DockFormer: Efficient Multi-Modal Receptor-Ligand Interaction Prediction using Pair Transformer

Ben Shor
Department of Computer Science
Hebrew University of Jerusalem
Jerusalem
ben.shor@mail.huji.ac.il

Dina Schneidman-Duhovny
Department of Computer Science
Hebrew University of Jerusalem
Jerusalem
dina.schneidman@mail.huji.ac.il

Abstract

Protein-small molecule interactions, or receptor-ligand interactions, are essential for understanding biological processes and advancing drug design. In this paper, we introduce DockFormer, a method that leverages multi-modal learning to predict both the binding affinity and structure of these interactions. DockFormer employs fully flexible docking, where no part of the receptor remains rigid, by adapting the AlphaFold2 architecture. Instead of relying on protein sequences and Multiple Sequence Alignments, DockFormer uses predicted receptor structures as input. This modification enables the model to concentrate on ligand docking prediction rather than protein folding, while preserving full receptor flexibility. The streamlined design also reduces the model size to just 8 layers, compared to AlphaFold2's 48 layers, greatly accelerating the inference process and making it more efficient for large-scale screening. Benchmark tests on the PoseBusters dataset demonstrated a 27% success rate, while on an affinity benchmark, DockFormer achieved a Pearson correlation of 0.91. This optimized architecture offers a valuable tool for rapid and accurate virtual screening in drug discovery.

1 Introduction

Interactions between proteins and small molecules, or receptor-ligand interactions, are critical for biological functions and drug design. A major challenge is the screening process, in which a huge database is searched for ligands that bind to a specific target receptor with high affinity. Currently, this screening process heavily relies on wet-lab experiments and physics-based computations [1], both of which are expensive and time-consuming. The importance of these interactions stresses a need for accurate *in silico* screening prediction methods.

The interaction prediction task is defined as the prediction of structure, prediction of affinity, or both. The prediction of bound interaction structure given the receptor holo- or apo- structure is generally known as the docking of a ligand in the receptor. Docking methods can be physics-based, such as GLIDE [2] and AutoDock [3] or deep-learning-based, such as DiffDock [4]. These methods use the receptor backbone structure as rigid and change the position and conformation of the ligand and possibly the side-chains. Prediction of both protein structure and the interaction is commonly referred to as co-folding. Co-folding methods, while potentially more accurate, are resource-intensive and complex [5, 6]. For many years, a major obstacle to co-folding methods was the lack of accurate methods to predict individual protein structures.

The release of AlphaFold2 [7] marked a significant advancement in protein structure prediction, utilizing a transformer-based architecture called EvoFormer to process single and pairwise sequence representations. In its successor, AlphaFold3, a leaner variation named PairFormer was used[5], where Multiple Sequence Alignments (MSA) are preprocessed separately. Although the success of

Machine Learning for Structural Biology Workshop, NeurIPS 2024.

AlphaFold2 was partly due to MSA, studies have shown that even without MSA, the model can predict protein interaction interfaces [8, 9], stressing the competence of this architecture. Despite these advances, structural models generated are not ideal for docking ligands, as evidenced by the reduced accuracy of methods like GLIDE (44% to 15%) and DiffDock(38% to 21%) when using AlphaFold2 predictions [1, 10]. This is because small-molecule binding can alter the protein conformation in mechanisms known as induced fit or population shift [11, 12]. Change can be small, for example, only in the side-chains, or significant when binding to the cryptic allosteric sites [13].

In addition to structure prediction methods, there are also approaches for predicting affinity without the bound structure. These methods can work either with the protein structure or just its sequences. While they perform well when trained on a specific protein, they struggle to generalize to new proteins outside their training set. This limitation often arises because, without the bound structure, the models cannot learn the specific interactions influencing affinity and instead tend to memorize small molecule features and their impact on affinity in specific proteins [14].

Multi-modal learning is the concept of a model that processes multiple different types of data. Here we apply multi-modal learning to train a multi-task model, DockFormer, that performs two prediction tasks of different data types - structure and affinity. In computer vision, it has been shown that if tasks are similar enough we could get improved accuracy in all tasks by using a single model, as opposed to different specialized models[15]. In the field of protein folding, multi-modal methods such as ESM3[16], have also been proven to be successful in obtaining state-of-the-art performance on multiple tasks such as structure prediction, structure design, and masked position prediction.

Here we present DockFormer, a multi-modal model that learns to predict both affinity and bound structure. DockFormer explores the ability of the PairFromer architecture to identify biological and physical interfaces by removing the MSA completely and letting the model predict contacts between receptors and ligands solely based on their geometrical and physicochemical compatibility. DockFormer tackles the trade-off between docking and co-folding by creating a model that folds the receptor, however, it expects as input the backbone distogram, that is the distances between each pair of Cα atoms in a prediction of the receptor structure or in a structure of an apo conformation. As the backbone distogram only supplies optional restraints to the structure, the model will still learn to co-fold the receptor and ligand. However, this allows the model to not learn the complex folding process of proteins and allows for simpler and less resource-intensive model inference.

2 Methods

DockFormer architecture is based on the AlphaFold2[7] architecture, with several significant differences to support ligands and enhance efficiency. The main difference is in the replacement of the MSA and template inputs with an input receptor structure. This change also results in making the Evoformer much simpler, and more similar to the PairFormer architecture presented in AlphaFold3[5]. The number of layers in the Evoformer portion of the model is reduced to 8 (instead of 48 in AlphaFold2). Another significant difference is the incorporation of tokens representing ligand atoms, in addition to the tokens representing amino acids. The last change is the addition of an affinity module that is used to train the model as a multi-modal model.

The first step is the input embedding. The input for DockFormer is an AlphaFold2 prediction, or an experimentally determined apo conformation, of the receptor structure and a reference structure of the ligand generated by RDKit [17]. The output of the embedding stage is a single embedding and a pair embedding. The single embedding is a 1D vector containing an embedding vector for each token - either an amino acid or a ligand atom. The pair embedding is a 2D vector that contains an embedding for each pair of tokens. The features that are used for embedding protein tokens are only the amino acid types. For ligand tokens, the features used are atom types, charge, chirality, and bond types between atoms. Additionally, the input structures are converted into intra-protein and intra-ligand distograms, that is the distances between each pair of tokens. For the protein, the distance between Cα atoms is calculated and matched to one of 15 bins spread evenly between 3.25Å and 20.75Å, similar to the recycling process in AlphaFold2. For the ligand, the distances between the atoms are matched to one of the 10 bins evenly spread between 0.75Å and 9.75Å. The encoding of each matched bin is added to the embedding in the pair embedding.

After generating single and pair embedding, those are passed through a transformer-based model, with 8 blocks, where each block is similar to a Pairformer block. Recycling is also applied similarly

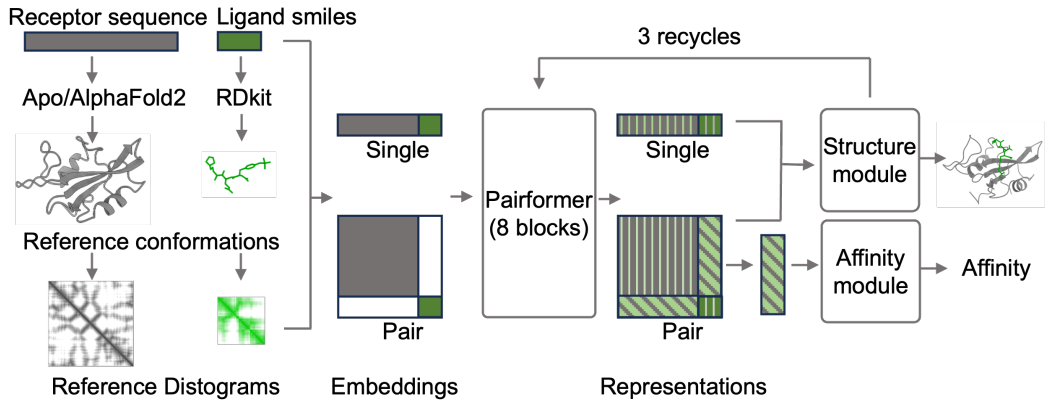

Figure 1: **The model architecture.** The inputs to DockFormer are receptor and ligand structure distograms. The structures are not required to be in the holo conformation, they can be an apo experimental structure or a structure prediction. For the ligand, a random reference conformation is generated. The distogram of the structures and the types of amino acids and atoms is then used to create a one-dimensional embedding for each input token referenced as single embedding and a two-dimensional input embedding for each pair of tokens referenced as pair embedding. These embeddings are the input for the deep learning-based model. The model consists of 8 Pairformer blocks that each output a single and pair representation. Both representations are used by the structure module to generate a structure. Only the parts in the pair representation that represent a pair of a ligand atom and a receptor amino acid are considered in the affinity module.

to AlphaFold2 with 3 recycles. The final single- and pair- representations that are outputted by the last block are passed to the structure module as implemented in AlphaFold2. For ligand tokens, the frame rotations are ignored, and only transformations of frames are used as the atom positions. An additional optional step is to replace the predicted ligand positions with the closest conformation from a set of 1,000 conformations by RDkit. This step prevents impossible bond distances or conformations.

In addition to the structure module, there is also an affinity module, which predicts the affinity as a classification task. The classification task is defined as 30 bins for affinities between 0 and 15, where the number represents the minus log of the nanomolar measured affinity. The final predicted affinity is a weighted average of the bins, weighted by the probability predicted for each bin. To predict the probability of each affinity bin, we apply a linear layer to every vector in the pair representation that represents a pair of a ligand atom and a receptor amino acid. Each such vector is weighted by a layer designated to weigh vectors for affinity relevance, and also by considering the logits predicted by the inter-contact auxiliary head.

An additional approach for affinity prediction was tested. In this approach, we add a classification token to the model input, similar to what is common in Natural Language Processing models [18]. The vector corresponding to the affinity token in the single representation is passed through a linear layer to predict the affinity bins. This approach has shown similar results to the interface-based approach. Although the interface-based approach is more complicated, it leverages data learned from structures, in the form of receptor-ligand contacts and encourages biological restraints on affinity prediction.

We have also implemented several new auxiliary heads and loss modifications. The inter-contact auxiliary head is trained to predict, given a pairwise representation, the binary classification task of whether the amino acid Cα atom and ligand atom have a distance of at most 5Å. The binding site prediction head predicts for a single representation of a receptor token whether this token is at most 5Å of the ligand. We have also added a variation of FAPE loss that considers all frames of the receptor, but only the ligand atoms, and by weighing this loss, we have forced the model to emphasize the contact between receptor and ligand.

The training data is composed of ~20,000 samples from the PDBBind 2020 dataset and ~30,000 PLINDER samples that had a correlated binding affinity, marked in the training split and released before September 2019 (to prevent data leakage to the PoseBusters benchmark). We first trained

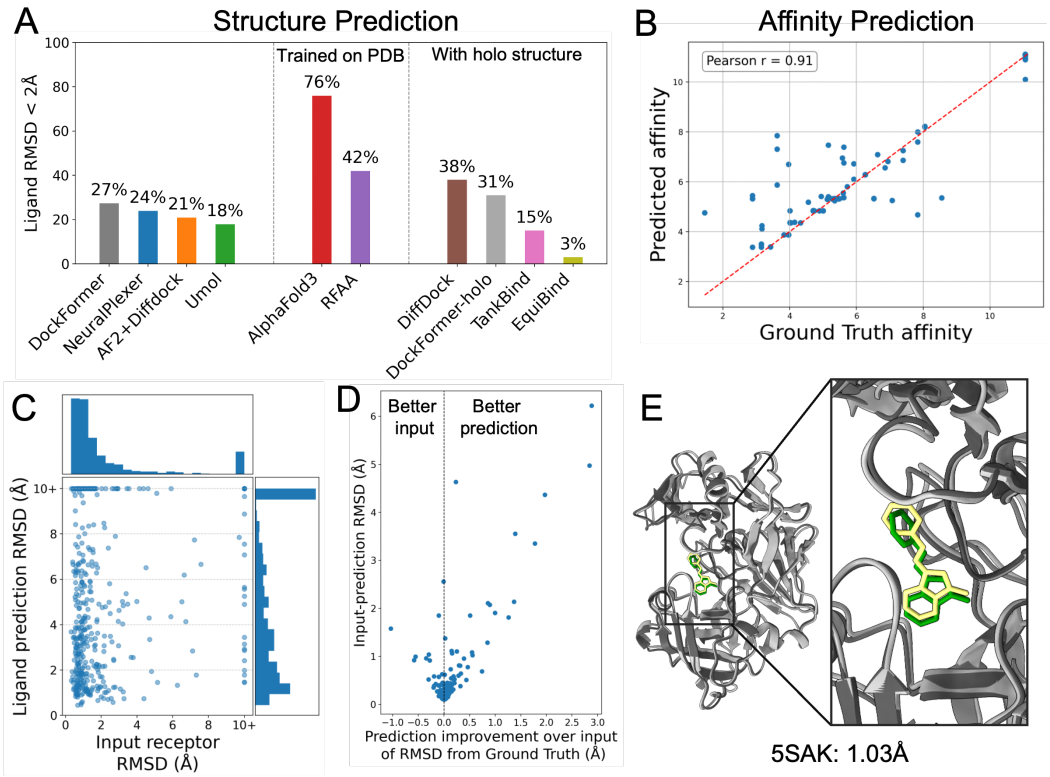

Figure 2: **Results on PoseBusters benchmark A** Percentage of predictions with ligand RMSD under 2Å, as measured on the PoseBusters benchmark (n=428). **B** Predicted affinity accuracy on benchmark derived from PLINDER test set (n=104). Axes are the minus log of the nanomolar affinity. **C** RMSD between predicted ligand positions when pockets aligned, compared to RMSD between ground truth receptor structure and input receptor structure. **D** The y-axis is the RMSD between the input receptor structure and the predicted receptor structure. The x-axis is the result of subtracting the RMSD between the input structure and the ground truth structure from the RMSD of the predicted structure and the ground truth structure. Positive x values represent an improvement in the accuracy of DockFormer prediction compared to the input structure. **E** Example of accurate prediction by DockFormer from the PoseBusters benchmark. Predicted receptor (light gray) and ligand (green) vs. ground truth receptor (dark gray) and ligand (yellow).

the model on the PDBBind samples, and only after 15,000 optimization steps started using the full dataset.

3 Results

To assess the performance of DockFormer in predicting structures, we have used the PoseBusters common protein-small molecule structures benchmark dataset[19]. This dataset contains 428 protein-small molecule interactions. For 27% of the structures, DockFormer has computed a model with ligand Root Mean Square Deviation (RMSD) of less than 2Å compared to the experimental structure, when aligned by protein pocket residues Cα atoms (a pocket residue is a residue with one of the atoms within 8Å of any of the ligand atoms). DockFormer is favorably comparable to methods that do not get the holo receptor structure as input and did not use the entire Protein Data Bank (PDB) dataset for training. Although AlphaFold3 and RosseTTAFold-AllAtom achieve significantly higher performance (Figure 2A) their training set was considerably larger, as they were trained on the entire PDB. An interesting comparison is between DockFormer and Umol[10] (accuracy of 0.27 vs. 0.18), as both models are based on the AlphaFold2 architecture and their main difference is the smaller

number of layers in DockFormer and the removal of MSA from the pipeline. This is evidence that the concept of using input structures over co-folding completely from scratch seems to be beneficial.

To further analyze DockFormer structure prediction, we have examined the relationship between input receptor structure accuracy (generated by AlphaFold2) and docking accuracy. We find that in 74% of the test set samples in PoseBusters the Cα RMSD to the ground truth structure is below 2Å, and only in 8% of the samples it is above 10Å(Figure 2C). In 98% lof samples, the input receptor structure does not change significantly after the prediction(Cα RMSD < 1Å). However, in most cases where there is a significant change in structure, it becomes more similar to the ground truth structure compared to the input (Figure 2D). This provides evidence that DockFormer refines the receptor structure to fit the ligand. To further analyze this point, we have tested DockFormer with an input receptor structure in the holo conformation, in a variation named DockFormer-holo. The performance improved slightly, from 27% to 31%, indicating that the DockFormer is not dependent on an accurate receptor structure.

One of the advantages of the DockFormer is its lean architecture designed for prediction efficiency. On the PoseBusters dataset, the median prediction time was 2.67 seconds, with an average of 6.11 seconds. Although these run times do not account for the generation of the input receptor structure, they remain highly relevant for screening processes where the same receptor is tested against multiple ligands, making the receptor structure generation a negligible factor.

In addition to structure prediction, to validate the affinity prediction we have used the test dataset from PLINDER. We have filtered out only structures released after 2019 and not present in the PDBBind dataset, to prevent data leakage. This resulted in a dataset of 104 receptor-ligand complexes with affinity. The Pearson correlation of predicted affinity compared to ground truth affinity for this dataset is 0.91(Figure 2B), which shows the ability of the model to predict affinity with high accuracy.

4 Discussion

We have presented DockFormer, a method for predicting structure and affinity of receptor-ligand interactions that integrates two key concepts into the AlphaFold2 [7] architecture. First, the method uses an approximated receptor structure, either apo conformation or prediction, as an input, enabling the model to specialize in the docking task over the protein folding task, while keeping the receptor structure flexible to capture conformational changes. This also enables reducing the model architecture to provide efficiency in training and inference. Second, by employing multi-task learning for simultaneous affinity and structure prediction, the network gains additional information that can further enhance its accuracy.

Despite these innovations, our structure prediction accuracy is currently lower compared to other state-of-the-art structure prediction methods such as, AlphaFold3[5] and RoseTTAFold All-Atom[6]. We suggest that these methods demonstrate higher performance partly because they have been trained on significantly larger datasets. Training on all available PDB structures, rather than a limited subset such as PDBbind[20], provides them with a wealth of structural data, which increases the ability to generalize, and also provides the model with more examples similar to the benchmark set. We initially trained DockFormer only on samples that had both affinity and structural data available, limiting our training set to approximately 50,000 samples. However, multi-modal training also enables training on partial data, and therefore in future versions we plan to incorporate larger PLINDER structure dataset[21], which offers around 300,000 structures for training, and the BindingDB affinity dataset[22], which contains about 2,000,000 affinity measurements.

Another key concept in several of the competing successful methods is diffusion-based architecture. We believe that diffusion architectures are particularly effective at optimizing subtle interactions between the ligand and the receptor pocket, resulting in more plausible and accurate structures. They also explicitly model side-chain interactions, whereas our method treats side-chains implicitly, which could contribute to their higher accuracy. In future versions, we intend to incorporate diffusion which may increase accuracy.

The foundational concepts behind DockFormer have the potential to be extended beyond receptor-ligand interactions. For instance, similar models could be developed for protein-protein interactions, including specialized models for challenging cases like antibody-antigen binding or protein-peptide interactions. In those cases, the additional modality of the model can be specificity. By leveraging

large databases such as STRING[23], which provide interaction labels, we can train models to predict whether proteins interact. The reduction DockFormer offers in model size may make specialized task-specific models more feasible.

5 Code Availability

The code for running and training the model, with weights trained on the PLINDER dataset, is available for running inference or download at: https://huggingface.co/spaces/benshor/DockFormer. This implementation builds upon the OpenFold[24] implementation of AlphaFold2[7].

References

[1] Masha Karelina, Joseph J Noh, and Ron O Dror. How accurately can one predict drug binding modes using alphafold models? *Elife*, 12:RP89386, 2023.

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

[3] Jerome Eberhardt, Diogo Santos-Martins, Andreas F Tillack, and Stefano Forli. Autodock vina 1.2. 0: New docking methods, expanded force field, and python bindings. *Journal of chemical information and modeling*, 61(8):3891–3898, 2021.

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

[5] Josh Abramson, Jonas Adler, Jack Dunger, Richard Evans, Tim Green, Alexander Pritzel, Olaf Ronneberger, Lindsay Willmore, Andrew J Ballard, Joshua Bambrick, et al. Accurate structure prediction of biomolecular interactions with alphafold 3. *Nature*, pages 1–3, 2024.

[6] Rohith Krishna, Jue Wang, Woody Ahern, Pascal Sturmfels, Preetham Venkatesh, Indrek Kalvet, Gyu Rie Lee, Felix S Morey-Burrows, Ivan Anishchenko, Ian R Humphreys, et al. Generalized biomolecular modeling and design with rosettafold all-atom. *Science*, 384(6693):eadl2528, 2024.

[7] John Jumper, Richard Evans, Alexander Pritzel, Tim Green, Michael Figurnov, Olaf Ronneberger, Kathryn Tunyasuvunakool, Russ Bates, Augustin Žídek, Anna Potapenko, et al. Highly accurate protein structure prediction with alphafold. *nature*, 596(7873):583–589, 2021.

[8] Richard Evans, Michael O'Neill, Alexander Pritzel, Natasha Antropova, Andrew Senior, Tim Green, Augustin Žídek, Russ Bates, Sam Blackwell, Jason Yim, et al. Protein complex prediction with alphafold-multimer. *biorxiv*, pages 2021–10, 2021.

[9] James P Roney and Sergey Ovchinnikov. State-of-the-art estimation of protein model accuracy using alphafold. *Physical Review Letters*, 129(23):238101, 2022.

[10] Patrick Bryant, Atharva Kelkar, Andrea Guljas, Cecilia Clementi, and Frank Noé. Structure prediction of protein-ligand complexes from sequence information with umol. *Nature Communications*, 15(1):4536, 2024.

[11] Peter Csermely, Robin Palotai, and Ruth Nussinov. Induced fit, conformational selection and independent dynamic segments: an extended view of binding events. *Trends in biochemical sciences*, 35(10):539–546, 2010.

[12] Kei-ichi Okazaki and Shoji Takada. Dynamic energy landscape view of coupled binding and protein conformational change: induced-fit versus population-shift mechanisms. *Proceedings of the National Academy of Sciences*, 105(32):11182–11187, 2008.

[13] Sandor Vajda, Dmitri Beglov, Amanda E Wakefield, Megan Egbert, and Adrian Whitty. Cryptic binding sites on proteins: definition, detection, and druggability. *Current opinion in chemical biology*, 44:1–8, 2018.

[14] Mikhail Volkov, Joseph-André Turk, Nicolas Drizard, Nicolas Martin, Brice Hoffmann, Yann Gaston-Mathé, and Didier Rognan. On the frustration to predict binding affinities from protein–ligand structures with deep neural networks. *Journal of medicinal chemistry*, 65(11):7946–7958, 2022.

[15] Trevor Standley, Amir Zamir, Dawn Chen, Leonidas Guibas, Jitendra Malik, and Silvio Savarese. Which tasks should be learned together in multi-task learning? In *International conference on machine learning*, pages 9120–9132. PMLR, 2020.

[16] Tomas Hayes, Roshan Rao, Halil Akin, Nicholas J Sofroniew, Deniz Oktay, Zeming Lin, Robert Verkuil, Vincent Q Tran, Jonathan Deaton, Marius Wiggert, et al. Simulating 500 million years of evolution with a language model. *bioRxiv*, pages 2024–07, 2024.

[17] Greg Landrum. Rdkit: Open-source cheminformatics software. 2016.

[18] Arman Cohan, Iz Beltagy, Daniel King, Bhavana Dalvi, and Daniel S Weld. Pretrained language models for sequential sentence classification. *arXiv preprint arXiv:1909.04054*, 2019.

[19] Martin Buttenschoen, Garrett M Morris, and Charlotte M Deane. Posebusters: Ai-based docking methods fail to generate physically valid poses or generalise to novel sequences. *Chemical Science*, 15(9):3130–3139, 2024.

[20] Renxiao Wang, Xueliang Fang, Yipin Lu, and Shaomeng Wang. The pdbbind database: Collection of binding affinities for protein- ligand complexes with known three-dimensional structures. *Journal of medicinal chemistry*, 47(12):2977–2980, 2004.

[21] Janani Durairaj, Yusuf Adeshina, Zhonglin Cao, Xuejin Zhang, Vladas Oleinikovas, Thomas Duignan, Zachary McClure, Xavier Robin, Daniel Kovtun, Emanuele Rossi, et al. Plinder: The protein-ligand interactions dataset and evaluation resource. *bioRxiv*, pages 2024–07, 2024.

[22] Tiqing Liu, Yuhmei Lin, Xin Wen, Robert N Jorissen, and Michael K Gilson. Bindingdb: a web-accessible database of experimentally determined protein–ligand binding affinities. *Nucleic acids research*, 35(suppl_1):D198–D201, 2007.

[23] Damian Szklarczyk, Rebecca Kirsch, Mikaela Koutrouli, Katerina Nastou, Farrokh Mehryary, Radja Hachilif, Annika L Gable, Tao Fang, Nadezhda T Doncheva, Sampo Pyysalo, et al. The string database in 2023: protein–protein association networks and functional enrichment analyses for any sequenced genome of interest. *Nucleic acids research*, 51(D1):D638–D646, 2023.

[24] Gustaf Ahdritz, Nazim Bouatta, Christina Floristean, Sachin Kadyan, Qinghui Xia, William Gerecke, Timothy J O'Donnell, Daniel Berenberg, Ian Fisk, Niccolò Zanichelli, et al. Openfold: Retraining alphafold2 yields new insights into its learning mechanisms and capacity for generalization. *Nature Methods*, pages 1–11, 2024.