# Smiles2Dock: a large-scale dataset for ML-based docking score prediction using AlphaFold structures

**Anonymous Author(s)**
Affiliation
Address
email

## Abstract

Docking is a crucial component in drug discovery aimed at predicting the binding conformation and affinity between small molecules and target proteins. ML-based docking has recently emerged as a prominent approach, outpacing traditional methods like DOCK and AutoDock Vina in handling the growing scale and complexity of molecular libraries. However, the availability of comprehensive and user-friendly datasets for training and benchmarking ML-based docking algorithms remains limited. Moreover, existing datasets rely on proteins with experimentally determined structures and known ligand binding pockets, making them unusable for the growing number of proteins with only predicted structures. We introduce Smiles2Dock, an open large-scale dataset for molecular docking that addresses this gap. We created a framework combining P2Rank for binding pocket prediction and AutoDock Vina for docking, enabling us to dock 1.7 million ligands from the ChEMBL database against 11 genetically validated proteins from AlphaFold, resulting in over 17 million protein-ligand binding scores. Since AlphaFold-predicted structures do not include known ligand binding sites, our use of P2Rank allows docking to be performed without any experimental structure information, a first at this scale. The dataset encompasses a diverse set of biologically relevant compounds and enables researchers to benchmark all major approaches for ML-based docking such as Graph, Transformer, and CNN-based methods. We also introduce a novel Transformer-based architecture for docking score prediction and set it as an initial benchmark for our dataset.

## Introduction

Molecular docking is a computational technique used to predict how a small molecule binds to a protein target [30]. By estimating the ligand's position and orientation within the binding site, docking helps assess how well a compound might interact with and modulate the protein's function [31]. Traditional molecular docking methods use scoring functions designed to estimate how strong the interaction is between a protein and a ligand based on their 3D arrangement. These scoring functions are based on physical and chemical principles that estimate binding strength from a predicted 3D pose of the ligand in the protein. On the other hand, docking score predictors are Machine Learning (ML) models that learn to predict docking scores directly from molecular graphs or sequences.

### Traditional docking

Docking is widely used in drug discovery to screen large libraries of compounds and prioritize those most likely to bind effectively, reducing the need for costly synthesis and experimental testing [14, 38]. Docking algorithms output binding scores and poses, which estimate binding affinity and suggest how a ligand fits into the protein's active site. These predictions guide the selection of promising

Submitted to the AI for Science workshop (NeurIPS 2025).

compounds for further development. Effective docking results forecast where and how well a ligand binds and provide insights into the nature of the binding affinity and specificity.[30].

As the scale of molecular libraries expands dramatically in drug discovery, the need for faster and more efficient docking tools has become key. Traditional scoring functions such as DOCK, which relies on geometric matching algorithms to fit ligands into protein binding sites; AutoDock Vina, which uses gradient-based optimization to predict binding poses; and Glide, which performs systematic searches across ligand conformations, orientations, and positions, have proven too slow to handle modern large-scale libraries [6, 43, 11]. In response, researchers are turning to machine learning (ML) docking score predictors - models trained to replicate docking scores generated by programs like AutoDock Vina. When deployed on GPUs, these models can predict binding outcomes significantly faster than traditional methods, achieving speedups of 10 to 100 times [7].

## ML-based docking

Machine learning-based docking algorithms can be broadly divided into two categories: (1) docking score predictors, which directly estimate the binding affinity or docking score between a protein and a ligand, and (2) end-to-end docking methods, which predict the score as well as binding poses and affinities from structural or sequence-based inputs. Each branch leverages different model architectures to address the complexity of molecular interactions. Several ML approaches have been tried. The most prominent one is Graph Neural Networks (GNNs), which directly model the molecular structure of proteins and ligands as graphs where atoms are nodes and bonds are edges [20, 44, 17].

An extension of Graph Neural Networks (GNNs) is Graph Convolutional Networks (GCNs), which apply convolutional operations to graph-structured data, allowing the model to capture the topological features of molecules and their potential interactions with proteins [41] and predict their docking score. Similarly, Transformer-based architectures, originally designed for natural language processing, have been adapted for molecular data by treating atoms or fragments as sequence elements. These models, which can be pretrained on large corpora of SMILES strings, effectively capture long-range dependencies within molecules and across molecular complexes, ultimately representing proteins and ligands as embedding matrices or vectors [16, 15, 4].

Computer vision-based approaches, such as 3D Convolutional Neural Networks (3D CNNs), extend the concept of convolution into three dimensions, making them well-suited for modeling the spatial structure of molecules and the 3D configuration of protein-ligand interactions [46, 19] and are typically used for predicting binding scores rather than binding poses. However, GNINA integrates deep learning with traditional docking pipelines to predict both poses and affinities, making it an example of an end-to-end docking model [29].

Finally, reinforcement learning has been explored as an end-to-end solution, particularly through the asynchronous advantage actor-critic (A3C) framework. These methods treat docking as a sequential decision-making process, with actor models guiding search strategies and critic models evaluating them, allowing direct prediction of binding poses and improving docking performance [3, 1].

## Datasets for molecular docking

The downside of ML-based methods is the amount of data required for training. To solve this, several groups have attempted to build open-source docking datasets by using docking software predictions as inputs for ML models [5, 12, 27, 42]. However, available large-scale docking datasets have several limitations, notably scale, ease of use and lack of generalizability. Some focused on a specific set of proteins linked to a certain disease (e.g. SARS-COV2 proteome), greatly reducing generalization capabilities for ML models trained on those. Others used a number of ligands not in the scale of modern compound libraries, which often have millions of data points, and did not use well-known extensively tested chemical libraries.

## Lack of experimental protein structures

Reliable docking requires accurate 3D structures of proteins, especially their binding sites. These are typically obtained through experimental techniques like X-ray crystallography or cryo-EM, which are slow, expensive, and often infeasible, particularly for membrane proteins or unstable targets. As a result, only about 17% of the human proteome has experimentally resolved structures [32]. Existing

ML docking datasets like DOCKSTRING and DUD-E rely on these experimentally determined proteins and annotated binding sites, making them unusable for the majority of proteins with only predicted structures. Smiles2Dock is the first large-scale dataset to enable docking on proteins without any experimental structural data. We use AlphaFold-predicted 3D structures, which cover nearly the entire human proteome, and apply P2Rank to predict binding pockets directly from the structure. This lets us perform docking using AutoDock Vina on 11 genetically validated AlphaFold targets and 1.7 million ChEMBL ligands, resulting in over 17 million docking scores. All binding site predictions are released for reuse. Ligands are represented using SMILES strings, supporting Transformer-based, graph-based, and 3D computer vision models. The full dataset is hosted on Hugging Face, and can be loaded with just two lines of Python code [25, 23].

# Results

## Correlation and variability of docking scores

We computed the Pearson correlation coefficient between docking scores (Figure 1). A subset of proteins including *slc30a8*, *dpp9*, and *ifih1* formed a highly correlated cluster ($r > 0.8$), suggesting shared ligand binding preferences and possibly similar pocket chemotypes. Proteins such as *adcy5* and *cfhr5* exhibited weak correlations ($r < 0.3$) with most others, reflecting distinct binding environments or limited cross-reactivity with the ligand set. Boxplot analysis (Figure 2) revealed most proteins showed compact interquartile ranges and moderate outlier counts, consistent with well-behaved docking score distributions. Proteins such as *cfhr5* demonstrated particularly tight distributions, whereas *dpp9* and *nrlp3* showed larger score spreads and several high-affinity outliers (scores $< -12$ kcal/mol).

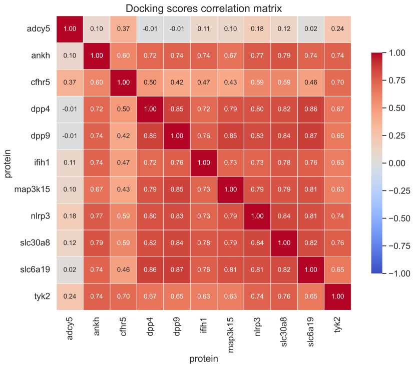

Figure 1: Correlations of docking scores per protein.

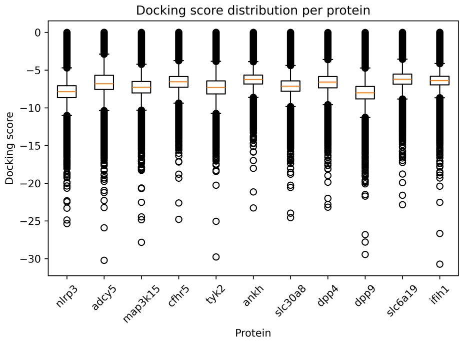

Figure 2: Boxplots of docking scores per protein.

## Distribution of docking scores

Our initial hypothesis after looking at Figure 3 was that scores for each protein were normally distributed. We performed Shapiro-Wilk tests to test for normality of scores distribution for each protein but all p-values were below the 0.05 significance threshold. [35]. We discovered, by computing a Q-Q plot (Figure 4), that the distribution was heavily right-skewed, which was also confirmed by computing the skewness of the distribution of scores for each protein, with values ranging from 5 to 20 (heavily right skewed) [28]. Finally, we tested for right-skewed distributions by performing a Kolmogorov-Smirnov test for goodness of fit using Log-Normal and Weibull distributions but again found p-values for all proteins below the significance threshold required [2].

## Metrics for evaluating models

To evaluate model performance in docking score prediction, we use ranking-based metrics that focus on the relative ordering of compounds rather than their absolute scores. This aligns with the practical

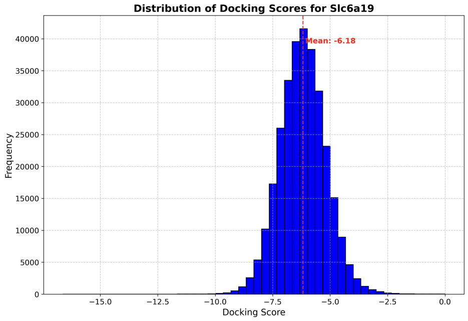

Figure 3: Scores distribution for protein Slc6a19.

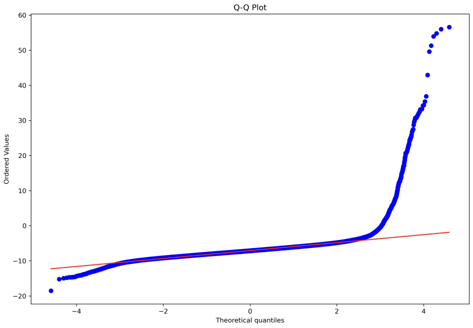

Figure 4: QQ plot for protein Slc6a19.

goal in drug discovery: prioritizing a small number of compounds for experimental validation, where identifying the top candidates is more important than accurately modeling the full score distribution. Spearman correlation is well-suited here, as it quantifies the agreement between predicted and true rankings across the dataset [45]. The Spearman rank correlation coefficient ($\rho$) is a non-parametric measure of statistical dependence that captures how well the relationship between two variables follows a monotonic trend. Given predicted and true docking score vectors $\hat{\mathbf{y}}$ and y, we denote their ranks by:

$$\text{rank}(\hat{y}_i) = R_i, \quad \text{rank}(y_i) = S_i.$$

Then the Spearman correlation is given by:

$$\rho = 1 - \frac{\sum_{i=1}^{n}(R_i - S_i)^2}{n(n^2 - 1)}.$$

The top-k overlap measures how well the model identifies the most promising compounds. This metric ranges from 0 to 1. A value of 1 indicates perfect agreement between the predicted and true top-$k$ ligands, while 0 indicates no overlap. Define $T_k$, the set of indices corresponding to the top-$k$ ligands based on the true scores, and $\hat{T}_k$, the set of indices corresponding to the top-$k$ ligands based on the predicted scores. Then the top-$k$ overlap is defined as:

$$\text{Top-}k\text{ overlap} = \frac{|T_k \cap \hat{T}_k|}{k}$$

### Hybrid model results

Table 1 presents Spearman correlation and top-k overlap metrics for various model configurations, varying protein model (PM), ligand model (LM), hidden layer (HL) sizes, and dropout rates. The best overall performance is achieved by the model with PM=128, LM=256, HL=64, and dropout 0.1, showing the highest Spearman correlation of 0.76 at the top 50% and maintaining good correlation (0.24) even at the top 1%. This suggests a good balance between model capacity and regularization. Larger models with higher dropout (e.g., PM=256, LM=512) show lower Spearman correlations and top-k overlaps, which may indicate over-regularization or difficulty training such large architectures on the dataset. Smaller models (e.g., PM=64, LM=64) perform moderately well but generally fall short of the best configuration. As expected, top-k overlap decreases with stricter cutoffs (from top 50% to top 1%), reflecting the increasing challenge of identifying the very best candidates. Overall, moderate-sized models with moderate dropout provide the most consistent and accurate ranking performance.

## Limitations

**P2Rank as a probabilistic framework for binding site prediction:** P2Rank uses an ML-based algorithm to predict the binding sites of each protein along with an associated probability. We used an arbitrary threshold to define what counted as a "valid" binding site. In the case where we had multiple binding sites above our 50% threshold, we only used the one with the highest probability.

| PM size | LM size | HL size | Dropout | Top 50% | Top 25 % | Top 10 % | Top 1 % |
|---:|---:|---:|---:|:---:|:---:|:---:|:---:|
| 128 | 256 | 64 | 0.1 | 0.76 / 0.45 | 0.64 / 0.33 | 0.51 / 0.16 | 0.24 / 0.02 |
| 256 | 512 | 128 | 0.2 | 0.63 / 0.05 | 0.34 / 0.18 | 0.18 / 0.16 | 0.10 / 0.05 |
| 64 | 64 | 128 | 0.2 | 0.73 / 0.38 | 0.55 / 0.34 | 0.41 / 0.20 | 0.21 / 0.06 |
| 128 | 256 | 128 | 0.2 | 0.44 / -0.22 | 0.15 / 0.06 | 0.04 / 0.16 | 0.05 / 0.07 |
| 128 | 128 | 256 | 0.1 | 0.62 / 0.03 | 0.32 / 0.16 | 0.16 / 0.16 | 0.10 / 0.05 |
| 256 | 512 | 64 | 0.2 | 0.67 / 0.15 | 0.43 / 0.20 | 0.28 / 0.15 | 0.16 / 0.04 |
| 256 | 256 | 256 | 0.2 | 0.64 / 0.08 | 0.37 / 0.18 | 0.21 / 0.16 | 0.13 / 0.06 |
| 64 | 128 | 64 | 0.2 | 0.56 / -0.07 | 0.25 / 0.11 | 0.10 / 0.15 | 0.07 / 0.04 |

Table 1: Spearman correlation (left) and top-k overlap (right) for different percentiles of top scores (PM = Protein model, LM = Ligand model, HL = Hidden layer).

**Conformational space exploration:** We used an exhaustiveness parameter of 8 and tried 5 different poses, the default values for Vina which are known in other studies for balancing accuracy and computational resource use. Increasing those further would not have been feasible but it could be beneficial for future studies to do a more thorough search. We also limited ourselves to one binding site per protein, both for computational resources and also to standardize the prediction task for ML researchers. However, it could be interesting to look at algorithms that can work on multiple binding sites at the same time.

## Methods

**Alphafold:** AlphaFold is an ML model developed by DeepMind designed to predict protein structures and solve the protein structure prediction problem, which involves determining a protein's three-dimensional shape from its amino acid sequence [21]. Its predictions have been extensively validated, with a reported root-mean-square deviation (RMSD) of around 1.5 Ångströms for many proteins, comparable to experimental methods like X-ray crystallography and cryo-electron microscopy, while being significantly cheaper and faster.

**ChEMBL:** ChEMBL is a bioactivity database maintained by the European Bioinformatics Institute, containing detailed information on the biological activity of 2.3M small molecules [13]. It is widely used for drug discovery and development, offering data on compound properties, target interactions, and pharmacological profiles.

**P2Rank:** P2Rank is an ML model for predicting ligand-binding sites on proteins by analyzing surface patches based on features like hydrophobicity, electrostatic potential, and geometric arrangement of atoms [24]. Each protein surface is segmented into patches, with the random forest model assessing the likelihood of each patch being a binding site based on the extracted features. An example of the binding pocket predicted by P2Rank for protein adcy5 can be seen on figure 6.

**AutoDock Vina:** AutoDock Vina is a popular molecular docking package widely used in computational chemistry for predicting the interactions between a protein and a ligand. It uses a scoring function to estimate the strength and stability of a ligand when docked into a protein's binding site.

### Dataset preparation

In our study, we developed a dataset of molecular docking scores using a comprehensive framework to ensure precise predictions of protein-ligand interactions, which can be seen in Figure 5.

### Ligand preparation

For ligands, we downloaded the ChEMBL database and used the 2.3M SMILES strings available. Out of this set, around 20% could not be processed by AutodockVina because of errors, either when converting SMILES strings to .sdf files using RDKit or due to atom types incompatible with Autodock. This left us with a set of approximately 1.7M ligands to dock. Then the ligands were deprotonated at ph 7.4 with OpenBabel. Finally, a 3D conformation was generated using RDKit and

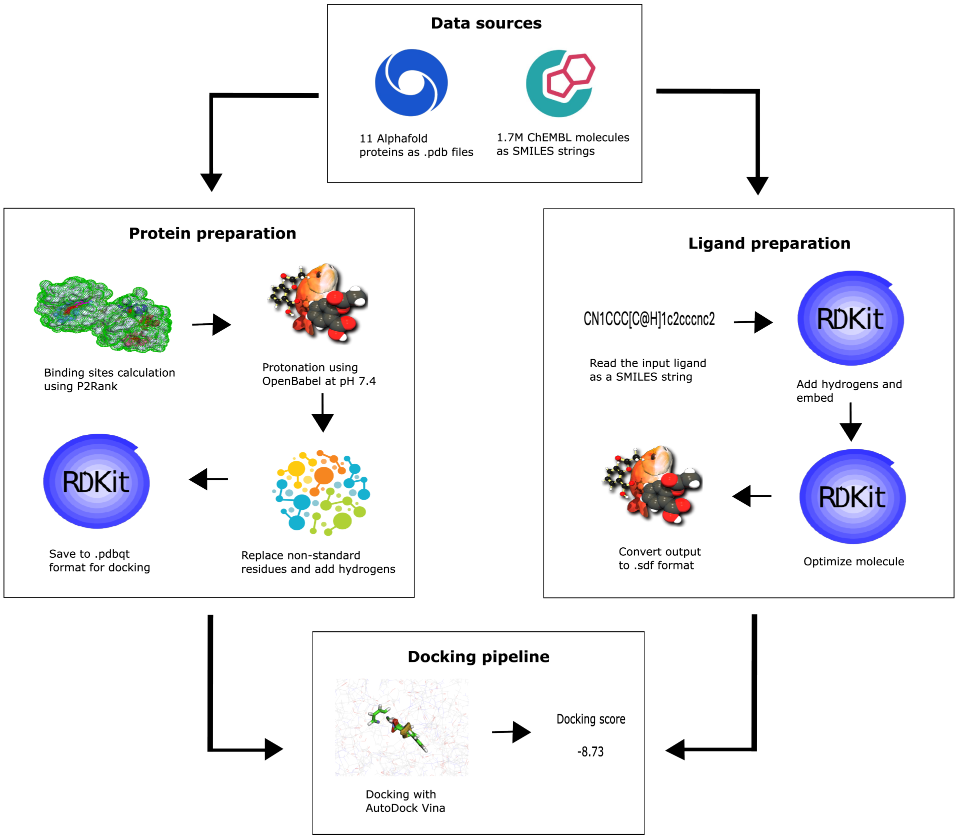
Data sources
11 Alphafold proteins as .pdb files
1.7M ChEMBL molecules as SMILES strings
Protein preparation
Binding sites calculation using P2Rank
Protonation using OpenBabel at pH 7.4
RDKit
Save to .pdbqt format for docking
Replace non-standard residues and add hydrogens
Ligand preparation
CN1CCC[C@H]1c2cccnc2
Read the input ligand as a SMILES string
RDKit
Add hydrogens and embed
RDKit
Convert output to .sdf format
Optimize molecule
Docking pipeline
Docking with AutoDock Vina
Docking score
-8.73

Figure 5: Diagram of the methodology followed for this project.

the ETKG algorithm, which we further refined using the classical force field MMFF94. Finally, we outputted a PDBQT file for each ligand, which is the file format required by AutoDock Vina.

## Protein preparation

For proteins, we started with a set of 30 proteins from the AlphaFold database based on their identification as therapeutic targets in previous genetic association studies [39, 37, 40, 8]. These are proteins where natural human mutations (identified through GWAS) have been linked to protection from or risk of disease. This selection strategy is based on work by Plenge et al. [34], which showed that using human genetics can improve the chances of translating a target into a successful therapy.For each protein, we looked at their average pLDDT scores and confidence levels from AlphaFold models and only selected proteins which had an average pLDDt score of High. Finally, we used P2Rank to predict each protein's binding sites and only kept proteins for which we had at least one site with a probability above 50%.

Protein structures were preprocessed using a custom pipeline built on PDBFixer and RDKit. Structures were loaded from AlphaFold .pdb files. Nonstandard amino acid residues were identified and replaced with their standard equivalents to ensure consistency. All heterogens, including ligands, cofactors, and metal ions, were removed from the structure. Water molecules were also removed to reduce noise in the structural representation whilst missing hydrogen atoms were added at physiological pH (7.4). Finally, the cleaned and protonated structure was written to disk in PDB format and parsed into an RDKit molecule object for further processing [9].

## Docking protocol

For each protein, we selected the binding site found by P2Rank with the highest probability. Then, using its coordinates, we built a cubic bounding region of 5 Å around the pocket using DeepChem.

Each box extends 5 Å in every direction from the center, resulting in a cube with 10 Å side length. The padding is automatically scaled based on the ligand's dimensions to ensure the box is large enough to fully contain it, while still preserving the 5 Å margin around the ligand.

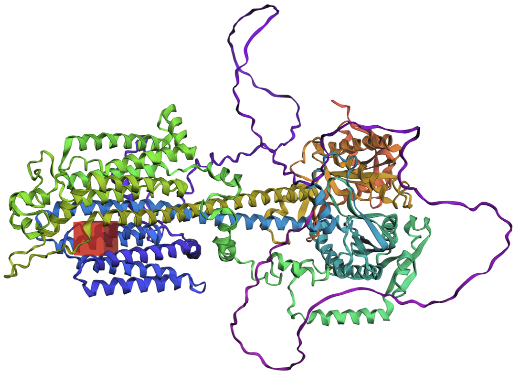

Figure 6: Binding pocket (in red) found by P2Rank on adcy5.

Lastly, we used AutoDock Vina through its Python extension to perform the docking, specifying 5 poses per ligand and an exhaustiveness level of 8 [10]. We only kept the best score for each protein-ligand combination out of the 5 poses (i.e. the lowest score). The computations were executed on a High-Performance Computing (HPC) cluster, taking approximately 45 days to complete and 600,000 CPU hours. We split the dataset into three folds by assigning all scores for certain proteins to separate subsets: 7 proteins for training, 1 for validation and 3 for testing. This setup provides a more realistic and challenging evaluation scenario, as models must generalize to unseen proteins rather than memorize patterns specific to a single target. It also avoids the overfitting risk of standard random splits, where the same protein might appear in both training and test sets.

## Transformer architecture

To inform ML researchers and benchmark our dataset, we built a novel Transformer method to predict docking scores. We followed an embedding-based approach and used two foundation models to encode the protein and the ligand and perform the docking (Figure 7).

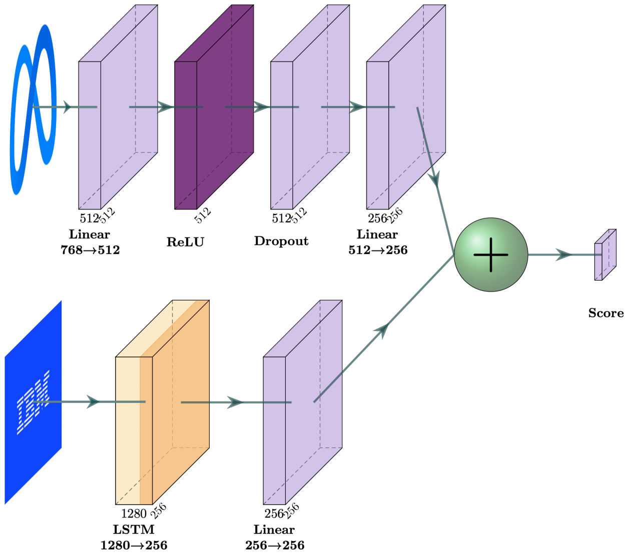

Figure 7: Architecture of the hybrid LSTM-FFN protein-ligand model.

The first model is ESM2, developed by Facebook AI Research (FAIR), which was pretrained on 250 million protein sequences comprising 86 billion amino acids [26]. Its learned representations capture both local biochemical properties and long-range structural patterns, including secondary and tertiary structures.

The second model is MolFormer, developed by IBM Research, which combines masked language modeling with a linear attention Transformer and rotary positional embeddings [36]. It was pretrained on 1.1 billion canonical SMILES strings from ZINC and PubChem. Canonicalization was performed using RDKit to ensure consistency in representation. The model learns compact embeddings of molecular structures and was fine-tuned on a range of downstream tasks.

We used both models to encode our set of 1.7 million molecules from ChEMBL and 16 proteins from AlphaFold. The ESM2 model generates an embedding matrix $\mathbf{E} \in \mathbb{R}^{n\times 1280}$, where $n$ is the length of the protein sequence. Given the variability in protein sequence lengths, we pad all embeddings to match the length of the longest protein, which is 1990, resulting in a final matrix $\mathbf{E}_{\text{padded}} \in \mathbb{R}^{1990\times 1280}$. The MolFormer model produces a fixed-size vector $\mathbf{V} \in \mathbb{R}^{768}$ for each molecule. Formally, for a protein sequence $P_i$ of length $n_i$ and a molecule $M$, their embeddings are represented as:

$$\mathbf{E}(P_i) = \begin{bmatrix} e_{1,1} & e_{1,2} & \cdots & e_{1,1280} \\ e_{2,1} & e_{2,2} & \cdots & e_{2,1280} \\ \vdots & \vdots & \ddots & \vdots \\ e_{n_i,1} & e_{n_i,2} & \cdots & e_{n_i,1280} \end{bmatrix} \in \mathbb{R}^{n_i\times 1280} \quad \text{and} \quad \mathbf{V}(M) = \begin{bmatrix} m_1 \\ m_2 \\ \vdots \\ m_{768} \end{bmatrix} \in \mathbb{R}^{768}$$

### Final hybrid model

The docking model, implemented using PyTorch [33], is designed to predict the interaction between ligands and proteins through a specialized architecture combining separate sub-models for ligands and proteins. The ligand sub-model is a feedforward neural network, starting with an input dimension of 768, matching the size of the MolFormer embedding. It includes two linear layers with a ReLU activation and a dropout layer for regularization. The protein sub-model uses an LSTM (Long Short-Term Memory) network to process sequential data, taking inputs with a dimension of 1280 to match the size of the input embeddings from ESM2[18]. The output of the LSTM is further processed through a linear layer to produce features that align in size with the ligand sub-model.

The model's forward pass processes the ligand and protein embeddings through their respective sub-models then concatenates these features into a combined vector. This vector is passed through a regression layer that outputs the docking score prediction. The training phase involves calculating the RMSE between predicted and actual scores and optimizing this loss using the Adam optimizer [22] with a learning rate of $1 \times 10^{-4}$. We trained our models on an HPC cluster using a multi-GPU setup with 8 Nvidia Tesla V100 (256 GB of VRAM in total). Each model was trained for 2 epochs on the train set and used the validation set to print out the RMSE while training to look for signs of overfitting.

## Conclusion

We introduce Smiles2Dock, an open large-scale comprehensive dataset for training and benchmarking ML-based protein-ligand docking algorithms from AlphaFold predicted structures. It uses well-known chemical data sources such as AlphaFold and ChEMBL, a diverse set of biologically relevant compounds on the same scale as modern molecular screening databases and is suitable for most major approaches explore such as CNN, graph and embedding based methods. Moreover, existing datasets rely on proteins with experimentally determined structures and known ligand binding pockets, making them unusable for the growing number of proteins with only predicted structures. It is easy to use for ML researchers and can be downloaded using two lines of code and a single library using the Datasets library from HuggingFace. We also introduce a novel Transformer-based architecture that uses ESM2 and Molformer to embed molecules and proteins in latent spaces and predict docking scores.

# References

[1] Tunde Aderinwale, Charles Christoffer, and Daisuke Kihara. Rl-mlzerd: Multimeric protein docking using reinforcement learning. *Frontiers in Molecular Biosciences*, 9:969394, 2022.

[2] Vance W Berger and YanYan Zhou. Kolmogorov–smirnov test: Overview. *Wiley statsref: Statistics reference online*, 2014.

[3] Bin Chong, Yingguang Yang, Zi-Le Wang, Han Xing, and Zhirong Liu. Reinforcement learning to boost molecular docking upon protein conformational ensemble. *Physical Chemistry Chemical Physics*, 23(11):6800–6806, 2021.

[4] Lee-Shin Chu, Jeffrey A Ruffolo, Ameya Harmalkar, and Jeffrey J Gray. Flexible protein–protein docking with a multitrack iterative transformer. *Protein Science*, 33(2):e4862, 2024.

[5] Austin Clyde, Xuefeng Liu, Thomas Brettin, Hyunseung Yoo, Alexander Partin, Yadu Babuji, Ben Blaiszik, Jamaludin Mohd-Yusof, Andre Merzky, Matteo Turilli, Shantenu Jha, Arvind Ramanathan, and Rick Stevens. Ai-accelerated protein-ligand docking for sars-cov-2 is 100-fold faster with no significant change in detection. *Scientific Reports*, 13:2105, February 2023.

[6] Ryan G. Coleman, Michael Carchia, Teague Sterling, John J. Irwin, and Brian K. Shoichet. Ligand pose and orientational sampling in molecular docking. *PLoS ONE*, 8(10):e75992, October 2013.

[7] Kevin Crampon, Alexis Giorkallos, Myrtille Deldossi, Stéphanie Baud, and Luiz Angelo Steffenel. Machine-learning methods for ligand–protein molecular docking. *Drug Discovery Today*, 27(1):151–164, January 2022.

[8] Christopher DeBoever, Yosuke Tanigawa, Malene E Lindholm, Greg McInnes, Adam Lavertu, Erik Ingelsson, Chris Chang, Euan A Ashley, Carlos D Bustamante, Mark J Daly, et al. Medical relevance of protein-truncating variants across 337,205 individuals in the uk biobank study. *Nature communications*, 9(1):1612, 2018.

[9] Peter Eastman, Jason Swails, John D Chodera, Robert T McGibbon, Yutong Zhao, Kyle A Beauchamp, Lee-Ping Wang, Andrew C Simmonett, Matthew P Harrigan, Chaya D Stern, et al. Openmm 7: Rapid development of high performance algorithms for molecular dynamics. *PLoS computational biology*, 13(7):e1005659, 2017.

[10] 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, July 2021.

[11] 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.

[12] Miguel García-Ortegón, Gregor N. C. Simm, Austin J. Tripp, José Miguel Hernández-Lobato, Andreas Bender, and Sergio Bacallado. Dockstring: Easy molecular docking yields better benchmarks for ligand design. *Journal of Chemical Information and Modeling*, 62(15):3486–3502, July 2022.

[13] Anna Gaulton, Louisa J Bellis, A Patricia Bento, Jon Chambers, Mark Davies, Anne Hersey, Yvonne Light, Shaun McGlinchey, David Michalovich, Bissan Al-Lazikani, et al. Chembl: a large-scale bioactivity database for drug discovery. *Nucleic acids research*, 40(D1):D1100–D1107, 2012.

[14] Daniel A Gschwend, Andrew C Good, and Irwin D Kuntz. Molecular docking towards drug discovery. *Journal of Molecular Recognition: An Interdisciplinary Journal*, 9(2):175–186, 1996.

[15] Linyuan Guo, Tian Qiu, and Jianxin Wang. Vitscore: a novel three-dimensional vision transformer method for accurate prediction of protein-ligand docking poses. *IEEE transactions on nanobioscience*, 2023.

[16] Linyuan Guo and Jianxin Wang. Vitrmse: a three-dimensional rmse scoring method for protein-ligand docking models based on vision transformer. In *2022 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, pages 328–333. IEEE, 2022.

[17] Ye Han, Fei He, Yongbing Chen, Wenyuan Qin, Helong Yu, and Dong Xu. Quality assessment of protein docking models based on graph neural network. *Frontiers in Bioinformatics*, 1:693211, 2021.

[18] Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural computation*, 9(8):1735–1780, 1997.

[19] Huaipan Jiang, Mengran Fan, Jian Wang, Anup Sarma, Shruti Mohanty, Nikolay V Dokholyan, Mehrdad Mahdavi, and Mahmut T Kandemir. Guiding conventional protein–ligand docking software with convolutional neural networks. *Journal of chemical information and modeling*, 60(10):4594–4602, 2020.

[20] Huaipan Jiang, Jian Wang, Weilin Cong, Yihe Huang, Morteza Ramezani, Anup Sarma, Nikolay V. Dokholyan, Mehrdad Mahdavi, and Mahmut T. Kandemir. Predicting protein–ligand docking structure with graph neural network. *Journal of Chemical Information and Modeling*, 62(12):2923–2932, June 2022.

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

[22] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

[23] Thomas Kluyver, Benjamin Ragan-Kelley, Fernando Pérez, Brian E Granger, Matthias Bussonnier, Jonathan Frederic, Kyle Kelley, Jessica B Hamrick, Jason Grout, Sylvain Corlay, et al. Jupyter notebooks-a publishing format for reproducible computational workflows. *Elpub*, 2016:87–90, 2016.

[24] Radoslav Krivák and David Hoksza. P2rank: machine learning based tool for rapid and accurate prediction of ligand binding sites from protein structure. *Journal of Cheminformatics*, 10(1), August 2018.

[25] Quentin Lhoest, Albert Villanova del Moral, Yacine Jernite, Abhishek Thakur, Patrick von Platen, Suraj Patil, Julien Chaumond, Mariama Drame, Julien Plu, Lewis Tunstall, et al. Datasets: A community library for natural language processing. *arXiv preprint arXiv:2109.02846*, 2021.

[26] Zeming Lin, Halil Akin, Roshan Rao, Brian Hie, Zhongkai Zhu, Wenting Lu, Nikita Smetanin, Robert Verkuil, Ori Kabeli, Yaniv Shmueli, Allan dos Santos Costa, Maryam Fazel-Zarandi, Tom Sercu, Salvatore Candido, and Alexander Rives. Evolutionary-scale prediction of atomic level protein structure with a language model, July 2022.

[27] Andreas Luttens, Israel Cabeza de Vaca, Leonard Sparring, Ulf Norinder, and Jens Carlsson. Rapid traversal of ultralarge chemical space using machine learning guided docking screens, May 2023.

[28] John I Marden. Positions and qq plots. *Statistical Science*, pages 606–614, 2004.

[29] Andrew T McNutt, Paul Francoeur, Rishal Aggarwal, Tomohide Masuda, Rocco Meli, Matthew Ragoza, Jocelyn Sunseri, and David Ryan Koes. Gnina 1.0: molecular docking with deep learning. *Journal of cheminformatics*, 13(1):43, 2021.

[30] Xuan-Yu Meng, Hong-Xing Zhang, Mihaly Mezei, and Meng Cui. Molecular docking: a powerful approach for structure-based drug discovery. *Current computer-aided drug design*, 7(2):146–157, 2011.

[31] Garrett M Morris and Marguerita Lim-Wilby. Molecular docking. *Molecular modeling of proteins*, pages 365–382, 2008.

[32] Marina A Pak, Karina A Markhieva, Mariia S Novikova, Dmitry S Petrov, Ilya S Vorobyev, Ekaterina S Maksimova, Fyodor A Kondrashov, and Dmitry N Ivankov. Using alphafold to predict the impact of single mutations on protein stability and function. *Plos one*, 18(3):e0282689, 2023.

[33] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Köpf, Edward Yang, Zach DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. Pytorch: An imperative style, high-performance deep learning library, 2019.

[34] Robert M Plenge, Edward M Scolnick, and David Altshuler. Validating therapeutic targets through human genetics. *Nature reviews Drug discovery*, 12(8):581–594, 2013.

[35] Nornadiah Mohd Razali, Yap Bee Wah, et al. Power comparisons of shapiro-wilk, kolmogorov-smirnov, lilliefors and anderson-darling tests. *Journal of statistical modeling and analytics*, 2(1):21–33, 2011.

[36] Jerret Ross, Brian Belgodere, Vijil Chenthamarakshan, Inkit Padhi, Youssef Mroueh, and Payel Das. Large-scale chemical language representations capture molecular structure and properties. *Nature Machine Intelligence*, 4(12):1256–1264, December 2022.

[37] Saori Sakaue, Masahiro Kanai, Yosuke Tanigawa, Juha Karjalainen, Mitja Kurki, Seizo Koshiba, Akira Narita, Takahiro Konuma, Kenichi Yamamoto, Masato Akiyama, et al. A cross-population atlas of genetic associations for 220 human phenotypes. *Nature genetics*, 53(10):1415–1424, 2021.

[38] Francesca Stanzione, Ilenia Giangreco, and Jason C. Cole. *Use of molecular docking computational tools in drug discovery*, page 273–343. Elsevier, 2021.

[39] Yosuke Tanigawa, Jiehan Li, Johanne M Justesen, Heiko Horn, Matthew Aguirre, Christopher DeBoever, Chris Chang, Balasubramanian Narasimhan, Kasper Lage, Trevor Hastie, et al. Components of genetic associations across 2,138 phenotypes in the uk biobank highlight adipocyte biology. *Nature communications*, 10(1):4064, 2019.

[40] Yosuke Tanigawa, Michael Wainberg, Juha Karjalainen, Tuomo Kiiskinen, Guhan Venkataraman, Susanna Lemmelä, Joni A Turunen, Robert R Graham, Aki S Havulinna, Markus Perola, et al. Rare protein-altering variants in angptl7 lower intraocular pressure and protect against glaucoma. *PLoS Genetics*, 16(5):e1008682, 2020.

[41] Wen Torng and Russ B Altman. Graph convolutional neural networks for predicting drug-target interactions. *Journal of chemical information and modeling*, 59(10):4131–4149, 2019.

[42] Viet-Khoa Tran-Nguyen, Célien Jacquemard, and Didier Rognan. Lit-pcba: an unbiased data set for machine learning and virtual screening. *Journal of chemical information and modeling*, 60(9):4263–4273, 2020.

[43] Oleg Trott and Arthur J. Olson. Autodock vina: Improving the speed and accuracy of docking with a new scoring function, efficient optimization, and multithreading. *Journal of Computational Chemistry*, 31(2):455–461, June 2009.

[44] Xiao Wang, Sean T Flannery, and Daisuke Kihara. Protein docking model evaluation by graph neural networks. *Frontiers in Molecular Biosciences*, 8:647915, 2021.

[45] Clark Wissler. The spearman correlation formula. *Science*, 22(558):309–311, 1905.

[46] Liangzhen Zheng, Jingrong Fan, and Yuguang Mu. Onionnet: a multiple-layer intermolecular-contact-based convolutional neural network for protein–ligand binding affinity prediction. *ACS omega*, 4(14):15956–15965, 2019.