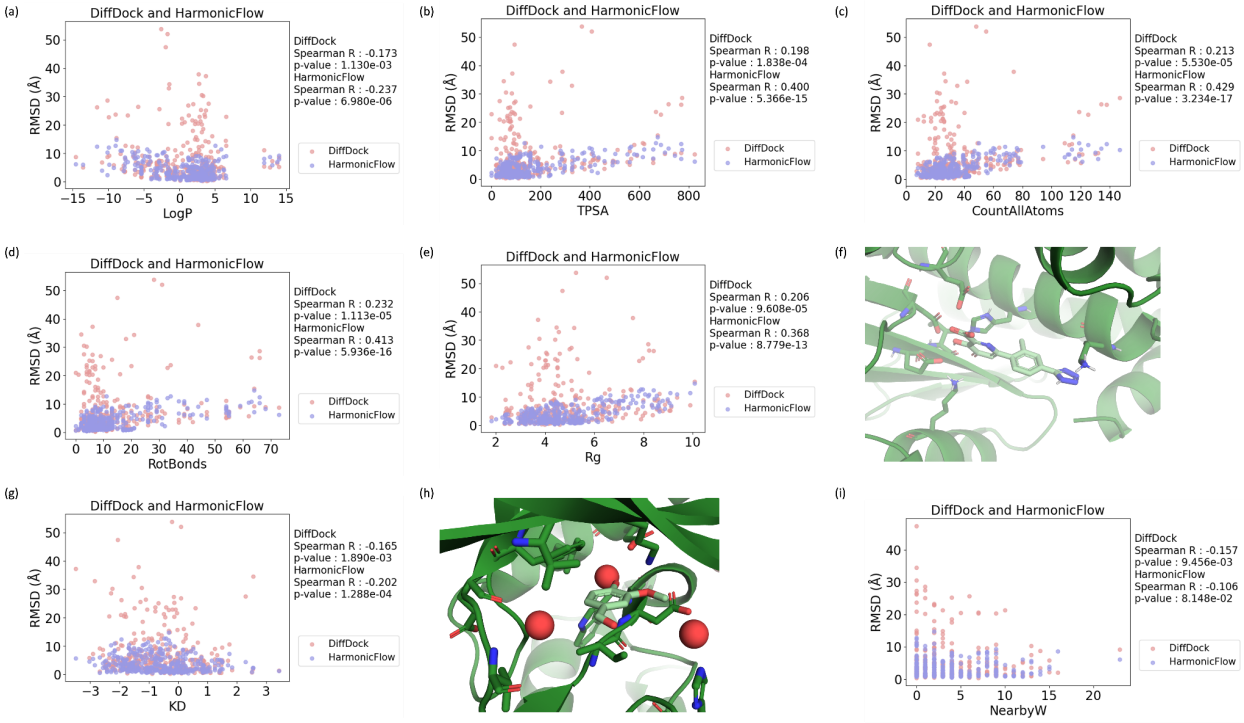
We are grateful to Hannes for suggesting to focus on DiffDock due to its being a more "polished docking pipeline" (Slack exchange). Thus, we initially investigated DiffDock in order to comment on the state of chemical drivers of deep learning docking. However, then we looked into HarmonicFlow drivers 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 1a and 1b), bound waters not influencing results (Figure 1i), binding site amino acid composition differing by RMSD performance (Figure 1j), 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, rotatable bond count, and radius of gyration (Figure 1c,d,e)

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.

It is important to note that the chemical drivers are almost definitely correlated with each other, which might confound some of the signal's source. For instance, TPSA reports on size and polarity, we believe, and it is more correlated than logP which we think is more size-independent (with a higher correlation increase for TPSA between DiffDock and HarmonicFlow than for logP). Future work could focus on disentangling these correlations.

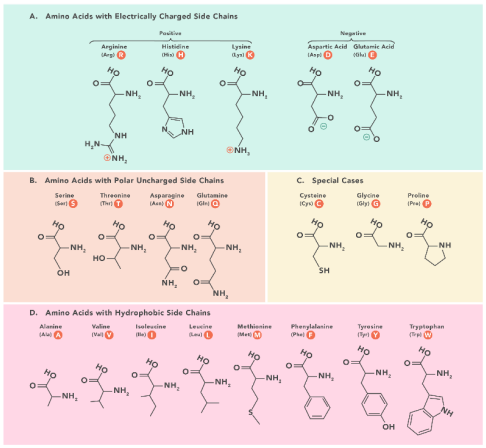
One incidental finding during iterations of data cleaning is that removing ligands which are seemingly more peptide-like (having peptide residue names) alters some results (Figure A4). This raises the question of whether there may be sub-populations of the test set where chemical drivers are more or less influential, which is a possible avenue for future work.




(j)

| Amino acid | *DiffDock* Count in RMSD < 2 Å binding sites | *DiffDock* Count in RMSD ≥ 2 Å binding sites | *DiffDock* Ratio | *HarmonicFlow* Count in RMSD < 2 Å binding sites | *HarmonicFlow* Count in RMSD ≥ 2 Å binding sites | *HarmonicFlow* Ratio |
|---|---|---|---|---|---|---|
| ALA | 49 | 120 | 0.41 | 75 | 94 | 0.80 |
| ARG | 49 | 176 | 0.28 | 63 | 162 | 0.39 |
| ASN | 44 | 141 | 0.31 | 64 | 121 | 0.53 |
| ASP | 115 | 296 | 0.39 | 177 | 234 | 0.76 |
| CYS | 16 | 57 | 0.28 | 25 | 48 | 0.52 |
| GLN | 16 | 106 | 0.15 | 27 | 95 | 0.28 |
| GLU | 41 | 194 | 0.21 | 77 | 158 | 0.49 |
| GLY | 140 | 332 | 0.42 | 236 | 236 | 1.00 |
| HIS | 30 | 127 | 0.24 | 55 | 102 | 0.54 |
| ILE | 72 | 135 | 0.53 | 103 | 104 | 0.99 |
| LEU | 60 | 190 | 0.32 | 107 | 143 | 0.75 |
| LYS | 32 | 121 | 0.26 | 50 | 103 | 0.49 |
| MET | 26 | 51 | 0.51 | 30 | 47 | 0.64 |
| PHE | 43 | 166 | 0.26 | 72 | 137 | 0.53 |
| PRO | 27 | 117 | 0.23 | 68 | 76 | 0.89 |
| SER | 47 | 188 | 0.25 | 63 | 172 | 0.37 |
| THR | 47 | 144 | 0.33 | 79 | 112 | 0.71 |
| TRP | 23 | 140 | 0.16 | 33 | 130 | 0.25 |
| TYR | 52 | 275 | 0.19 | 103 | 224 | 0.46 |
| VAL | 62 | 157 | 0.39 | 92 | 127 | 0.72 |

(k)

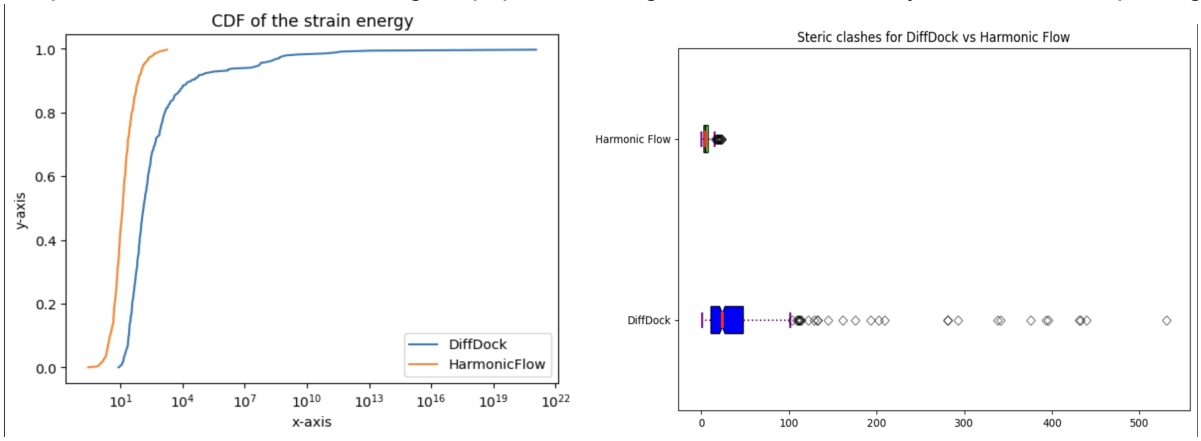



Amino acid structures copied from https://www.technologynetworks.com/applied-sciences/articles/essential-amino-acids-chart-abbreviations-and-structure-324357

**Figure 1.** *HarmonicFlow andDiffDock chemical feature analysis(a-e) Correlations of RMSD with (a) ligand logP (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 count of rotatable bonds (analysis was also done in [Corso et al., 2022] for DiffDock) (e) ligand radius of gyration (f) Schematic of selection of binding site residues for hydropathy analysis. PDB ID 6E6W [Credille et al., 2019], from processed PDBBind [Liu et al., 2017] linked on the FlowSite GitHub. Nearby residues to the ligand shown as sticks. (g) Correlation of average binding site (from PDBBind ligand pose, not docked ligand pose) residue Kyte-Doolittle hydropathy with docking RMSD performance. (h) Schematic of identification of binding site waters, with waters shown in red spheres. PDB ID 6HOU [Cozza et al., 2020], from RCSB. (i) Correlation of binding site water count with docking RMSD performance. (j) 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. (k) Amino acid structures, copied from a website.*

*(C) Physical feasibility and docking performance*

For this analysis we used the Posecheckers package [Harris et al.] to analyze the performance of DiffDock and Harmonic Flow in predicting physically feasible protein-ligand binding interactions by analyzing the strain energy and steric clashes of the generated poses. We went through the entire test set of both models to see how well the molecules it generated on that set performed. We also do some comparison to the results of the original paper and the generative models they used to test the package.




(a) (b)

**Figure 2:** *This compares the Steric clashes and Strain Energies for DiffDock and HarmonicFlow. We can see that Harmonic Flow seems to perform much better in both metrics. Note that graph a is Logarithmic*

When looking at the strain energy (figure 2a) we found that on average it out performed the diffusion based models from the paper with a median strain energy of about 129 kcal/mol. We can also see that the distribution of strain energies seems to have been more concentrated in a specific region with all the energies falling mainly in the range of 10-1000 kcal/mol, but we do have some energies that explode up to really high values. One seems to have even gone all the way up to 10^21 kcal/mol, but this seems to be an error. This could be a result of the proteins and ligands that we were analyzing in our initial dataset being different from the ones in the paper we initially looked at, but this seems to imply that the orientations of the protein and ligand poses makes more sense with Diffdock that with the other diffusion based docking models. And looking at the curve specifically DiffDock also seems to have outperformed many of the physics based simulations meaning that Diffdock does a surprisingly effective job at finding energetically favorable orientations. The most surprising part was the fact that Harmonic Flow outperformed even Diffdock. The highest energy configuration has a strain energy of 1784 kcal/mol which is not much higher than the median energy of the diffusion based models from the paper and we can see that the median energy for the Harmonic flow trainer was 12.675 kcal/mol. This is an amazingly low energy even beating some of the physics based models from the paper. This seems to be a little bit anomalous, so in the future we could attempt to do further analysis to see if we can get better results, but it seems the Harmonic flow is amazing at generic energetically favorable protein ligand configurations

When we look at the steric clashes that are in the generated poses. We find that Diffdock seems to perform much worse than all the other models (figure 2b). It has a median of 23 steric clashes and some of the generated poses have almost 500 steric clashes. In comparison to the models in the paper we see that most of the models there have a median of around 10 steric clashes and very few cases that have close to 500 clashes. Now again, we are using a different dataset, so we can't come up with a very good set of things to learn from this, but it indicates that Diffdock might be very good for getting an idea of a feasible pose energetically, but may need quite a bit of tweaking to put the molecules in their specific correct places, which is very in-line with the beliefs of one of the co-collaborators in creating the tool Hannes, so this is a sensible perspective to some up with. The most surprising thing is how few steric clashes we find from Harmonic flow. HF has a median of 5 steric clashes which is almost nothing DiffDock. In fact the highest number of clashes we get from Harmonic flow is actually 23 which is the same as the median number of steric clashes we get from DiffDock. This is very unique because it indicates that Harmonic flow also generates molecules that aren't overlapping with one another very

much. These are very impressive results which makes us more excited about the prospect of utilizing Harmonic flow to generate good molecules.

Another analysis we did with the steric clashes was to find the correlation between the steric clashes and the RMSD of the molecule. We wanted to make sure that we weren't just finding which molecules were the most incorrect by looking at the steric clashes and we found that there was a correlation value of 0.27 between the two indicating that these steric clashes are actually meaningful to look at outside of the RMSD. Additionally, there was little correlation between the RMSD and the strain energy. The Pearson correlation was about 0.18. From this we can conclude that we are performing meaningful analysis.

In all we can conclude that Harmonic flow is a very promising tool in terms of generating molecule positions that make sense. This means that we can feel confident in using it for the rest of the paper to make good docking poses.

***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 HarmonicFlow inference output ligand poses to PDBBind ligand poses on two 3D metrics: radius of gyration and asphericity (Figure 3). For each ligand, we found percent error comparing the PDBBind pose metric and the inference metric. For radius of gyration, we found percent error (as well as the raw difference) shows some correlation with worse docking performance. HarmonicFlow is underestimating the radius of gyration, creating ligand poses that are more folded on themselves and less extended than the corresponding PDBBind ligand poses.

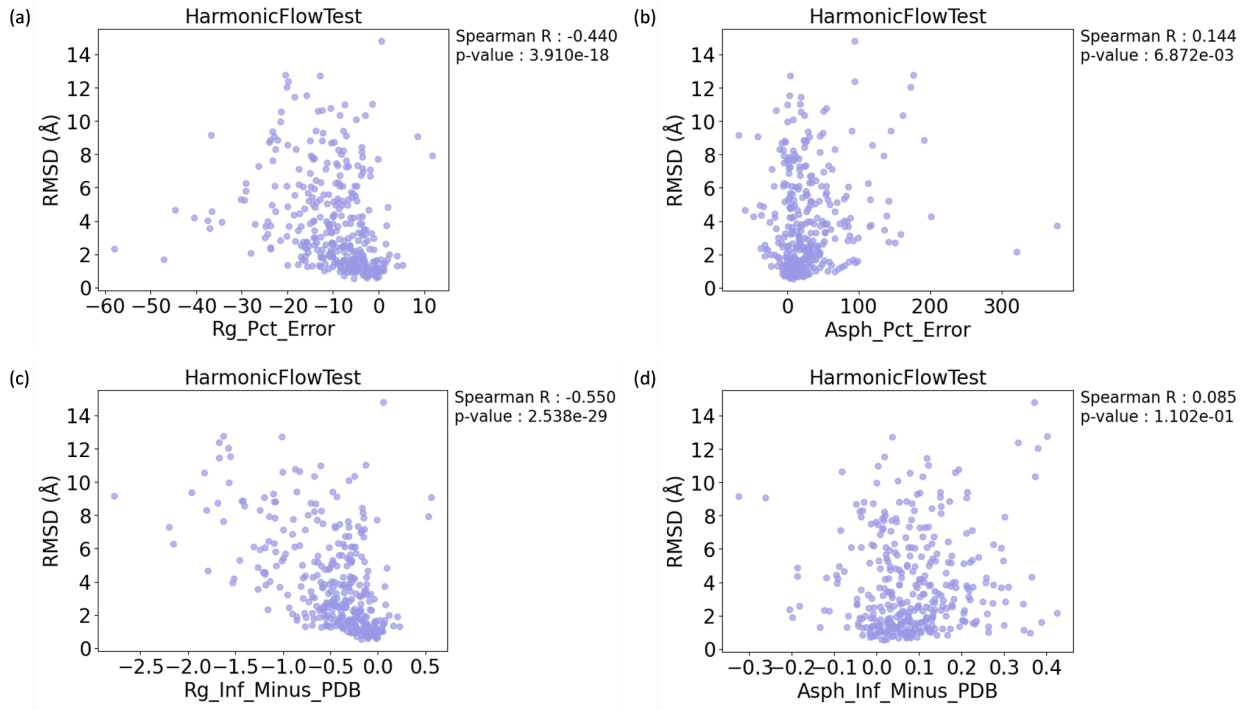



**Figure 3.** *Scatterplots of RMSD performance versus comparison of 3D metrics between the PDBBind structure ligand and the HarmonicFlow inference pose, with metrics: (a) radius of gyration percent error (b) asphericity percent error [Todeschini et al., 2003] (c) difference in radius of gyration between inference pose and PDBBind pose (inference value minus PDBBind value) (d) difference in asphericity [Todeschini et al., 2003] between inference pose and PDBBind pose (inference value minus PDBBind value)*

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 the 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.

Figure 4 shows the training and validation loss and RMSD metrics during the training process. We observed that HarmonicFlow with RDKit prior shows lower and faster convergence on loss during training. In addition, the HarmonicFlow with RDKit prior almost outperform HarmonicFlow with Harmonic prior in the first few epoch in almost all of the validation metrics, including median validation RMSD, median validation centroid RMSD, the proportion of validation RMSD < 2, etc.


**Figure 4.** *The pink line is the HarmonicFlow with Harmonic prior. The green line is the HarmonicFlow with RDKit prior. The plot top-left shows the training loss during training. The plot top-right shows the validation loss during training. The plot bottom-left shows the median validation RMSD during training. The plot bottom-right shows the proportion of validation RMSD < 2 during training.*

However, after the first few epochs, HarmonicFlow with RDKit prior started to underperform HarmonicFlow with Harmonic prior in median RMSD and proportion of validation RMSD < 2, which are two main metrics for evaluating the ligand poses. The HarmonicFlow with Harmonic prior eventually shows better and faster convergence on both median validation RMSD and proportion of validation RMSD < 2. The convergence is also smoother and less bouncy with Harmonic prior. Possible explanation for divergent result in loss and RMSD may be predicting the ground truth structure in a single step with RDKit prior is easier, but it is also more likely to end up in an out-of-distribution structure during inference because the prior is less “smooth” than Harmonic prior.

On the testing set, HarmonicFlow with RDKit prior still shows better loss but worse median RMSD and proportion of validation RMSD < 2. This might imply the RDKit prior included some correct longer-range information, but also some other noisy and incorrect longer-range information. Since a single RDKit prior creates both correct and incorrect inductive bias for the model, it might be beneficial to consider generating prior using multiple conformers to reduce potential incorrect longer-range information.

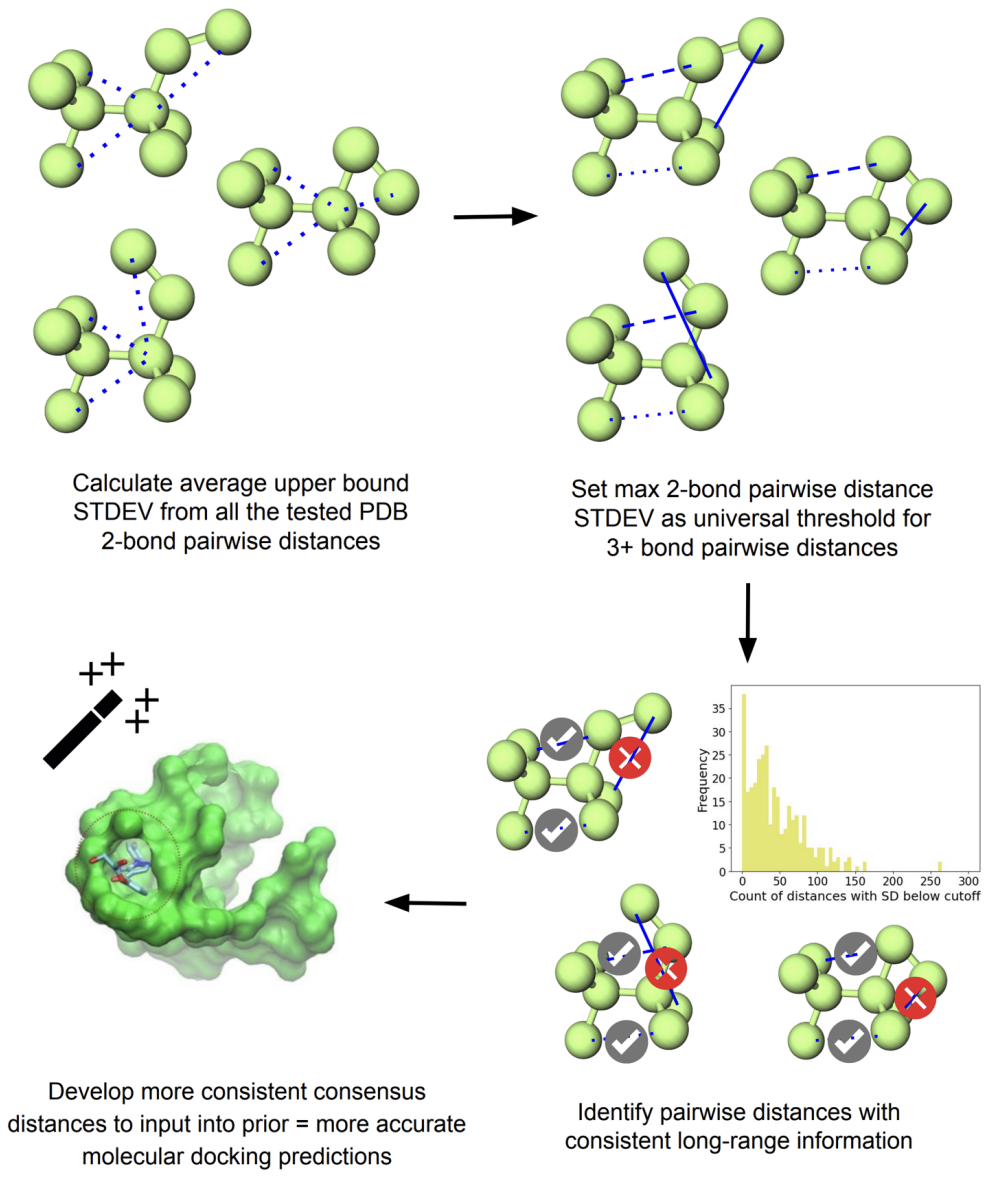



**Figure 5.** *The framework of the long-range consensus distance prior. First, the average upper bound standard deviation of the two-bond pairwise distances was calculated from all the tested PDB files and set as the universal threshold for all structures. Then, the universal threshold was applied to the long-range pairwise distance (3+ bond away); the analogous long-range pairwise distances across conformers are highlighted by line style/dashes. From this, we could identify which of the pairwise distances provided inconsistent values across conformers based on its STDEV and exclude them when creating our consensus distances. These reformed consensus distances were then utilized into the prior for leveraging multiple conformers while mitigating the noise that comes from the imperfections each conformer produces, as identified in the RDKit prior.*

Thus, our second approach to constructing the prior is to leverage consensus distances from multiple conformers generated in RDKit. Inferring the 3D ligand structure emphasized how the conformer is still a guess: there is potential in adding conformer data into the prior, but also the need to de-noise the inaccurate longer-range information when creating the consensus distances. Utilizing multiple conformers, we can try to identify and implement long-range information that doesn't vary across conformers; consistent long-range pairwise distances leveraged into the prior will mitigate the problems found in the previous prior-modulation by utilizing long-range information that is, mostly, correct.

Considering how HarmonicFlow performs very well in predicting closer pairwise distances, there is small variation in the pairwise distances of heavy atoms two bonds away. Figure 5 shows how we can leverage the upper bound of these pairwise distances to cap the wider, more spread distribution of STDEVs for the long-range (3+ bond distance) pairwise distances. Notably, the upper bound used is consistent across all structures and averaged from the maximum of the STDEV of all heavy atoms two bonds away from the 300+ PDB files utilized in testing. By identifying the pairwise distances that meet this STDEV threshold, we eliminate the ambiguity in the longer-range information produced by HarmonicFlow. From this, consistent pairwise distances for longer-range information can be utilized when creating consensus distances to integrate into the prior.

While this second approach is still being trained, the specifics behind it include generating 10 conformer structures in RDKit. From these conformers, the pairwise distance matrices for each conformer are calculated. The same STDEV upper bound from the two bond distances is applied to limit the pairwise distances that may be out of range and producing noise for the 3+ bond distances. With these, the average distances are converted to a position matrix to reconstruct the 3D structure as prior for the mode.

This new prior is being tested by, again, retraining HarmonicFlow on pocket-level docking with both the original Harmonic prior and the capped long-range information of consensus distances using multiple conformers. We are using distance-pocket and time split to generate training, validation, and testing sets. Results to be released when HarmonicFlow has completed its retraining with this prior.

It should be noted that it is likely (by chemical intuition) many of these consensus distances correspond to aromatic rings, though we have not yet visualized to which atom pairs the consensus distances correspond. Seeing how many consensus distances do and do not correspond to aromatic rings may help also inspire other ideas for future priors. If the consensus distances include many non-intuitive ones, then this indicates the importance of the consensus distance analysis. If the consensus distances are mostly aromatic rings, this indicates that another possible prior could be fixing aromatic ring geometry and the consensus approach may not be relevant beyond finding aromatic rings.

**Future Goals**

The chemical drivers analysis further amplifies the relevance of understanding generalizability, a topic which Hannes had suggested investigating and which we mentioned in our proposal. 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. Because chemical features influence performance, this raises the question of whether a test set entry's having a similar ligand (or pocket- though pocket features are most correlated with performance) to one in the training set would help the test set entry reach a low RMSD. We found similarity between the test and training set on different drivers (Figure A2) on a "bulk" level. This is however insufficient to address whether there are particular test set entries very much like training set entries. The DiffDock paper [Corso et al., 2022] included a highest Tanimoto similarity analysis. It may also be interesting to explore other generalizability metrics, such as number of training set ligands with Tanimoto similarity above a certain cutoff to a test set ligand, and their effect on performance. Some work has been done on the relevance of protein domain [Corso et al., 2023], opening the door to further generalizability studies of the protein and/or protein-ligand complex. On a related note, as discussed in the main text, understanding subpopulation performance (such as all peptide-like ligands) may help understand generalizability.

Other future goals for following up on chemical drivers include using drivers to predict docking performance and docking fragments. As discussed with Jeremy, understanding combinations of drivers' effects on performance through logistic regression could help predict whether a docking method will perform well on new ligands and proteins. This could perhaps help contribute to a confidence score or tune docking run parameters to find those most suited to a ligand depending on its projected ease of prediction (e.g., using more computationally costly methods for ligands estimated to be harder to predict. The finding smaller ligands have better performance raises the question of whether running deep learning docking on fragments and then somehow integrating results could improve performance. (Fragment docking is not a novel idea, e.g., [Yanagisawa et al., 2022].)

Applications of the newly developed prior include more accurate pharmacophores for drug development. This approach allows for a comprehensive analysis of ligand structures. By modulating the prior with multiple conformers and applying a universal threshold based on our two-bond pairwise distances, we effectively took methods to address the underestimation of the radius of gyration. This refined modeling technique ensures that both short and long-range interactions are consistently represented, enhancing the accuracy of our pharmacophore models. This advancement could potentially shorten the drug discovery timeline and lead to the development of more effective therapies, paving the way for personalized medicine by allowing for the tailoring of drug designs to specific protein-ligand interactions.

One interesting analysis that we would've loved to be able to do given more time would be to analyze the results of using multiple different priors to see if we could not only get a better RMSD, but get more physically feasible results by incorporating our chemical knowledge to the protein ligand poses that we were dealing with. Due to time constraints we weren't able to expand this idea more, but there is so much we could've done with this idea. Another interesting concept that was beyond the scope of our project would've been applying some sort of Molecular dynamic simulation to the molecule after generation of the pose to see if we could improve the sensibility of the results that we come up with even more.

We have experimented HarmonicFlow with single RDKit prior and will finish experiment on prior using multiple conformers from RDKit. It is important to design a prior that provides correct inductive bias and minimizes incorrect inductive. We could incorporate more property constraints when designing the prior. Beyond multiple conformers and consensus distances, we will explore deep learning methods to generate more realistic priors that will potentially improve performance. It is also interesting to develop more realistic and comprehensive metrics that can be used to measure the performance of the protein docking model and then investigate how well HF performs with different priors from.

**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].

**GitHub Repo**
https://github.com/hynsam/MLCB2023_Final_Project

**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=FhFgIOZbtZ

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

Credille, C. V., Morrison, C. N., Stokes, R. W., Dick, B. L., Feng, Y., Sun, J., Chen, Y., & Cohen, S. M. (2019). SAR Exploration of Tight-Binding Inhibitors of Influenza Virus PA Endonuclease. Journal of Medicinal Chemistry, 62(21), 9438-9449. https://doi.org/10.1021/acs.jmedchem.9b00747

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

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

Yanagisawa, K., Kubota, R., Yoshikawa, Y., Ohue, M., and Akiyama, Y. (2022). Effective Protein–Ligand Docking Strategy via Fragment Reuse and a Proof-of-Concept Implementation. ACS Omega 7(34), 30265-30274. https://doi.org/10.1021/acsomega.2c03470

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

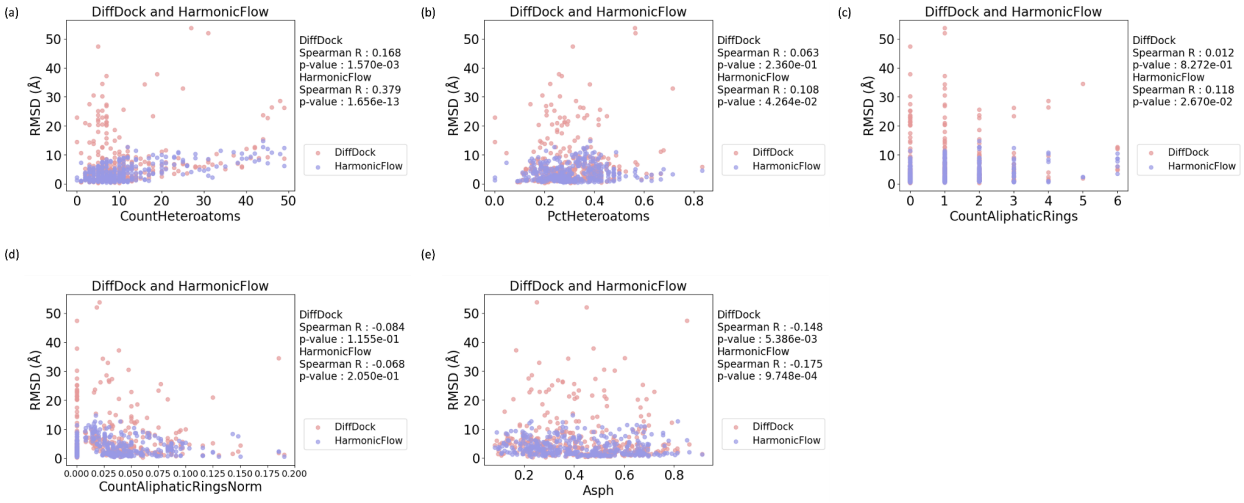



**Figure A1.** Additional HarmonicFlow and DiffDock chemical feature analysis. Correlations of RMSD with (a) ligand heteroatom count (b) percent of ligand heavy atoms which are heteroatoms (c) ligand aliphatic ring count (d) ligand aliphatic ring count to heavy atom count ratio (e) ligand asphericity [Todeschini et al., 2003]

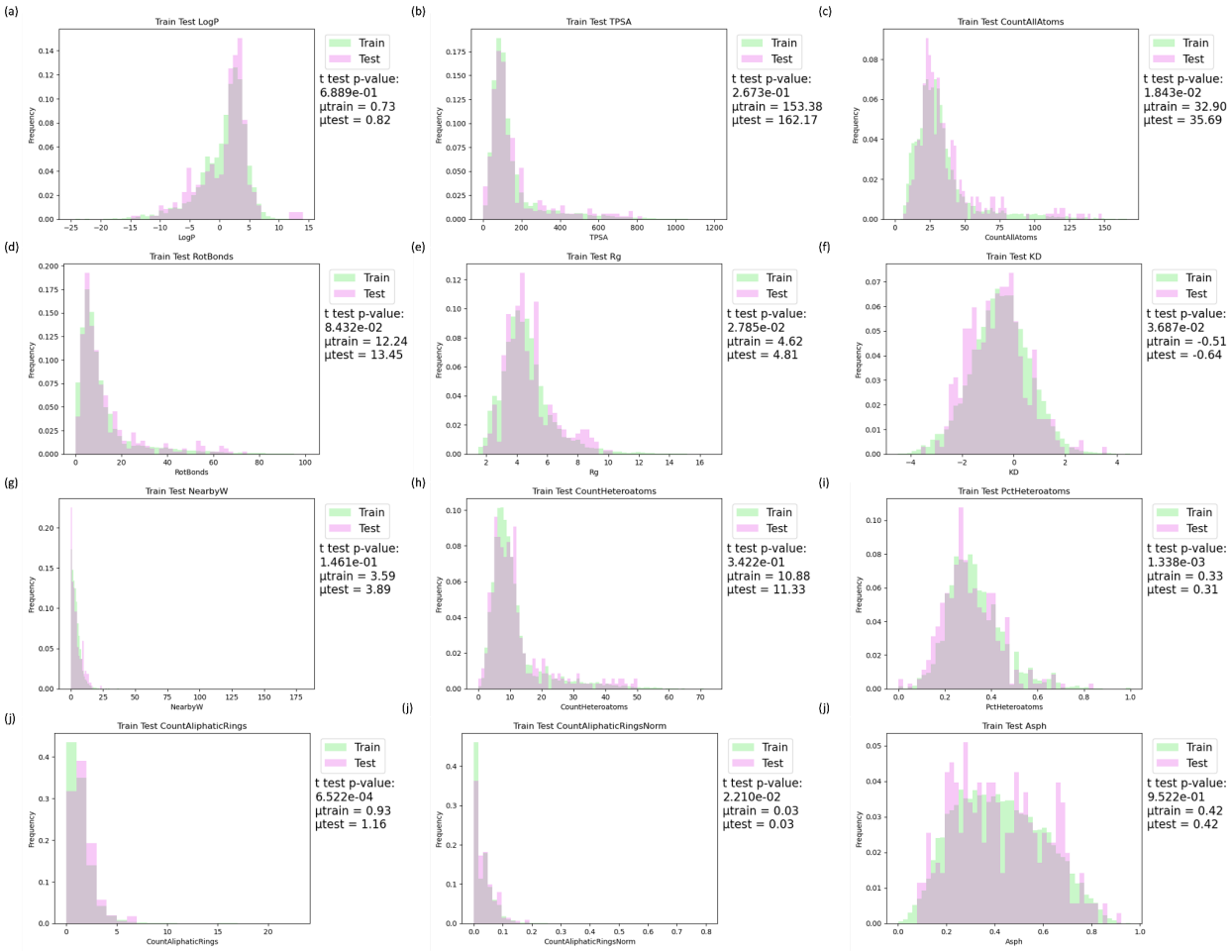



**Figure A2.** Test and training set distributions of chemical drivers: (a) ligand logP (b) ligand topological polar surface area ligand heteroatom count (c) ligand heavy atom count (d) ligand count of rotatable bonds (e) ligand radius of gyration (f) average binding site (from PDBBind ligand pose, not docked ligand pose) residue Kyte-Doolittle hydropathy (g) binding site water count with docking RMSD performance. (h) ligand heteroatom count (i) percent of ligand heavy atoms which are heteroatoms (j) ligand aliphatic ring count (k) ligand aliphatic ring count to heavy atom count ratio (l) ligand asphericity [Todeschini et al., 2003]

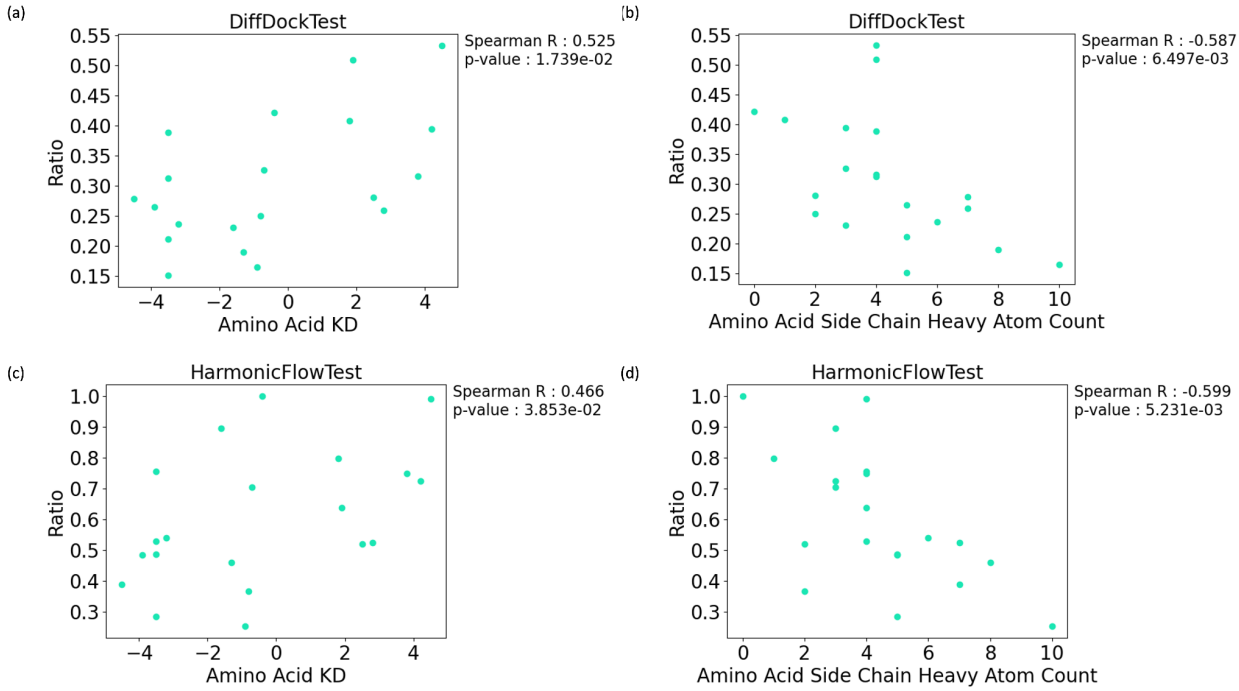



**Figure A3.** Scatterplot of ratio of each amino acid's instances in low RMSD complex binding sites to instances in high RMSD complex binding sites, plotted against (a) amino acid Kyte-Doolittle hydropathy for DiffDock (b) amino acid side chain heavy atom count for DiffDock (c) amino acid Kyte-Doolittle hydropathy for HarmonicFlow (d) amino acid side chain heavy atom count for HarmonicFlow

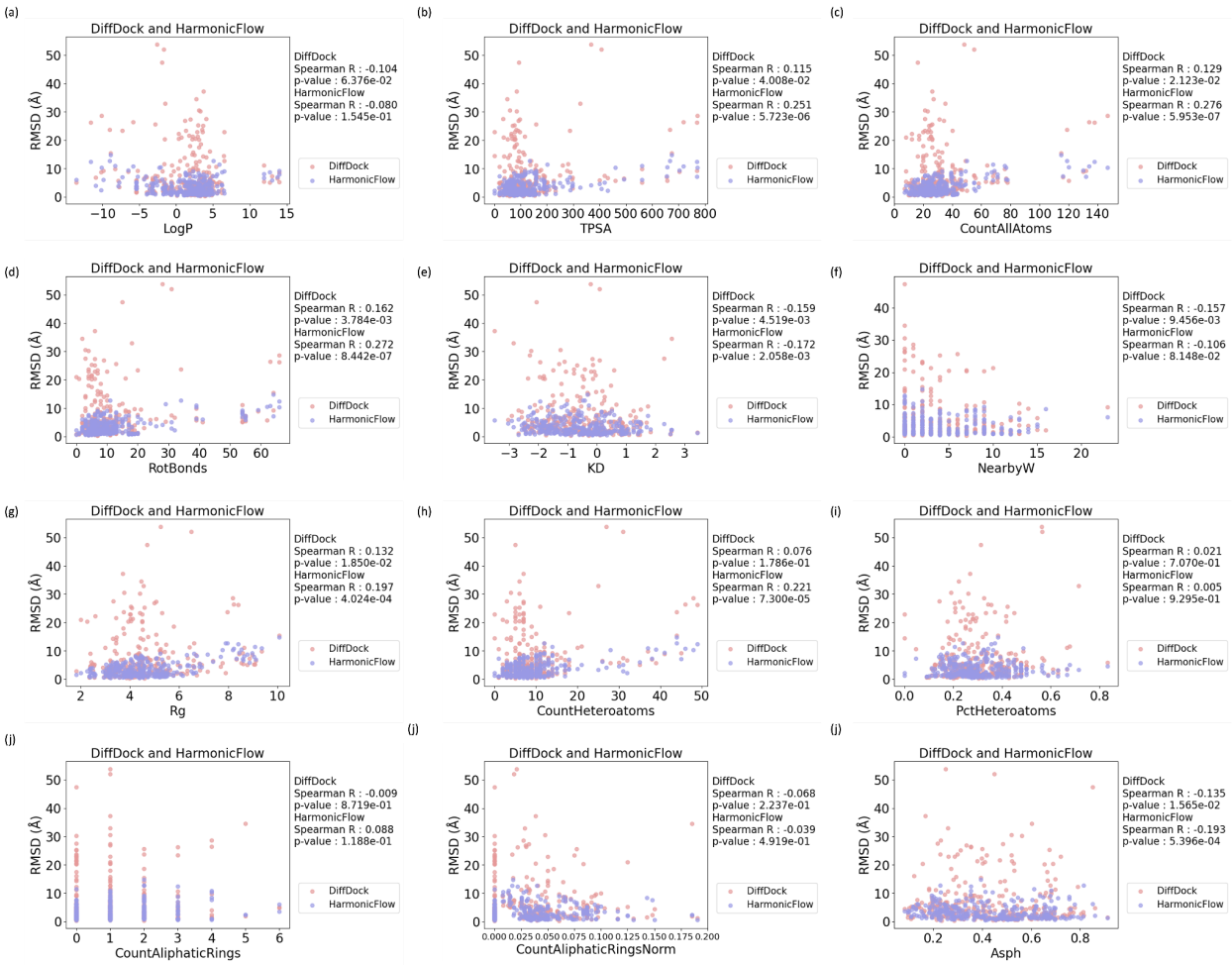



**Figure A4.** Driver correlation scatterplots excluding ligands with amino acid names (assumed- though not guaranteed- to be more peptide-like). (a) ligand logP (b) ligand topological polar surface area ligand heteroatom count (c) ligand heavy atom count (analysis for all ligands was also done in [Corso et al., 2022] for DiffDock) (d) ligand count of rotatable bonds (analysis for all ligands was also done in [Corso et al., 2022] for DiffDock) (e) ligand radius of gyration (f) average binding site (from PDBBind ligand pose, not docked ligand pose) residue Kyte-Doolittle hydropathy (g) binding site water count with docking RMSD performance. (h) ligand heteroatom count (i) percent of ligand heavy atoms which are heteroatoms (j) ligand aliphatic ring count (k) ligand aliphatic ring count to heavy atom count ratio (l) ligand asphericity [Todeschini et al., 2003]