UNIVERSITY OF
CAMBRIDGE

Sequence-based Binding Kinetics Prediction: A Case Study of Medin

Author: Carolina Moretti Ierardi

Supervisor: Professor Michele Vendruscolo

This dissertation is submitted in partial fulfillment of the requirements for the MPhil in Computational Biology

Abstract

The selection of candidate compounds in drug discovery is crucial to find strong leads with higher chances of going to later clinical trials phases. Deep learning presents a useful tool to predict the binding affinity (pk_D) of a compound to a target protein, and screen large molecular libraries. pk_D is defined as the difference between the unbinding rate and the binding rate ($pk_{off} - pk_{on}$) of a protein-ligand complex. Although these two kinetic parameters are often correlate with drug potency [1], efficient methods to predict these rates based on a complex's structure are lacking. In this study, we extend the architecture of the Ligand-Transformer [2] to predict pk_{on} and pk_{off}, and apply the model to Medin, an amyloid protein aggregated in blood vessels and related to different forms of dementia [3]. An evaluation of our model's performance showed a Pearson's $r_{pkoff} = 0.79$, $r_{pkon} = 0.74$ and $r_{pkd} = 0.73$, where ground truth binding kinetic data was available. We also predicted pk_D values with the difference of pk_{off} and pk_{on} with r = 0.75, where only ground truth pk_D data was available. A prediction of Medin's structure revealed this protein is ordered rather than intrinsically disordered, contrary to previous reports. Thirty molecules with high binding affinity to this protein were selected. We predicted the selected molecules form a binding pocket stretching across the protein's β-hairpin structure and thus, highly topologically interconnected molecules are more likely to have higher binding affinity to Medin. Overall, we have created a model with good performance in predicting binding kinetic parameters, and identified important binding mechanisms in Medin. These mechanisms and their effect on amyloid fibril formation will be examined in aggregation assays.

Declaration

I hereby declare that this dissertation entitled "Sequence-based Binding Kinetics Prediction: A Case Study of Medin" is the result of my own work and includes nothing which is the outcome of work done in collaboration except as declared in the Preface and specified in the text. I further state that no substantial part of this dissertation has already been submitted, or, is being concurrently submitted for any such degree, diploma or other qualification at the University of Cambridge or any other University or similar institution except as declared in the Preface and specified in the text. I confirm that I have read and understood the Faculty of Mathematics Guidelines on Plagiarism and the University-wide Statement on Plagiarism.

Author: Carolina Moretti Ierardi
August, 2024

Acknowledgements

I would like to thank Professor Michele Vendruscolo for his support and guidance throughout the project. Thank you also to Shengyu Zhang, Sebastian Pujalte Ojeda and Michaela Brezinova for all the assistance, input and support given throughout the project. I would also like to thank my friends and my family.

Contents

1 Background **7**

Protein structure prediction . . . 7
Intrinsically Disordered Proteins . . . 9
Drug discovery for IDPs . . . 10
Motivation for the project . . . 10
Prediction of Binding Kinetics . . . 11
Case study: Drug Design for Medin . . . 13

2 Methods **14**

Datasets collection and pre-processing . . . 14
Binding Kinetics Data . . . 14
PDBbind dataset . . . 15
Protein Representations from AlphaFold2 . . . 15
Ligand Representations from GraphMVP . . . 15
Ligand-Transformer Kinetics predictions . . . 16
Medin Case Study Methodology . . . 18
IUPred . . . 18
Binding Affinity Predictions . . . 19
Molecular Clusters . . . 19
Protein Binding Sites . . . 20
Target Specificity . . . 20
PaDEL Molecular Descriptors . . . 20
Random Forest . . . 21

3 Results **22**

Prediction of Binding Kinetics . . . 22
Data Exploration . . . 22
Internal validation . . . 23
External validation . . . 27
Case study: Drug Design for Medin . . . 29

Medin Structure 29
Molecular library screening 31
Predicted binding site 34
Medin Specificity 35
Molecular Descriptors Driving Medin Binding 37

4 Discussion **40**

Bibliography **44**

A Supplementary figures **53**

Chapter 1

Background

The field of protein structure prediction recently achieved a breakthrough with AlphaFold2 [4], a deep learning model using end-to-end learning to predict a protein's 3D structure from just its amino acid, or residue, sequence. This method has been used to predict the structure of unmapped proteins in combination with experimental methods [5], generate molecules for drug discovery [6], and design novel proteins [7].

Although AlphaFold has good performance on a subset of folded proteins, questions remain as to whether its predictions for intrinsically disordered proteins (IDPs) are also accurate. Moreover, we have yet to use the model to find drugs targeting IDPs. In the next section, we will provide background on protein structure prediction, IDPs and drug discovery for them as well as motivate the current project.

Protein structure prediction

Proteins are RNA-encoded molecules composed amino-acid residues joined by peptide bonds. They are needed for basic life maintenance and needed for DNA replication, providing cell structure, making up the immune system, among other functions. Protein structure prediction has been a great challenge in computational biology. In 1970, it was found that the three-dimensional structure of a protein could be determined by its amino-acid sequence [8]. Due to a great effort from genome sequencing, there are now over 200 million protein sequences in UniProt [9]. However, the sequences themselves contain limited information about a protein's function, which is instead determined by its three-dimensional structure [10]. Although there are established methods to derive a protein structure experimentally, these are expensive and time-consuming [11], creating a large difference between the number of known protein sequences and the number of protein structures mapped. Currently, the number of protein structures available on Protein Data Bank (PDB) [12] amounts to 0.1% of the number of protein sequences available in UniProt. Thus, protein structure modelling presents an opportunity to map protein structures at a

faster rate than experimentally possible.

Protein structure modelling can be of two types: template-based modelling (TBM) or template-free modelling (TFM). TBM methods rely on homologue structure mapping. The methods use proteins with structures already on the PDB that have similar sequences to a given query protein, assuming similar sequences will lead to similar structures. Because of this, the quality of TBM depends on template similarity. Scholnick and Zhou have argued that this method fails in achieving high accuracy levels as some proteins have specific structural characteristics which break the assumption that similar templates have similar structures, claiming for new methods to predict protein structures [13].

TFM, however, relies on rules of physics, energy functions and sampling methods to map a protein's structure. Examples of such methods include molecular dynamics and fragment assembly. In molecular dynamics, a structure's native state is the conformation with the lowest free energy. Based on the residue properties, methods can combine energy functions from typical bond lengths and angles, dihedral angles and van der Waals interactions to minimize a global energy function and refine the crystal structure of a protein. A challenge of this procedure is its computational expense and difficulty to apply to a protein sequence. In contrast, fragment assembly aims to reduce the conformational space whilst accurately mapping the local structures of a protein. To achieve this, existing structures from the PDB are fragmented into smaller residue sequences and used to create the structure of a query protein based on their sequence similarity [14]. A popular tool that uses this method is Rosetta [15]. Using fragment assembly and simulated annealing Monte Carlo simulations, this tool finds fragments most similar to the query protein. It also considers many physical properties of folding proteins to yield an accurate structural prediction.

Recently, AlphaFold2 [4] achieved a major breakthrough by accurately predicting the 3D structures of proteins previously mapped using experimental approaches and available on the PDB. As a TFM method, AlphaFold2 was the first to achieve accuracy as high as a TBM method. This model uses deep learning to predict the 3D structure of a protein from its sequence. AlphaFold2 is trained using data from PDB. To predict a novel protein's structure, it uses multiple sequence alignment (MSA) to find similar proteins to the query protein according to its sequence. The model uses attention mechanisms to process the MSA and refine the prediction. This way, it can calculate the inter-residue distances and angles. The model then uses gradient descent to obtain 3D coordinates for each atom in the protein. AlphaFold outputs a predicted structure with an associated confidence score for each protein segment; low-confidence regions tend to overlap with sections of the protein that are disordered. Therefore, there is a higher level of uncertainty regarding AlphaFold's predictions for IDPs, which is discussed in more detail in the next section.

Intrinsically Disordered Proteins

IDPs do not have a stable three-dimensional structure under physiological conditions. Instead, these proteins have a high rate of conformational flexibility [16]. It has been found that at least 33% of eukaryotic proteins are IDPs [17] and around 70% of proteins in the human proteome are intrinsically disordered regions (IDRs) of at least 30 residues [18]. Hence, most proteins are a combination of ordered and disordered regions. Rather than having one stable structure, IDPs/IDRs' structures are best described as ensembles of conformations. This means IDPs/IDRs have a set of conformations they can adopt that may change spontaneously or as a function of their environment. Recent research has described this ensemble as having statistical weights that follow a Boltzmann distribution [19]. Thus, the probability of being in a given protein conformation depends on the state's energy and temperature.

Because IDPs/IDRs lack a stable structure, they represent an exception to the sequence-structure-function rule that ordered proteins follow [20]. On the contrary, it has been suggested these proteins follow a sequence-ensemble-function rule [18]. The ensemble properties of an IDP can be defined by its residue patterns, which describes how clustered residues of a chemical group are within a sequence. Variations in residue patterns determine the ensemble properties through the electrostatic interactions of attraction and repulsion. These properties in turn play a key role in the protein's function. A recent study [21] has performed *in vivo* experiments showing that a shift in the dimensions of IDRs can cause cell volume change. Although this was performed synthetically, it demonstrated that changes to the ensemble can lead to different molecular functions.

Importantly, IDPs play roles in different human disorders such as cancer, neurodegenerative and cardiovascular diseases [16], leading to the concept of D^2, or "Disorder to Disorders" [22]. Examples of these IDPs include p53 and c-Myc, which are oncoproteins, α-synuclein, involved in Parkinson's disease and tau and amyloid-β, which lead to Alzheimer's disease. Commonly, neurodegenerative-related proteins are IDPs and very often aggregate, leading to disease progression [23]. Therefore, studying the molecular mechanisms of IDPs in diseases is extremely important to understand how to better treat and diagnose these conditions, hence these proteins have become common drug targets. To illustrate, it has been found that using a nutlins molecule to target p53 leads to the protein's apoptosis and inhibits the cell growth in tumours [24]. Compounds such as curcumin, ferulic acid and safranal have been found also to inhibit the aggregation of α-synuclein by binding to the protein [25]. The conformational fluctuations, however, make the binding of IDPs/IDRs unstable and hard to predict. Instrumental in causing many diseases, IDPs represent a significant opportunity for drug discovery, but their inherent instability poses technical challenges to targeting their structures.

Drug discovery for IDPs

A common way to measure the suitability of a drug or molecule to target a specific protein is to measure the strength of the interaction between two partners, such as a ligand and a receptor, referred to as the binding affinity of the complex [26]. Binding affinity is usually measured by a system's equilibrium dissociation constant (K_D). This is defined by the likelihood of the bound ligand-receptor complex separating into free ligand and receptor. It is related to kinetic rate constants, k_{on} and k_{off}, through the equation

$$K_D = K_{off}/K_{on} \tag{1.1}$$

where k_{off} corresponds to the rate at which a ligand dissociates from its receptor and k_{on} is the rate at which the ligand binds to the receptor [27]. It is a common practice to transform these values into logspace so the scale of these variables is more interpretable, to obtain

$$pK_x = -log_{10}(\frac{K_x}{1M}), K_x \in [K_D, K_{on}, K_{off}] \tag{1.2}$$

where $1M = 1molar$.

Therefore, the pK_D is directly related to the binding affinity between a ligand and its receptor.

Interestingly, the binding affinity between IDPs/IDRs and other molecules is highly modulated by their structure. For instance, it has been shown that changing the residual helicity, the proportion of alpha helixes in the structure, of p53 resulted in stronger binding to Mdm2, a common binder of this protein [28]. Other research [29] found that modulating the secondary structure of IDPs/IDRs with ions in solution can change the binding affinity to ligands by up to sixfold. The paper also demonstrated this change is caused by impact on both the association (K_{on}) and dissociation rates (K_{off}) of the complex. Examples such as these showcase the importance of IDPs/IDRs' ensemble properties not only for their function but also for their kinetic relationship with other molecules in the environment.

Motivation for the project

Thus far we have established the challenge of protein structure prediction in computational biology and the breakthrough of AlphaFold2. Further, we characterised IDPs/IDRs and how they can follow sequence-ensemble-function relationships. Finally, the involvement of IDPs/IDRs in diseases and the importance and difficulties of considering them as drug targets were discussed.

Despite the most accurate description of IDP/IDR structures to be conformational ensembles, it has been found that AlphaFold2 is able to predict inter-residue distances

in disordered proteins. Importantly, for proteins with IDRs that only fold under specific conditions, AlphaFold2 predicts the conditionally folded state with an 88% precision and a 10% false positive rate [30]. When we consider such proteins were minimally represented in the training data, these results are surprising.

Using this, a recent project has aimed to design drugs targeting disordered proteins using a deep learning model [2]. More specifically, the Vendruscolo group built a transformer that uses protein representations from AlphaFold2 and ligand representations from GraphMVP [31], pre-training framework uses knowledge of the 3D structure of a molecule to inform its 2D representation. This is advantageous for the model as it contains information on the structural features of the protein and ligand in complex. Once the protein and molecular features are computed, they are used as inputs into the model. Its initial component involves feature encoders that re-process the representations of proteins and ligands. The second component is a cross-modal attention network that facilitates the exchange of information between the protein and ligand representations. The third component contains two downstream predictors: one focused on predicting binding affinity (pK_D) and the other on predicting distance within the protein-ligand complex.

The Ligand-Transformer has been experimentally validated using α-synuclein and amyloid-β (Aβ). In these cases, molecules were screened using the model and obtained predicted binding affinity values. The molecules with the highest affinities were tested *in vitro* in aggregation assays to examine whether the binding of such molecules will delay aggregation in the sample. The expectation is that if the aggregation is delayed, this will have positive effects on delaying symptoms and disease progression in neurodegenerative disorders.

The goals of the present project are twofold. The first is to extend the model to predict not only pK_D but also pK_{on} and pK_{off}. The second is to apply this model to a novel protein, Medin [32], which is found to co-aggregate with Aβ in Alzheimer's disease [3]. Motivation and background for the specific research questions are given within each subsection below.

Prediction of Binding Kinetics

As mentioned in equation 1.1, binding affinity (pK_D) is defined by the ratio of dissociating molecules and associating molecules within a certain time frame. Although binding affinity is primarily used to determine how well two agents in complex bind, it is important to understand the characteristics of the individual association and dissociation rates in drug discovery. In particular, pK_{off} is important to quantify how quickly a molecule will dissociate from a protein; lower values of pK_{off} may reveal the need for lower doses of a drug. In cases where the drug may have adverse toxic effects, a shorter residence time (higher pK_{off}) will be required [33]. In contrast, pK_{on} reveals the molecular recognition

speed. Ideal values for this parameter can depend on whether the target disease is acute or chronic, for instance [34].

The ratio between these two rates can be the same with widely different kinetic behaviours underlying them. In some cases, the driver of ligand/drug potency is k_{off}, instead of K_D [35, 36]. Drug-target residence times have also been found to promote drug efficacy in studies done with specific inhibitors of G-coupled protein receptors (GPCRs) [37]. In light of this, there is a need to include binding kinetics into the pipeline of drug discovery [38]. In cases where the target's mechanism of action is unknown, it may be useful to sample ligands with varying kinetic characteristics to test. When there is a known desired mechanism for the ligand, one can select leads based on binding kinetics as well as binding affinity estimations. Hence, including binding kinetics in the process of lead selection and optimization in drug discovery is an essential step to improve drug efficacy.

Experimentally measuring binding kinetics is costly and time-consuming. Therefore, estimation of these parameters can take advantage of computational methods. To date, little to no predictors of pK_{on} are found in the literature. In contrast, the existing predictors of pK_{off} focus on specific protein case studies. Examples of this include MD-based models predicting HSP90 proteins' pK_{off} with $R^2 = 0.66$ [39]; Random Forest models to classify HIV-1 inhibitors with 74% accuracy [1]; deep learning models to predict RPK1 proteins with $R^2 = 0.74$; and a Partial Least Squares Regression model to predict 20 protein kinase Type II inhibitors with $R^2 = 0.56$. Approaches to predict a diverse set of protein-ligand complexes to date have only been found by Fedorov et al. [40] who collected data and used a Random Forest model to predict pK_{off} of protein-ligand complexes with $R^2 = 0.6$.

To our best knowledge, there are few deep learning models to predict general protein-ligand binding kinetics parameters, likely due to the lack of experimental data on these parameters. Nonetheless, because the Ligand-Transfomer already encodes information on the binding of the complex, we expect it to extend the predictions to other binding kinetic parameters. Thus, we extend the Ligand-Transformer [2] to predict binding kinetic parameters in addition to binding affinity values. We obtained data from the KIND dataset [41] and Fedorov et al. [40] to compose our own dataset. To our knowledge these are the largest datasets containing experimental measures of pK_D, pK_{on} and pK_{off} for protein and small-molecule complexes. We use the latest weights of the trained Ligand-Transformer and perform transfer learning, slightly modifying the network's architecture to also predict binding kinetics data. Thereafter, we make predictions on previous training data of the model, which does not contain ground truth pK_{on} or pK_{off}, and measure performance by comparing to ground truth pK_D using Equation 1.1, assessing how generalisable are our model's predictions.

Case study: Drug Design for Medin

Medin is a form of aortic medial amyloid present in the bloodstream of over 90% of individuals above 50 years old [32]. This IDP is an amyloid precursor formed by the proteolysis of Lactadherin, a mammary epithelial cell, and made up by part of its coagulation factor-like domain. Previous research has shown that patients with vascular dementia have increased Medin levels in the bloodstream compared to other cognitively impaired individuals [42]. It has also been shown that age-associated vascular dysfunction was eliminated in genetically-engineered mice lacking Medin, pointing to the pathological role of the protein [3]. The same study showed the role of this protein in Alzheimer's disease is also evidenced by the co-aggregation of Medin with Aβ. Patients with Alzheimer's have higher levels of Lactadherin expression and both these levels and Medin expression levels are correlated with the severity of vascular Aβ deposition.

As mentioned previously, many proteins that are neurodegenerative disease drivers tend to aggregate in pathology. Many amyloids go from a soluble state in polypeptide form to compose ordered fibrils in insoluble form, which are hard to make soluble again. These fibrils are deposited in the tissue, which lead to pathogenesis [43]. Medin is no exception to this process, aggregating into β-sheet-rich fibrils *in vitro* within 50 hours [44]. Studies have found that the amino acids 42-49 of Medin are highly amyloidogenic and that removing this C-terminus abolishes amyloid forming potential in the protein [45]. It has also been found that substituting Asp^{25} with Aspargine stabilized Medin oligomers, which led the authors to conclude the amino acid Asp^{25} drives medin aggregation [44].

The clear role of Medin in different types of dementia has drawn attention to study its structure. Due to the aforementioned difficulties in studying amyloid proteins in their soluble states, computational methods such as molecular dynamics (MD) or fragment assembly have been employed to infer its structure. To this end, researchers [46] used Rosetta and QUARK, two fragment assembly software tools, as inputs into MD simulations, to infer Medin's structure. They showed the simulations consistently found a 4-stranded β-sheet forming a β-hairpin from residues 21-35 of the protein. Indirect experimental methods to infer a protein's structure also agreed with the simulation results.

To the best of our knowledge, there are no available drugs targeting Medin nor are there successful attempts to find molecules that bind to this protein. In this study, we will describe the Medin structural input into the Ligand-Transformer and use the model to screen small molecules to find those with high binding affinity to the protein. We also aim to characterise the protein's binding sites and the compounds' specificity. With this method, we hope these selected molecules will stabilise Medin fibrils found in vascular dementia, and co-aggregated with Aβ in Alzheimer's disease.

Chapter 2

Methods

Datasets collection and pre-processing

Binding Kinetics Data

We obtained publicly available data from two sources: the KIND dataset [41] and Fedorov et al. [40]. Both of these had protein-ligand data triplet information (pK_D, pK_{off} and pK_{on}). KIND dataset comprises 3812 complexes from 21 publications and the K4DD (available here). Federov et al. sourced data from multiple publications, obtaining 501 complexes in total.

During preprocessing we ensured all data were transformed into logspace according to Equation 1.2. Furthermore, only complexes with information on all three parameters were kept. The KIND dataset did not have direct information regarding the protein sequence used in each experiment. To obtain these, the sources of publication were reviewed to find UniProt/GenBank protein IDs and specific segments of proteins used in the experiments. Where this information was unavailable, we assumed the whole protein sequence was used in the experiment. In the case of the Fedorov et al. dataset, each complex was accompanied of a .pdb file, from which the protein sequences were extracted.

All ligands in the dataset had SMILES (Simplified molecular-input line-entry system), line notations that describe a molecule's structure in ASCII strings. All compound SMILES were obtained in their canonical form to ensure no repeats due to non-canonical SMILES representations. To remove duplicates from the data, the following steps were followed: if a complex with multiple entries had a .pdb file available, this entry was kept. If no entries had .pdb files available, and were all from the KIND, the entry with the median pK_D was kept. These measures were taken as the directly available information from .pdb files is more reliable than our manual curation process. Finally, to limit the computational power in training, proteins with sequences over 800 residues were excluded from the analysis.

As the dataset was relatively small, a cross-validation was used to guarantee the model

performance was not due to a specific split of the data. To split the dataset into training, validation and testing, we used a stratified split for protein type and assay method that were used to experimentally determine these parameters. This was done due to differences in the distribution of kinetic parameters according to the type of the protein and assay method (see Results). We left 10% of the data for testing and the remaining 90% were used the 5 fold cross-validation.

Before training, the distributions for pK_{on} and pK_{off} were scaled to have mean $\mu = 0$ and standard deviation $\sigma = 1$. Only these features were scaled as their predictions stem from the same head in the model, and the model was pre-trained to predict pK_D, meaning scaling the new sample would deteriorate model performance significantly. The scaling was done using the mean and standard deviation of the training set and the same transformation was applied to the validation set. This ensures no data leakage between the training and validation in each fold of the cross-validation. After the final epoch, predicted and actual values were re-scaled back to their original distribution.

PDBbind dataset

The PDBbind dataset was used to train the model for external validation, along with KIND. PDBbind is a dataset containing over 20,000 pK_D values for protein-ligand complexes. The pre-processed data, split into training and validation, was obtained directly from the authors of Zhang et al. (2023) [2]. The split was done randomly, leaving around 10% for testing (1,720 complexes) and the remaining for training (19,480 complexes).

Protein Representations from AlphaFold2

AlphaFold2 was employed in ColabFold [47] to pre-compute protein features used in this study. The procedure follows that of of Zhang et al. (2023) [2]. ColabFold batch performs multiple sequence alignments (MSA) for a given query protein. There are then three outputs: the structure representation of the protein, which is a result of the structure module in AlphaFold2; the single representation of the protein, corresponding to a linear projection of the first row of the MSA; the pair representation of the protein, related to the inter-residue distances. The latter two are derived from the Evoformer of AlphaFold2. The structure, single and pair representations are then used as inputs into the Ligand-Transformer.

Ligand Representations from GraphMVP

GraphMVP [31] was used to produce molecular representations in this study. The procedure followed was the same as in Zhang et al. (2023) [2]. Briefly, the pre-trained model transforms 2D ligands into graph representations. We input molecules' SMILES

and the model outputs include atom representations for each atom in the molecules and bond representations for each pair of atoms in the molecule.

Ligand-Transformer for prediction of binding kinetics

The Ligand-Transformer for prediction of binding kinetics follows a very similar architecture to the original model. Figure 2.1 reproduces the model architecture seen in Zhang et al. (2023) [2], highlighting the addition of the Kinetics Head, used to predict pK_{on} and pK_{off}. This head follows the same architecture as the Affinity Head, with the exception of the output layer. Briefly, the Kinetics Head first aggregates embeddings of protein and molecule representation into average embeddings for the protein and molecule. These two are then concatenated to obtain a complex representation. After dropout and normalization, the complex is input into a network with three linear layers and ReLU activation. The output layer has two nodes, one for each kinetic parameter, pK_{on} and pK_{off}.

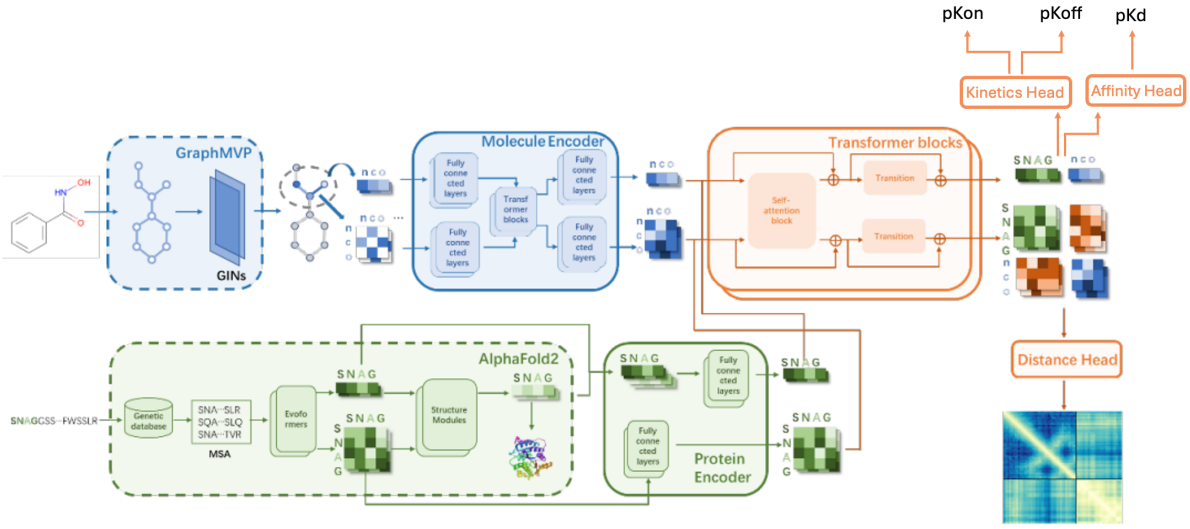

Figure 2.1: **Ligand-Transformer Kinetics architecture** The figure is adapted from Figure S1 from Zhang et al. (2023) [2]. The architecture of the model is shown with the addition of the Kinetic Head. Briefly, the Ligand-Transformer models a protein-ligand complex as a heterogeneous graph, combining residue and atoms from the protein and ligand, respectively. The outputs from AlphaFold2 and GraphMVP are re-encoded into a complete graph, which is then processed in a 12-layer transformer network, updating node and edge representations using self-attention with pair bias. The final output is processed by the Affinity Head for binding affinity prediction and Kinetic Head for binding kinetics prediction. The Distance Head outputs information about the residue-to-ligand distances of the complex as well as information regarding predictions of the protein structure.

Given this architecture, there are two ways to predict binding affinity. The first approach is to directly predict pK_D with the Affinity Head, which has been shown to perform well in previous versions of the model. An alternative approach is to use the

predictions of pK_{on} and pK_{off} to estimate pK_D. If we combine Equations 1.1 and 1.2, we find these values are related by

$$pK_D = pK_{off} - pKon. \tag{2.1}$$

Therefore, we can also use the outputs from the Kinetics Head in order to predict binding affinity. Because there is more data available for binding affinity than binding kinetics, using the predictions of $pK_{off} - pKon$ and measuring the error against the ground truth pK_D will allow us to measure how well the Kinetics Head performs beyond the KIND dataset. Although this is not a direct estimation of the parameters, it can indicate of how well the model performs on unseen data. The definitions of the different affinity calculations are described in Table 2.1.

Table 2.1: Affinity measures defined in the study.

Terms	**Definitions**
True Affinity	pK_D from the experimental dataset
Predicted Affinity	pK_D predicted from the Affinity Head
True Calculated pK_D	pK_{off} - pK_{on} from experimental dataset
Predicted Calculated pK_D	pK_{off} - pK_{on} from prediction of Kinetics Head

The KIND dataset is composed of data obtained in experimental studies. It is common practice to measure the binding parameters for a same protein in complex with many ligands, and the same ligand in complex with many proteins. Therefore, although our dataset contains a unique set of over 3,000 complexes, the same protein or ligand appears in more than one data point. For this reason, we chose to evaluate the performance of the model in two ways. The first uses a "warm setting" [48], meaning both the drug and target in a given complex in the validation set are in the training set, although in different complexes. Therefore, during validation the binding of a protein p is known for a number of drugs d or vice versa. This usually means the predictions will not generalise beyond the training set. The second training method adopted here uses a "cold drug-target setting" where neither protein or drug in the validation set are in the training set. It is expected that models with this setting perform better on unseen data.

To perform the "warm setting" training, the KIND dataset was used to fine-tune a pre-trained version of the Ligand-Transformer. A hyperparameter search was performed using only this dataset and then a 5-fold cross-validation was applied. The hyperparameter search was done with a subset of 30% of the training data and used a Bayesian Optimisation using the skopt library (available here) to decrease the search space. The hyperparameter search tuned the dropout rate of the Affinity and Kinetics Head and learning rate. The learning rate setting could range between $1e{-}7 - 1e{-}4$ and dropout rates could range between $0 - 0.4$. The final hyperparameter set was $Dropout_{affinity} = 0.16$; $Dropout_{kinetics} = 0.09$ and

Learning Rate = 1e−5. During training, predictions were made for each each validation set, composed only of KIND data. The model with the best performance was used to train the entire dataset and predict the test set.

To perform the "cold setting" training, we used the PDBbind in addition to the KIND. Once again, fine-tuning was done using a pre-trained version of the Ligand-Transformer. A new hyperparameter search was done for this combined dataset to adjust the dropout rate for both Kinetics and Affinity Head, and learning rate, once again using a subset of 30% of the data. The hyperparameter search in the same as the "warm-setting" model. The final hyperparameter set was $Dropout_{affinity} = 0.3$; $Dropout_{kinetics} = 0$ and Learning Rate = 1e−4. During training, at each step a data point was sampled from one of the datasets. There was a 50% chance this data point would be from KIND and 50% chance this data point would be from PDBbind. Once $\sim$ 20,000 samples were selected, one epoch of training was performed. In general, one epoch contained the entire KIND and the remaining data was from PDBbind. The model was then trained for 15 epochs.

The performance of the model was measured using the following metrics: R^2 score, Spearman's rank correlation coefficient (ρ) and Pearson's r correlation coefficient using implementations from the torchmetrics library (available here) and mean absolute error (MAE) according to the equation

$$\text{MAE}(y_{predicted}, y_{true}) = \frac{\sum_{i=0}^{N-1} |y_{predicted} - y_{true}|}{N} \tag{2.2}$$

and MSE using

$$\text{MSE}(y_{predicted}, y_{true}) = \frac{\sum_{i=0}^{N-1} (y_{predicted} - y_{true})^2}{N} \tag{2.3}$$

Medin Case Study Methodology

IUPred

We used IUPred2A [49] to predict Medin's disorder. The online tool (available here) was used to input the protein's sequence, found in residues 245 - 294 of Lactadherin (UniProt ID: Q08431) as the query protein. The "short-disorder" setting is used to capture short segments (< 30 residues) the protein, since the protein is only 50 residues long. We also use the context-dependent prediction setting with ANCHOR2 to predict whether a region will undergo disordered binding. In other words, whether the presence of a binding partner is likely to promote transition from ordered to disordered state.

Binding Affinity Predictions

The Ligand-Transformer can be used to run inferences between a query protein and a library of ligands. This directly predicts the binding affinity and inter-residue distances for a protein-ligand complex. We used the Medin sequence, as per the previous section, as the input protein sequence. The ZINC20 library (available here) of purchasable in-stock compounds was used as the molecular library. This library was chosen as it is one of the largest compound libraries freely accessible. The subset of in-stock compounds comprises ∼13 million molecules. Some of these were excluded from the analysis due to issues extracting the molecular features with RDKit (http://www.rdkit.org/). The final screening sample size was 13,663,278.

Protein features were extracted using the method described in Section "Protein Representations from AlphaFold2". Molecular features were extracted using the method described in Section "Ligand Representations from GraphMVP". We then performed inferences to obtain predicted binding affinities for each molecule and Medin. This was done with the original Ligand-Transformer (no Kinetics Head), as the model with the added Kinetics Head is still in development. The original Ligand-Transformer [2] has undergone improvements, and this latest version was applied without and with hyperparameter tuning, Models 1 and 2, respectively.

Molecular Clusters

Based on the screening, we aimed to select 30 molecules to be experimentally validated. These molecules will, in future work, be placed in aggregation assays with Medin to analyse the impact they have on protein amyloid fibril formation. Because neither Model 1 nor Model 2 have been experimentally validated, we chose to select molecules with high predicted affinity in both, or one of the models (see Results). For the selection, in addition to having high binding affinity to Medin, the molecules should also be different amongst themselves in order to better explore the chemical space. To this end, we used the 50,000 molecules with highest predicted binding affinity values to perform clustering. This value was chosen as the maximum number of molecules tolerated by Chemfp, a tool that handles big datasets in chemoinformatics (available here), without obtaining a license.

In Chemfp, we computed Morgan fingerprints [50] for each molecule. Based on these, Tanimoto similarity indexes were extracted for each pair of molecules and these were used as inputs for Butina clustering [51]. Briefly, Morgan fingerprints encode the atom groups of a molecule into a binary vector. Its user-specified parameters are the radius of the atom groups and length of the vector, here radius = 2 and bits = 2048 were used. Tanimoto similarity index refers to the similarity between two molecules' fingerprints. Using the pairwise similarity of each pair of molecules, Butina clustering forms clusters based on a

user-specified similarity threshold, at present a threshold of 0.6 similarity was used. For this particular study, we chose to have 50 final clusters. This meant once the algorithm was complete with the specified threshold, the smallest cluster members were merged into other clusters until there were 50 clusters in total. This number was chosen to ease the selection process of the final 30 molecules, while maintaining sufficient dissimilarity between clusters.

Protein Binding Sites

The Distance Head of the Ligand-Transformer provides information about the predicted distances between each residue of the query protein and the ligand. This is an indication of where the ligand binds in the protein's structure. For the 30 selected molecules, we obtained information about the residue-ligand distances and the probability of binding per residue. We also wished to visualise these residues in a Medin structure. Since experimental data on the protein's structure is not available, we asked the authors of Davies et al. (2017) [46] to share the protein structure they found through MD. The residues with an average probability of binding above 80% across all molecules were considered as the most likely binding sites, as they had high probability of binding in all 30 molecules.

Target Specificity

Although the binding of the selected molecules to Medin was already examined, we wished to measure how each selected molecule bound to every protein in the human proteome. To achieve this, Model 2 was used to predict binding affinities of the 30 selected molecules to proteins in the human proteome. Proteins from the Reference Proteome (UniProt ID: UP000005640), which contains 19,842 proteins in total, were downloaded. To limit computational power needed, protein sequences with over 800 residues were excluded, resulting in the analysis of 17,800 proteins. In order to identify the probability of a protein having a higher binding affinity value to a given molecule than Medin. We used the equation

$$p = \frac{\hat{p}}{N} \tag{2.4}$$

where $\hat{p}$ is the number of samples with a higher binding affinity to a given molecule than Medin and $N = 17,800$, the total sample size.

PaDEL Molecular Descriptors

Our next goal was to unveil the "black-box" model we used. We wished to identify if there were any molecular descriptors that could explain how well a molecule bound to Medin.

We extracted PaDEL Molecular Descriptors [52] for a subset of the ZINC library. This software computes over 1900 descriptors for a given molecule, including 1D, 2D or 3D descriptors.

A continuous relationship was investigated by sampling a subset of molecules using systematic sampling. For a ranked list of binding affinities to Medin in Model 2, molecules were sampled in a fixed interval so the final sample would be ∼100,000 molecules. This value was chosen as a trade off between computational expense and a representative sample of the data. We extracted PaDEL descriptors for these molecules and tested the relationship to binding affinity with a Random Forest Regressor.

A categorical relationship was investigated by sampling the 50,000 molecules with highest and lowest binding affinity predictions in Model 2. The same procedure was followed as described above for a Random Forest Classifier.

Random Forest

The PaDEL descriptors were used as features in the Random Forest models using RandomForestRegressor and RandomForestClassifier functions from the scikit-learn implementation. In each case, molecules with any NaN values for any of the features were excluded from the analysis. In total, 90,738 and 90,132 molecules were kept in the regression and classification models, respectively.

The datasets were randomly split leaving 10% of the data to the test set. Hyperparameters for the model included max depth = 2, and other default arguments in scikit-learn functions. No hyperparameter optimization was done at this point as the goal was not to maximise model performance, but rather investigate the molecular features which could be driving binding affinity to Medin.

Chapter 3

Results

Prediction of Binding Kinetics

Data exploration

The full curated dataset contained 3822 unique protein-ligand complexes. After data exclusion based on sequence length, the final dataset used in this study comprised 3537 complexes. We will refer to this input dataset as KIND, for simplicity. Figure 3.1A-C shows the distribution of parameters in the final dataset, indicating no obvious outliers in the data. We also see the correlations between the three parameters (3.1D-F). There is a strong negative correlation between pK_D and pK_{on} (r = -0.73), possibly indicating that improved affinity measures have underlying rapid binding rates.

The final data contained complexes with 5 different types of protein: kinases, G-coupled protein receptors (GPCR), heatshock proteins (HSP) and enzymes. Some of the protein entries had no accompanying data, and were labelled as "Unknown". We also had data from the experiments used to determined the parameters in the data. Although the vast majority of data are from Fluorescent ligand-binding assays, other experimental methods included RadioLigand and Surface Plasmon Resonance (SPR, Figure 3.2). Because we can see a difference in metrics for protein types and assay methods, cross-validation splits were stratified for these variables (see Results).

Once the splits were made, we wanted to also ensure the chemical space covered by the test and validation sets was representative of training set. Figure 3.3 shows the chemical space covered by the test set relative to the training set to demonstrate good chemical space coverage in the test set. The absence of any clear concentration of data points in the graph, indicates no molecules are over-represented in the test set. The same was done for each split of the cross-validation (Figure A.1).

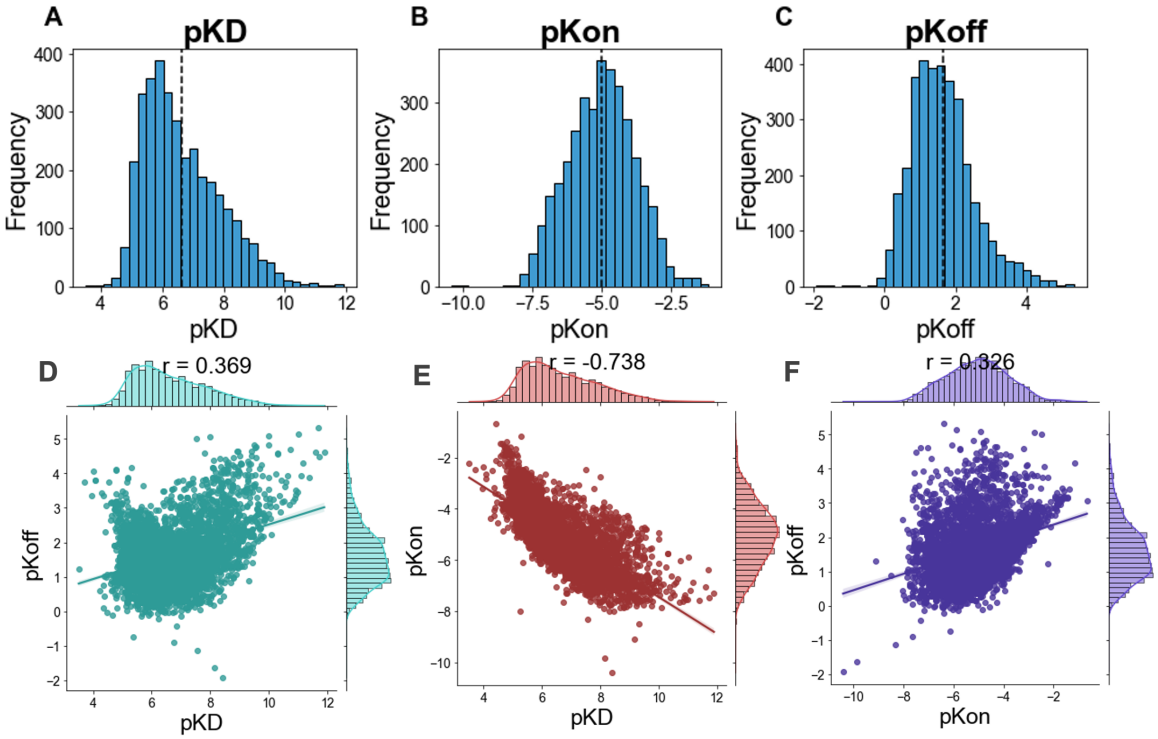

Figure 3.1: **Binding kinetic parameters in dataset** Histogram with pK_D (A), pK_{on} (B) and pK_{off} (C) for the KIND dataset. We also show correlations between each pair of metrics (D-F).

Internal validation

First, we wished to verify that predictions for the binding kinetics parameters could be made for complexes within the KIND dataset. The dataset contains several entries of the same protein in complex with different ligands as well as information on the same ligand bound to different proteins. Therefore, with hyperparameter tuning and training only within the dataset, we do not expect predictions to be accurate beyond the dataset. Nevertheless, this is a helpful examination of the model's capacity to predict pK_{on} and pK_{off}, which was not previously tested for the Ligand-Transformer. Thus, a pre-trained version of the model, previously trained on the PDBbind for 72 epochs, was fine-tuned to our input dataset.

A hyperparameter search was done with a subset of the KIND dataset to adjust dropout and learning rate before fine-tuning. We then performed training with cross-validation only within the KIND. Figure 3.4 shows a subset of the metrics used to evaluate model performance in the validation sets. The model convergence can be seen for all three metrics after 35 epochs of fine-tuning. Performance was best for pK_{off} out of the three metrics, which is surprising considering the model was previously trained only to predict pK_D and there is a strong correlation between pK_D and pK_{on} but not pK_D and pK_{off} (Figure 3.1). After the final epoch, the model performance for the best model in the cross-validation

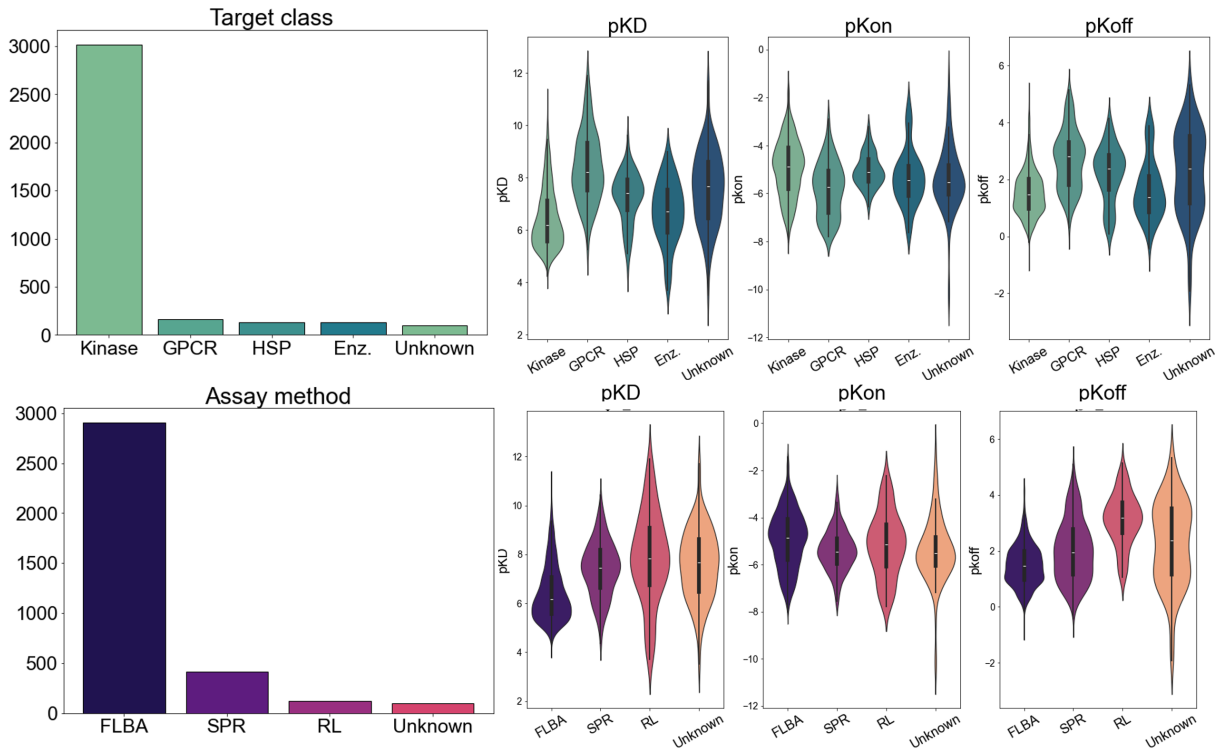

Figure 3.2: **Protein and Assay Type in KIND Dataset** GPCR - G-coupled protein receptors; HSP - heatshock protein; Enz. - enzyme; FLBA - Fluoresent-binding lignad assay; SPR - Surface Plasmon Resonance; RL - RadioLigand. On the left, the number of proteins from each class and experiment type. On the right, the distribution of kinetics for each protein class and experiment type. Based on this information, stratified splits were chosen.

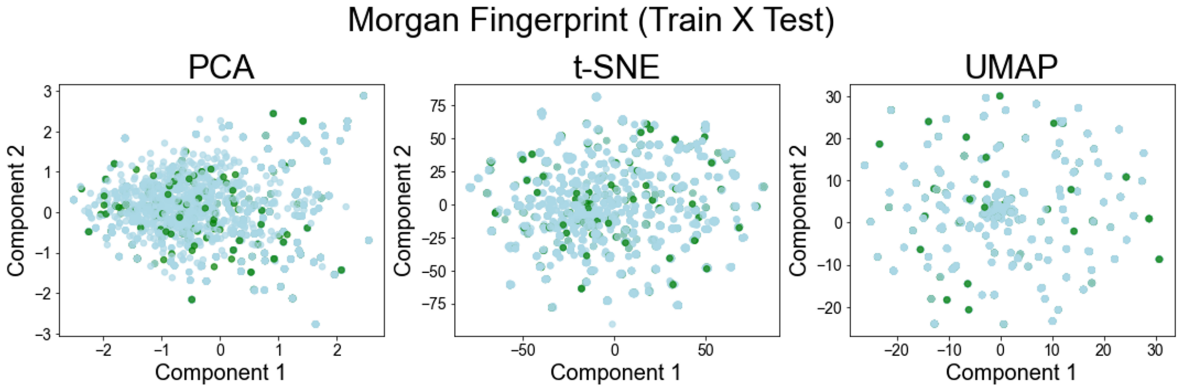

Figure 3.3: **Chemical space coverage of test set** Graphs illustrate the chemical space occupied by the test set (green) in relation to the train set (blue). Morgan fingerprints for each molecule were computed (radius = 2; bits = 124) and dimensionality reduction techniques used to visualise the space. The spread of data points in the test set indicate similar diversity in molecules compared to the training set.

is shown in Table 3.1. In this table, pK_{on} and pK_{off} were scaled back to their original distribution (see Methods).

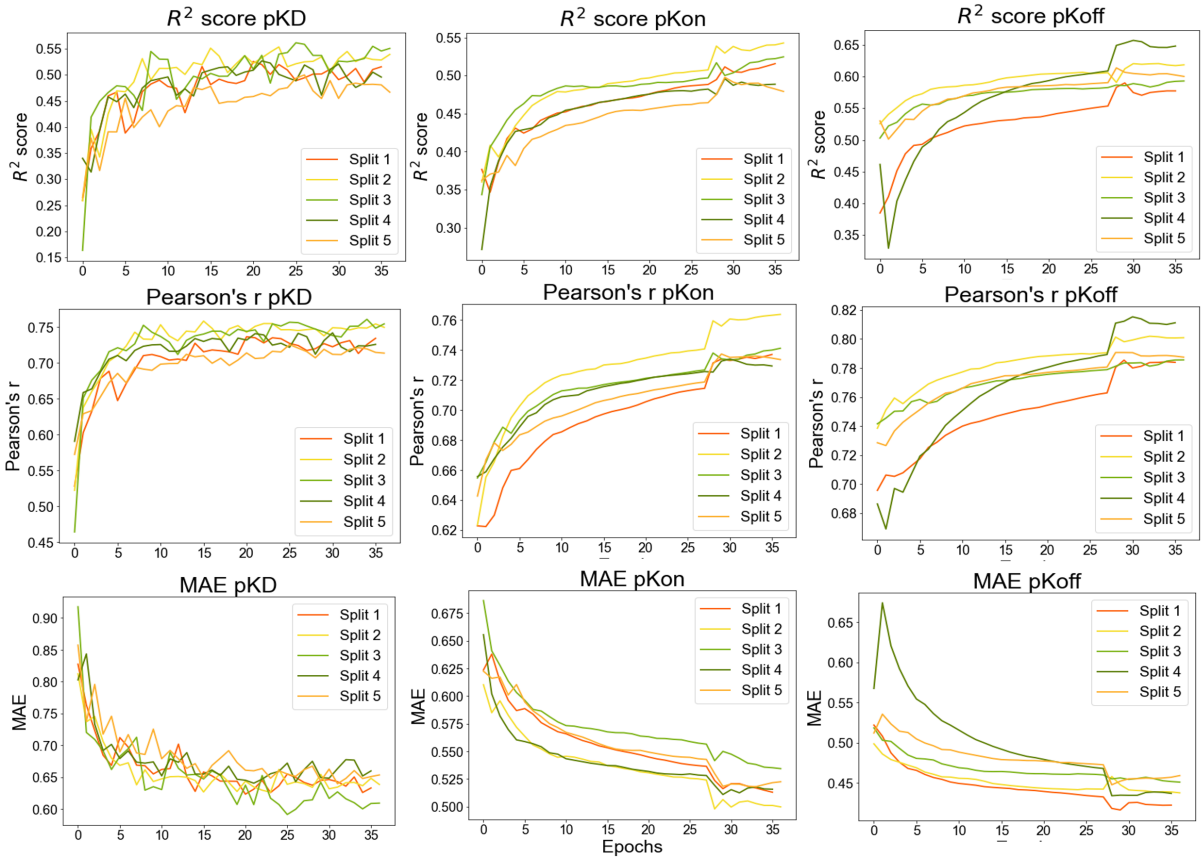

Figure 3.4: **Internal validation metrics.** The figures show R^2 scores (top), Pearson's r (middle) and MAE (bottom) for each split in the validation set.

Table 3.1: Metrics for internal validation model.

Metric	Pearson ↑	Spearman ↑	R^2 ↑	MAE ↓	RMSE ↓
pKD	0.736	0.7122	0.516	0.632	0.859
pKon	0.743	0.7306	0.519	0.614	0.836
pKoff	0.795	0.7842	0.609	0.426	0.654

As mentioned previously, predicted pK_D values can be calculated in one of two ways: either through the Affinity Head or the Kinetics Head of the model. Therefore, we used ground truth values of pK_D, and pK_{on} and pK_{off} to compare to either prediction method. The latter two metrics are related through Equation 2.1 to give "Calculated Kd" (Table 2.1). Figure 3.5 shows the average Pearson's r correlation coefficients across validation sets of the cross-validation. We find the performance of the Affinity Head and the Kinetic Head in predicting the true pK_D (Figures 3.5A and C) values is nearly identical (r = 0.73 and r = 0.72, respectively). Furthermore, the predicted calculated pK_D, from the model's Kinetic Head corresponds almost perfectly with the predicted pK_D, from the model's Affinity Head (Figure 3.5B, r = 0.99). Finally, there are slight errors between the true pK_D values and the true calculated pK_D values (Figure 3.5D, r = 0.98), likely

due to experimental measurement errors. These results indicate our model's Kinetic Head performs very similarly to the Affinity Head in the KIND dataset. Additionally, the performance of the Kinetic Head reaches experimental precision when predicting pK_D values. Hence, the model can predict binding affinity through the Kinetic Head using only validation data from within the original input dataset.

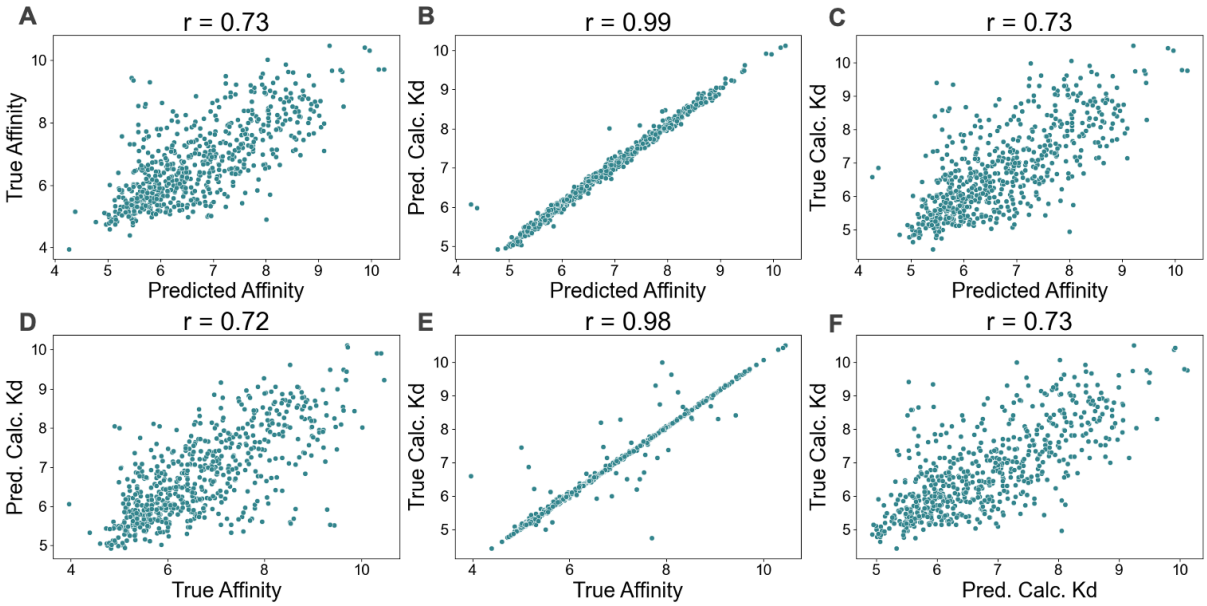

Figure 3.5: **Binding affinity prediction methods.** The plots show correlation between binding affinity values. These could either be true values, obtained from the data pK_D values (True Affinity) or calculated from data's pK_{off} and pK_{on} (True Calc. Kd). They can be predicted from the model in the Affinity Head (Predicted Affinity), or the Kinetics Head (Predicted Calc. Kd).

External validation

Our next goal was to accurately predict binding kinetics for unseen proteins and ligands in the training set. Once again, we used a previous version of the Ligand-Transformer, fine-tuning the model to this new task. We performed hyperparameter search and fine-tuning using data from the PDBbind, one of the largest databases with pK_D values for protein-ligand interactions, as well as the KIND dataset, one of the largest datasets with pK_D, pK_{on} and pK_{off} data on protein-ligand complexes.

The hyperparameter search was performed to adjust the learning rate, and the dropout rate for the Kinetics and Affinity Heads, using a Bayesian strategy with a subset of 30% of the total training data. Once the best model was selected, both datasets were used to train the model. In the case of the KIND, the Affinity Head and Kinetics Head were used in training and validation. For the PDBbind, only the Affinity Head was used in the training, but during validation the Kinetics Head was also used to predict pK_{on} and pK_{off} for these data points. Equation 2.1 then allowed us to measure the Kinetics Head performance by comparing the calculated affinity to the ground-truth affinity. This way, we were able to measure the model's performance on a wide range of protein-ligand complexes, going beyond the Kinetics Head training set.

Figure 3.6 shows the evolution of the model performance during the 15 epochs of fine-tuning. We can see the PDBbind metrics deteriorate compared to the original model. We also see the model performs similarly to the splits in the internal validation for the Kinetics Head in pK_{off} ($r_{ext} = 0.81$ and $r_{int} = 0.79$) and pK_{on} ($r_{ext} = 0.74$ and $r_{int} = 0.74$) and on the Affinity Head ($r_{ext} = 0.74$ and $r_{int} = 0.74$). The Kinetics Head converges similarly for KIND and PDBbind to predict pK_D (Figure A.2). The full table metrics (Table 3.2) are given for the data in the final epoch. The table refers to "$C.pK_D$" as the "Calculated pK_D". Again, we notice a higher accuracy for pK_{off} (r = 0.814) compared to the other metrics, but similar performance otherwise (r = 0.746 for pK_D, r = 0.737 for $C.pK_D$ and r = 0. 748 for pK_{on}). The model performs slightly better with data from PDBbind (r = 0.789 for pK_D and r = 0.752 for $C.pK_D$) compared to KIND.

Table 3.2: Metrics for external validation model.

Dataset	Metric	Pearson ↑	Spearman ↑	R^2 ↑	MAE ↓	RMSE ↓
KIND	pKD	0.746	0.715	0.492	0.664	0.891
	C. pKD	0.737	0.712	0.507	0.652	0.878
	pKon	0.748	0.738	0.511	0.628	0.844
	pKoff	0.814	0.801	0.658	0.381	0.536
PDBbind	pKD	0.789	0.782	0.511	0.97	1.296
	C. pKD	0.752	0.762	0.427	1.08	1.401

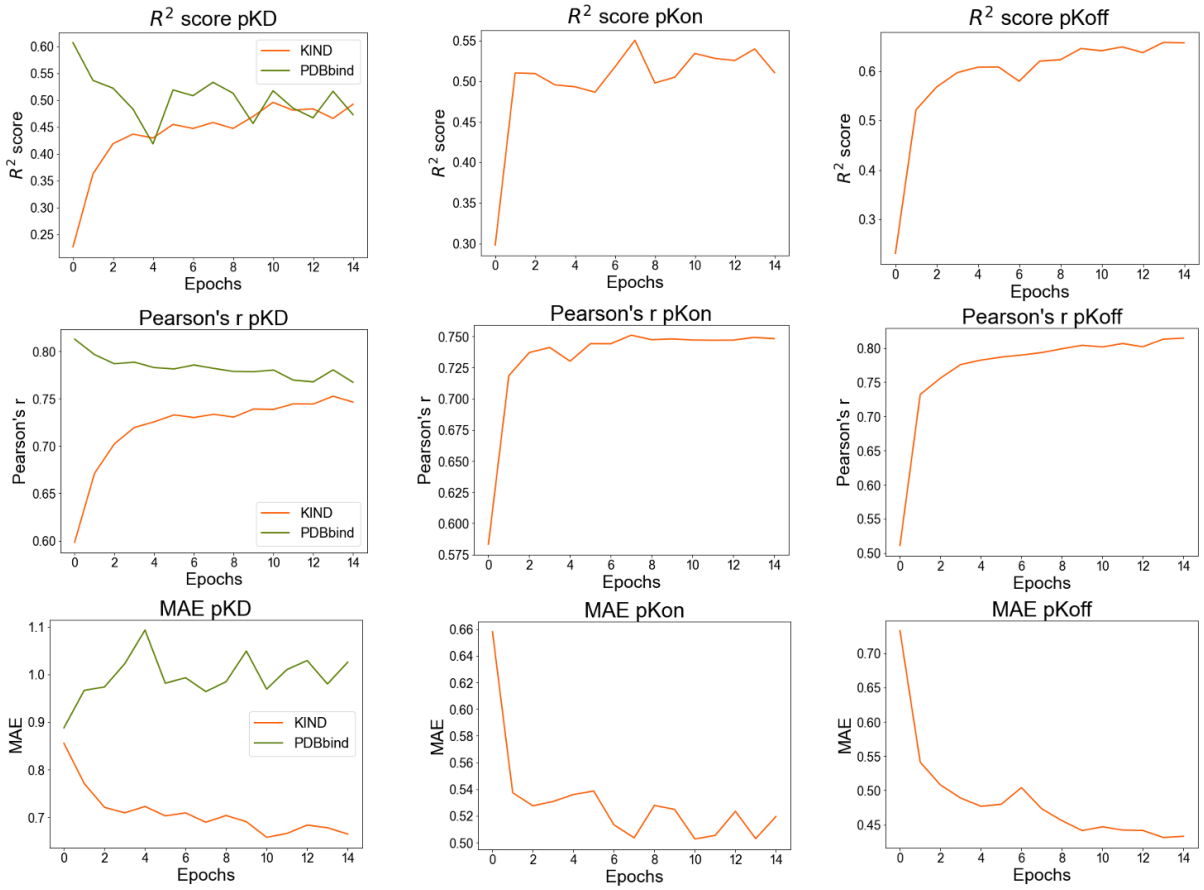

Figure 3.6: **External Validation Model Metrics.** The figures show R^2 scores (top), Pearson's r (middle) and MAE (bottom) for the validation set of the extended model. Where validation information for the PDBbind was available (green), this was plotted alongside the KIND information (orange).

Similar to the internal validation, we again compare the methods for calculating binding affinity. As mentioned previously, affinity values can be the ground truth values from the data or predicted values from the model. The predicted values can either stem from the Affinity Head, which directly uses the ground truth affinity value in training, or the value can stem from the Kinetics Head, which uses ground truth values for only a portion of the dataset, where this information is available. Figure 3.7 shows very similar Pearson's correlation scores in KIND when predicting affinity using the Affinity Head (r = 0.75) or the Kinetics Head (r = 0.74; Figures 3.7 A-B). The predictions are also similar for PDBbind data points (r = 0.79 for Affinity Head and r = 0.75 for Kinetics Head; Figures 3.7 D-E). Similar to the internal validation, we find the predicted affinities from the Affinity Head correspond very closely to those from the Kinetics Head (r = 0.98 for KIND and r = 0.95 for PDBbind; Figures 3.7 C and F). These results indicate the model is able to generalise well and predict kinetics parameters for unseen data.

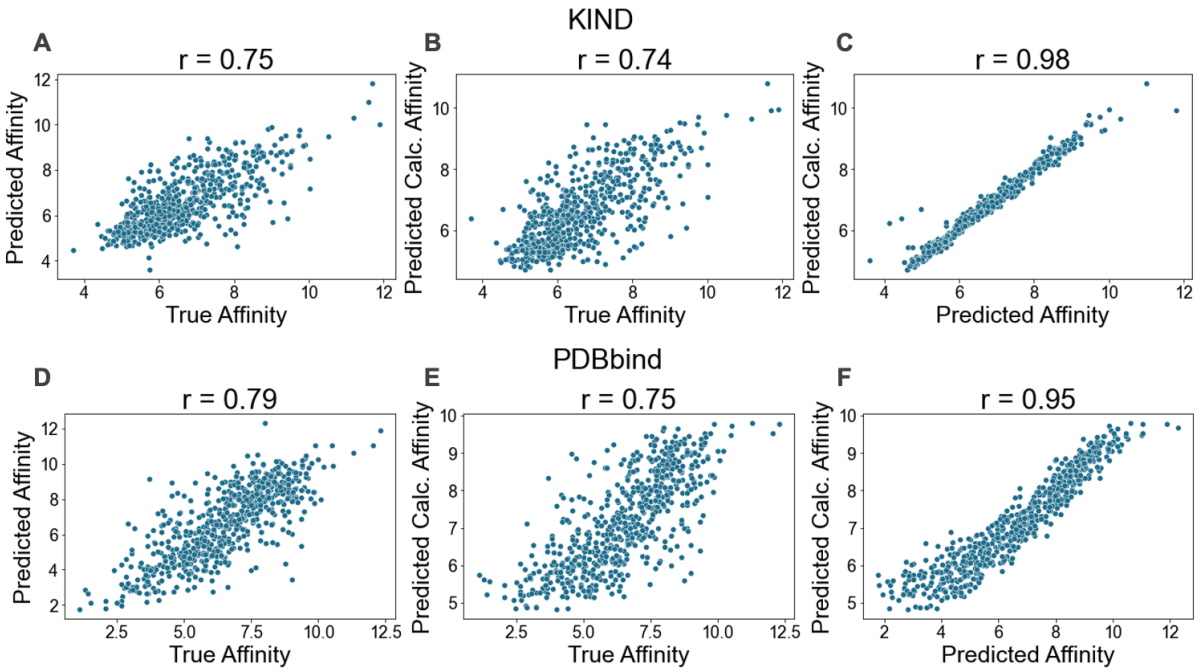

Figure 3.7: **Binding Affinity Prediction External Validation.** The scatter plots show the Pearson's r correlation coefficient for the binding affinity values. These could either be true values, obtained from the data pK_D values or calculated from data's pK_{off} and pK_{on} or they can be predicted from the model in the Affinity Head or the Kinetics Head.

Case study: Drug Design for Medin

Following investigations of the prediction of binding kinetics, we aim to apply the Ligand-Transformer to Medin. We define characteristics of the input Medin structure into the model, select molecules with high binding affinity to Medin and discuss the predicted binding sites and specificity of the selected molecules. Finally, we also look into the probable underlying molecular characteristics that drive binding to this protein.

Medin Structure

One of the outputs of the outputs of the Structure module of AlphaFold2 is the structural representation of the protein. This is composed of inter-residue distance matrices of the protein. We provide a matrix of the mean and standard deviation for each inter-residue point (Figure 3.8a and b, respectively), which are representations of the structure input used in the Ligand-Transformer. For illustration purposes, we also input the Medin sequence into AlphaFold2 and coloured by confidence score (Figure 3.8c). It should be noted this is just one representation of the protein, rather than the representation used in the model.

Even in this initial average representation of the protein, we can faintly see areas of non-adjacent residues with a predicted shorter inter-residue distance. This possibly

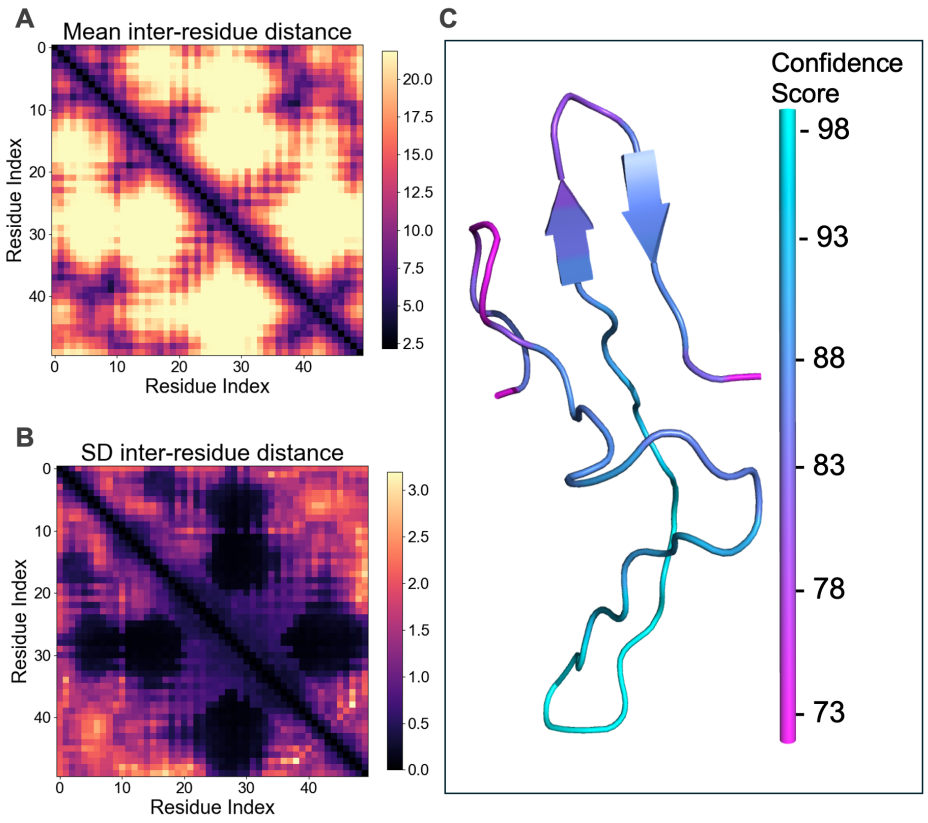

Figure 3.8: **Distance matrix for AlphaFold2 Medin prediction.** The figure shows the average (A) and standard deviation (B) predicted distance matrices for Medin from Colabfold. We also show the best model for Medin structure output from Colabfold (C).

indicates the formation of β-sheet hairpins, which have been previously described in the protein [46]. Interestingly, despite the reported intrinsic disorder in Medin, the confidence scores provided by AlphaFold2 are relatively high, with a minumum of 73%.

Based on this, we chose to investigate Medin's disorder. IUPred2A is a disorder predictor that takes into account a protein's energy [53], which can be derived by its amino-acid sequence. Because disordered proteins have amino-acid biases, this energy calculation is difficult to perform. Thus, the algorithm is able to calculate the protein segments' likelihood to be disordered. Figure 3.9 reveals Medin seems to be mostly ordered with the exception of its termini, which show over 80% probability of being disordered. No residues show high probability of being disordered.

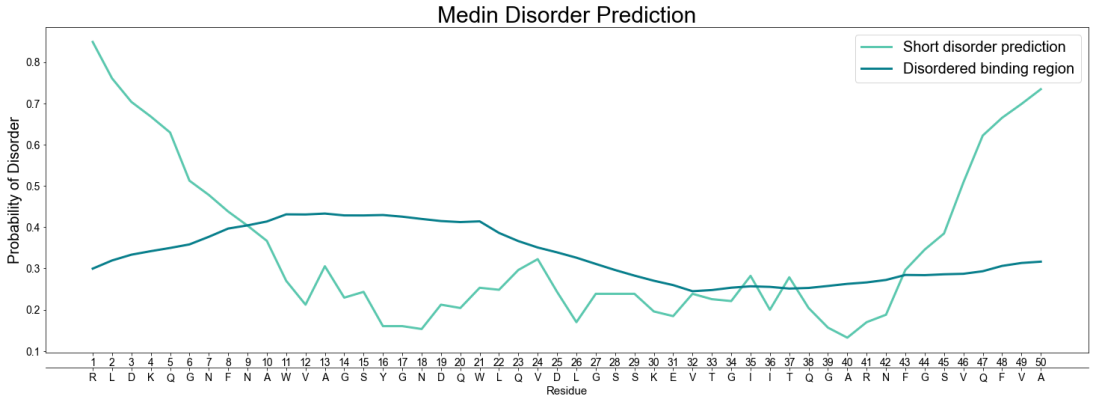

Figure 3.9: **Medin disorder prediction.** Graph showing the likelihood of short segments of the protein to be disordered. Prediction of short disorder segments reveal only the 5 residues of each termini show probability of disorder above 50%. Graph also shows that no residues are likely to be disordered binding regions.

Molecular library screening

Because the Ligand-Transformer for Kinetics is still in development, we used the latter to screen large molecular libraries. We performed a forward pass of the model to predict the the binding affinity of the molecules to Medin. Inferences were obtained for two versions of the model, Model 1 and Model 2 (see Methods). We can notice that for Model 2, the affinity predictions had a narrower distribution than for Model 1 (Figure 3.10).

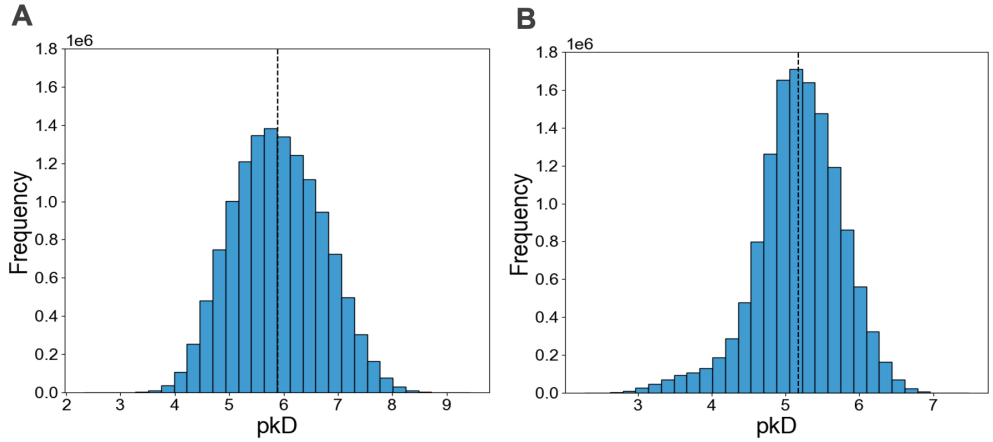

Figure 3.10: **Binding affinity distribution for ligand complexes from Medin and ZINC library.** The figure shows the values for binding affinity for Model 1 (A), without hyperparameter search, and Model 2 (B), with hyperparameter search. The dotted lines represent the mean affinity in each model.

To maximise the chemical space covered by the high-affinity molecules, we used Butina clustering, a method that computes molecules' Morgan fingerprints and their associated pairwise Tanimoto similarity index to form molecular clusters. We used the 50,000 molecules with highest binding affinity to Medin to perform clustering, as this was the

maximum number of molecules available to cluster on Chemfp (see Methods). This procedure was done for the outputs from Model 1 and Model 2 (Figure 3.11). The Butina clustering algorithm output 50 clusters for each model and those clusters were considered as representative samples of the chemical space.

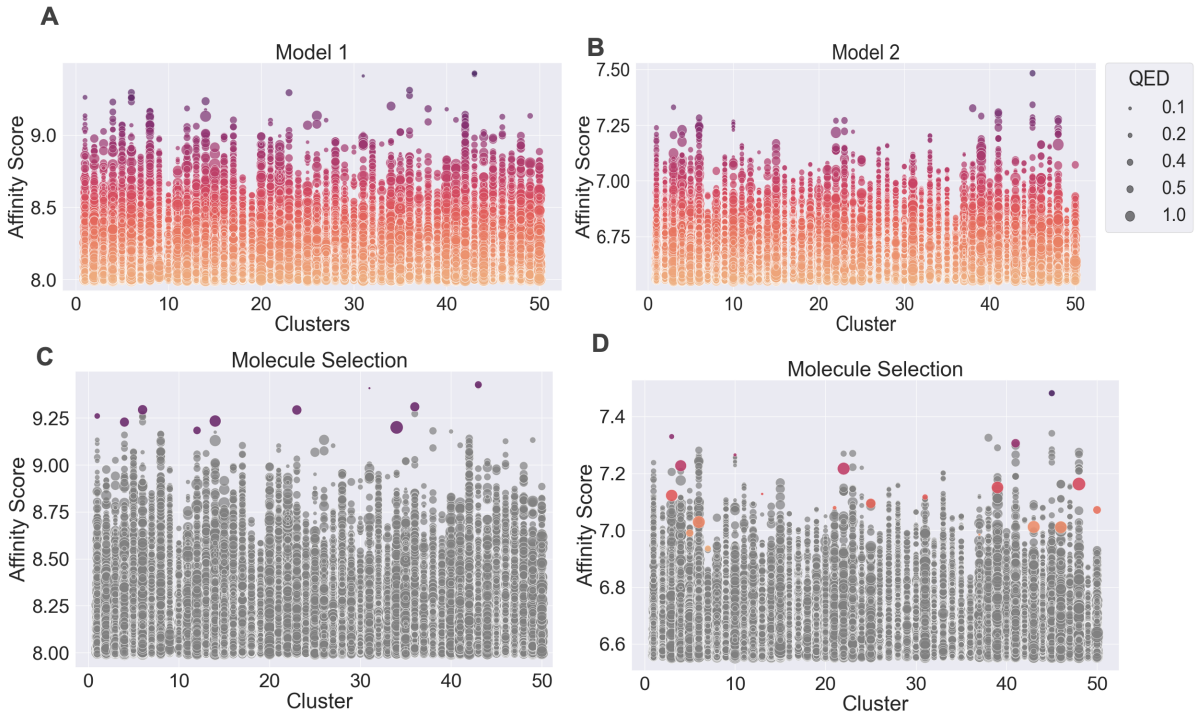

Figure 3.11: **Butina Clustering for Model Selection.** The figure shows the Butina clustering output for the top 50,000 from Model 1 (A,C) and Model 2 (B,D). The plots show the clusters on the x-axis and the affinity scores on the y-axis. The size of the molecules corresponds to their quantitative estimate of drug-likedness (QED) score. The colours of the molecules correspond to their affinity (A-B) and to their selection status (C-D), with gray corresponding to non-selected molecules. Selection was done for the molecules with the best affinity score in each cluster, with the 10 top molecules in Model 1 and Model 2. The remaining 10 molecules were chosen according to those in the top 10,000 molecules for both Models 1 and 2.

Our goal was to select 30 molecules to be experimentally validated. To select these molecules, we combined results from Model 1 and Model 2. In both cases the molecules with the highest affinity in each cluster were chosen and from that set, the 10 molecules with the highest affinity were used in the final selection. The remaining 10 molecules were selected as they were in the top 10,000 molecules for both models. Figure 3.12A and Table A.1 provide information about the final selection. For the selected molecules, we show the quantitative estimation of drug-likedness (QED), despite this metric not being considered for selection. We also confirmed the selected molecules were sufficiently different from each other using Tanimoto similarity matrix (Figure 3.12B). Although molecules M3 and M15 show higher similarity score, they were kept as they had very different affinity predictions.

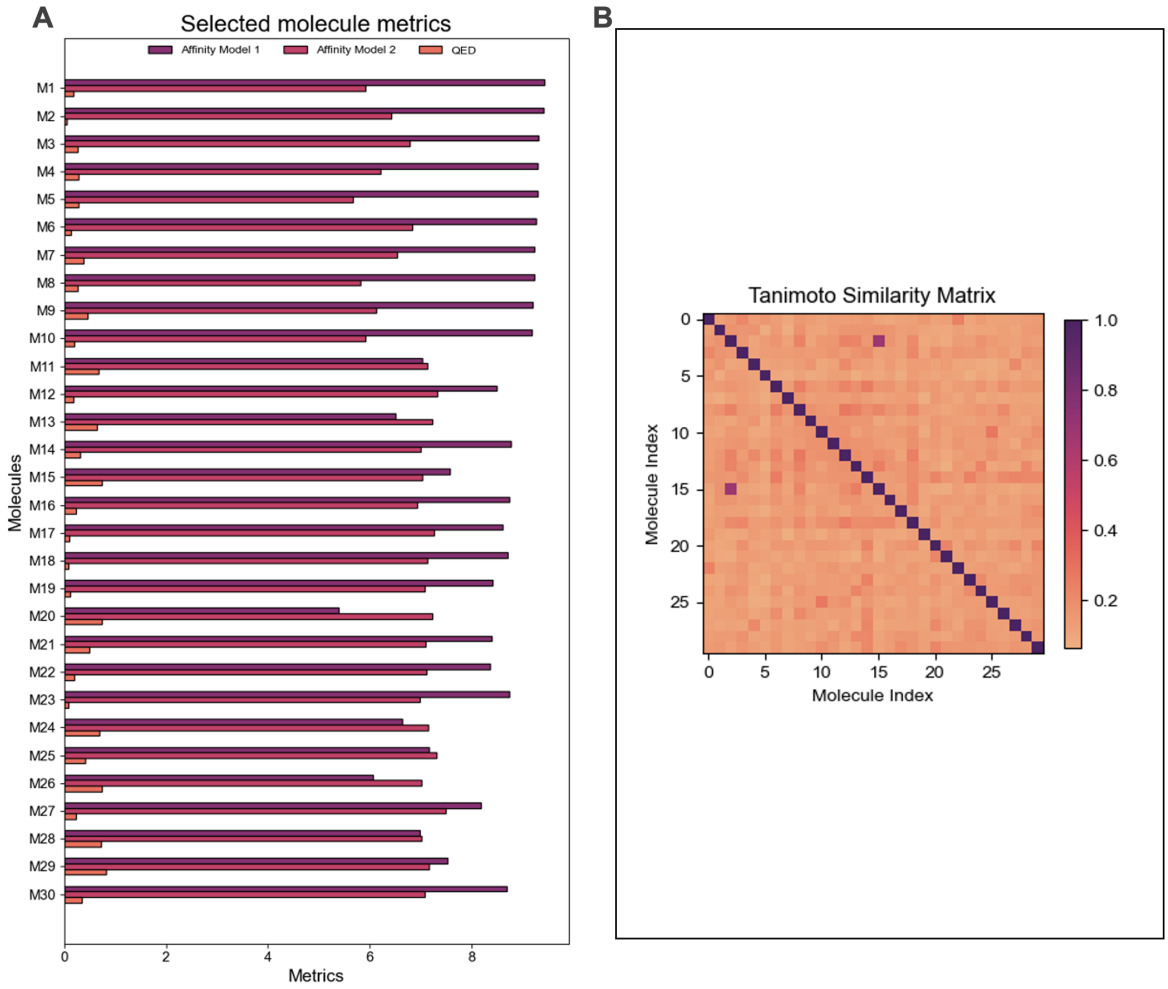

Figure 3.12: **Final molecule selection.** (A) The figure shows the predicted affinity measures according to Model 1 and Model 2. M1-10 were selected according to predictions for Model 1, explaining the high affinity predictions in Model 1 but not Model 2. M11-M20 were selected as being in the top 10,000 ranked molecules in both models. Finally, M21-30 were selected according to Model 2, showing lower affinities for Model 1. We also show the QED for each molecule. (B) Matrix of Tanimoto similarity index for molecules selected for experimental validation. Molecules 3 and 15 were kept despite their high similarity index, as their predicted affinities were sufficiently different in both models.

Finally, we obtained predicted pK_{on} and pK_{off} for the selected molecules and compared results to those obtained in Model 2 of the original Ligand-Transformer. As expected based on the difference in performance of the Affinity Head on the PDBbind data (Figure 3.6) after fine-tuning, there are some differences in the prediction of binding affinity in the Affinity Head of the Ligand-Transformer and the Ligand-Transformer Kinetics (r = 0.59; Figure 3.13A). The predicted binding affinity values from the Kinetics Head also seem generally higher than those from the Affinity Head of the Ligand-Transfomer (Figure 3.13B) and of the Ligand-Transformer Kinetics (Figure 3.13C), although the latter were highly correlated (r = 0.87).

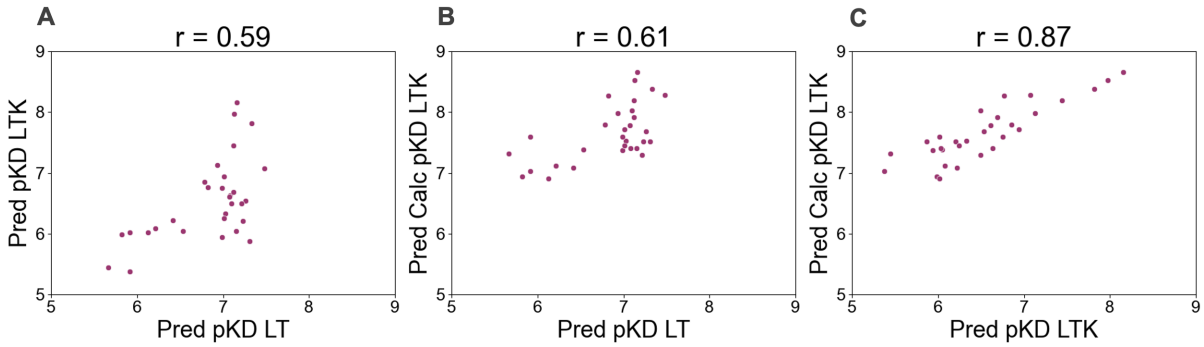

Figure 3.13: **Binding Kinetics Prediction for Selected Molecules.** Pred pKD LT: predicted binding affinity with Affinity Head in Ligand-Transformer; Pred pKD LTK: predicted binding affinity with Affinity Head in Ligand-Transformer Kinetics; Pred Calc pKD LTK: predicted binding affinity with Kinetics Head in Ligand-Transformer Kinetics. We show the correlations between the different affinity measures in the Ligand-Transfomer Model 2 and the Ligand-Transformer Kinetics.

Predicted binding sites

The Distance Head of the Ligand-Transformer provides information about the predicted binding sites of a protein-ligand complex. We wished to identify the binding sites for the final selection of 30 molecules. For each complex, the predicted distance from each residue to the ligand and probability of binding to ligand per residue were found. The distance from each residue in the protein to the ligand (Figure 3.14A) shows at least one contact point in Trp^{11}, where the distance is < 5 Å. The probability of binding per residue (Figure 3.14B) is mostly consistent across molecules with some binding sites, corresponding to the lowest distances to residue, having high probability of binding for all molecules.

To visualise the binding sites on the Medin structure, the authors of Davies et al. (2017) [46] kindly shared their predicted structure for Medin from MD simulations. We chose to visualise binding sites with average binding probability above 80% to ensure all molecules had high probability of binding to these sites. The highlighted residues are Trp^{11}, Leu^{22}, Ile^{35} and Thr^{37}. The aforementioned residues are shown in the protein structure (Figure 3.15A) and their side chains are highlighted to precisely show the likely binding pocket (Figure 3.15C). The distance matrix for the MD structure (Figure 3.15B) once again shows the non-adjacent residues at a close distance, confirming the β-hairpin structure observed in the AlphaFold output (Figure 3.8A).

We can see the binding sites lie close in distance to each other, indicating the presence of a binding pocket in Medin. The binding pocket forms across the β-sheet hairpin. Interestingly, when observing the structure of the selected molecules (Figure A.3), most molecules appear to be elongated. Thus, it is likely that long molecules are selected as having high binding affinity due to their elongated shape, in order to bind across Medin's β-sheet hairpin.

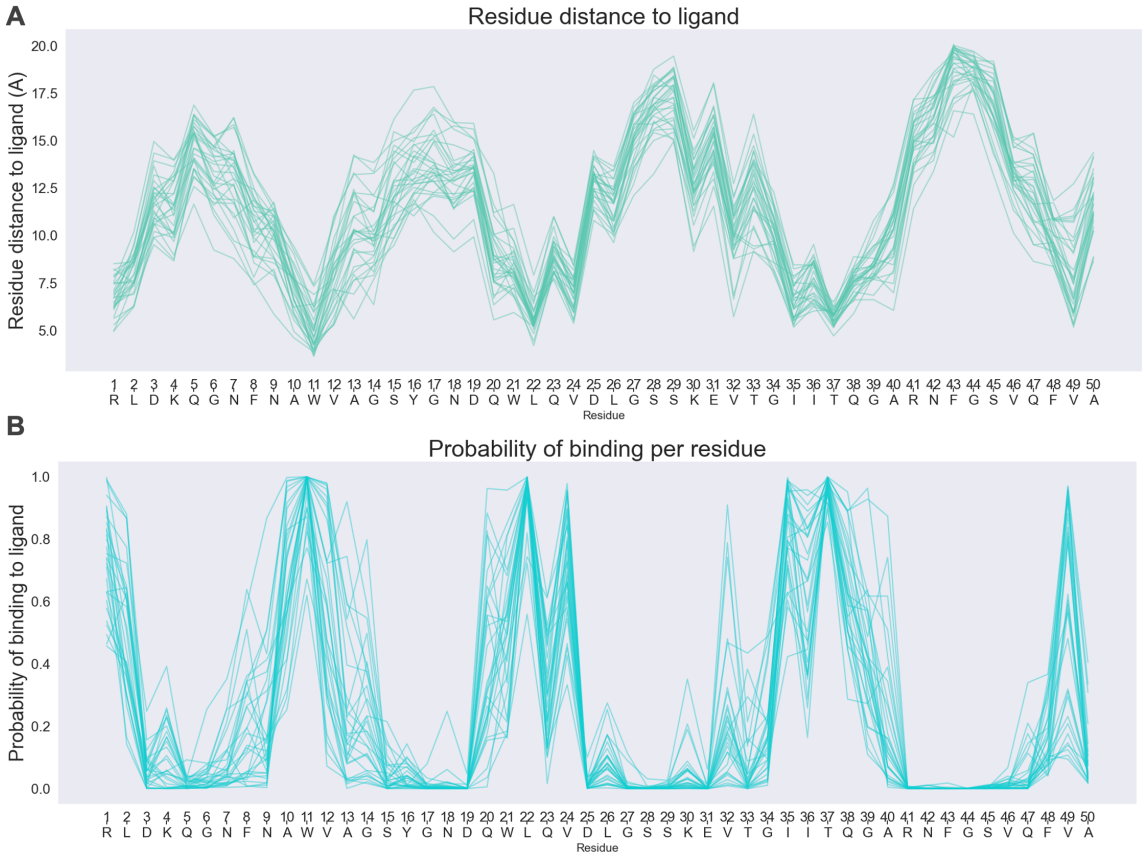

Figure 3.14: **Binding Site Prediction for Selected Molecules.** The figure shows the predicted distance from each residue to each ligand (A) and probability of binding to the ligand (B), each ligand portrayed as a line in the graph.

Medin Specificity

The binding affinity to Medin only elucidates part of its mechanism of action as it disregards the competitive binding between Medin and other proteins in the human proteome. The β hairpin structure found in Medin is not an uncommon motif in the human proteome and at least 50,000 unique β hairpin motifs are documented in the PDB [54]. Therefore, we aimed to find how specific binding to Medin is for our selected molecules. These were then screened against the human proteome using Model 2. Figure 3.16 shows the distribution of binding affinities of each molecule to the human proteome, with affinity to Medin highlighted as the dashed orange line and Lactadherin, Medin's precursor protein, in pink. Each plot shows a p-value indicating the probability that a given molecule will have a higher binding affinity to another protein than to Medin. For majority of the molecules (73%), Medin had a stronger binding affinity than Lactadherin, indicating the binding pocket is likely shielded by the Lactadherin folding. We can see molecules M1-M10 show much higher p-values, likely due to the fact they were chosen only based on predictions from Model 1. In contrast, the molecules M11-M30 show higher binding affinity to Medin

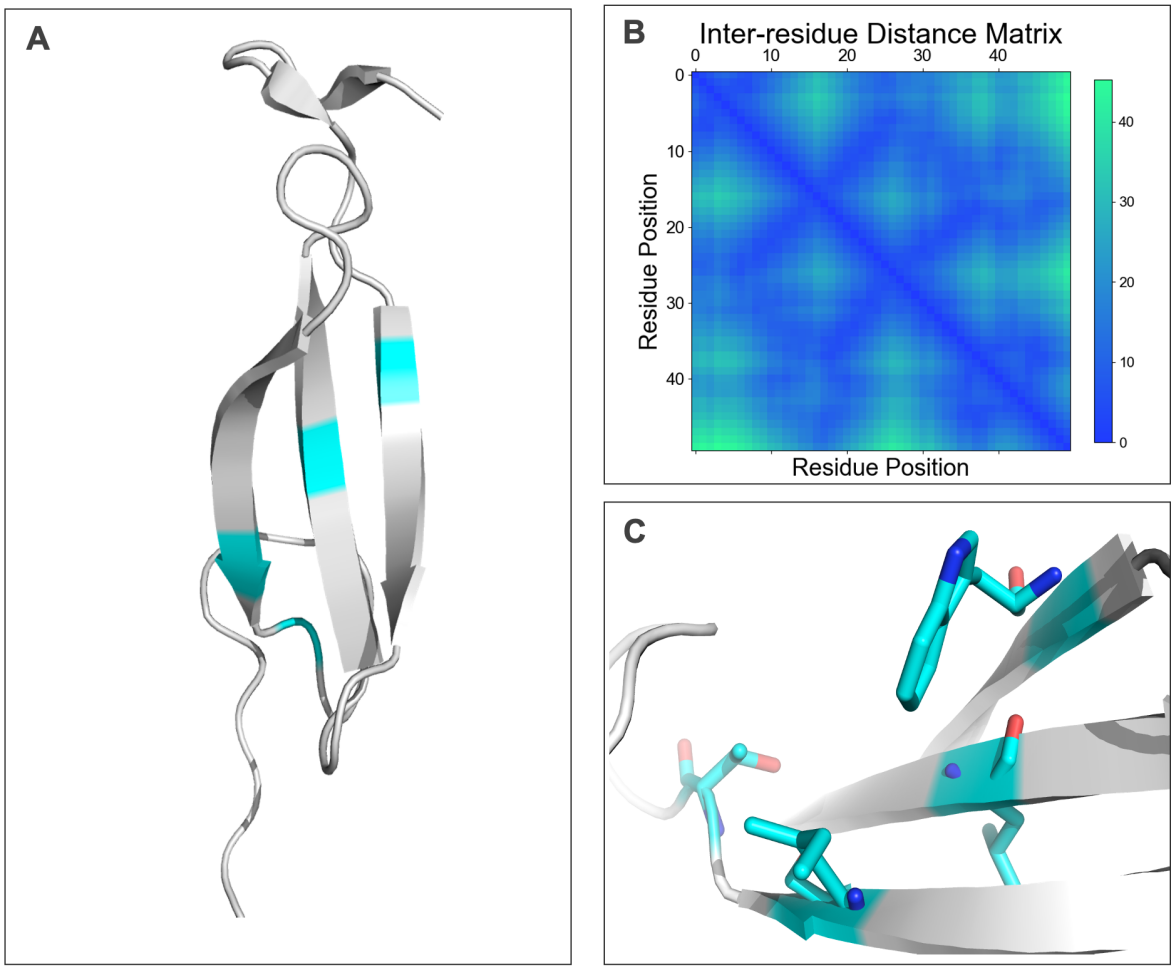

Figure 3.15: **Inter-residue distance MD simulation Madine et al.** (A) The figure shows inter-residue distance for the predicted Medin structure from Davies et al. (2017) [46]. Highlighted in cyan are the residues with highest probability of binding affinity: Trp^{11}, Leu^{22}, Ile^{35} and Thr^{37}. (B) The inter-residue distance matrix is derived from the protein structure in A. (C) The side chains for the highlighted residues are shown.

relative to the whole proteome. For 2 molecules, M13 and M24, the probability of binding to other proteins is lower than 10% for 7 molecules it is lower than 20%. Nonetheless, it is not surprising these results do not show overall high specificity to the target, Medin, as molecules were not selected using this criterion.

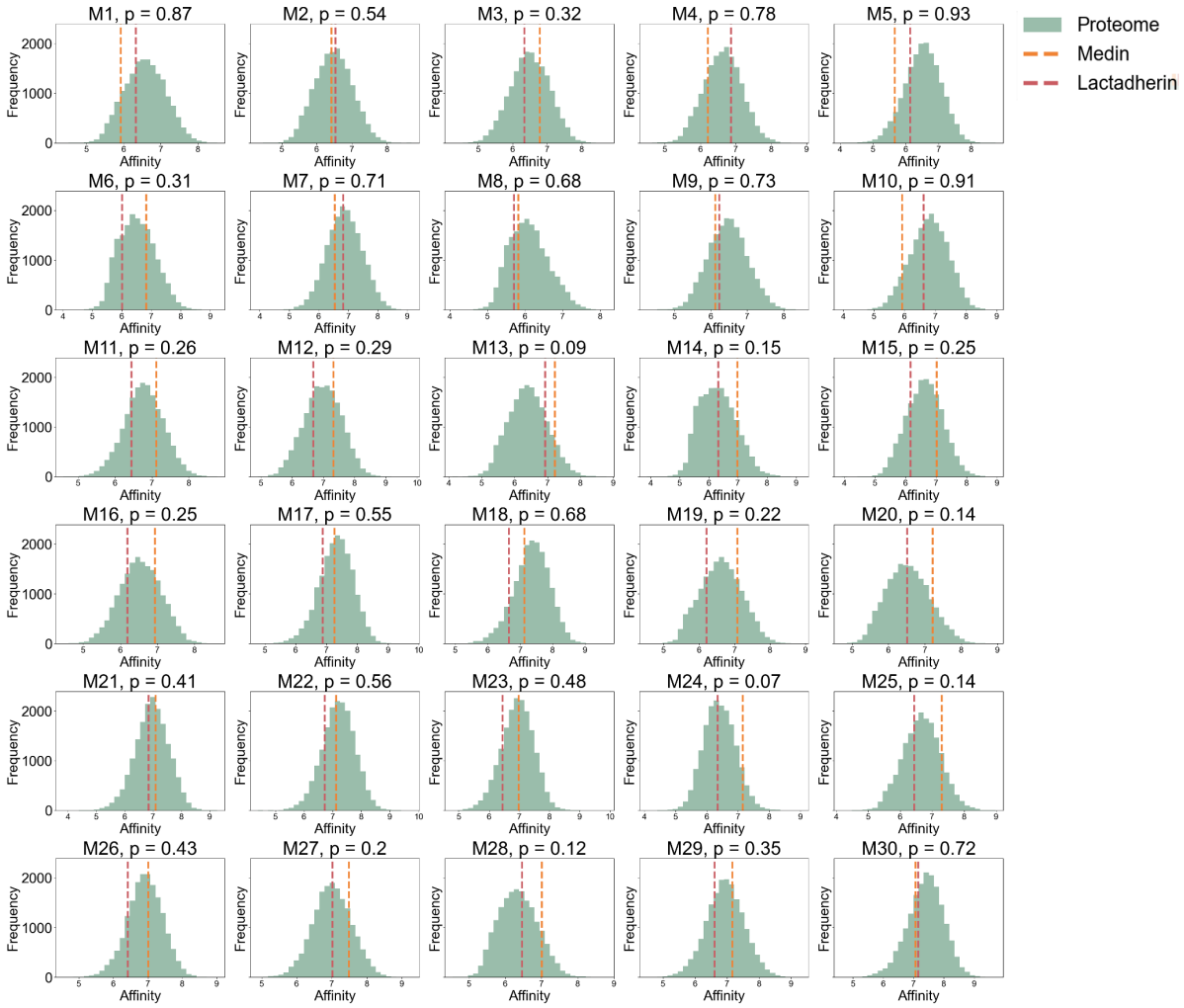

Figure 3.16: **Binding specificity for selected drugs to Medin.** Histograms show the distribution of binding affinities of molecule to the human proteome. The binding affinity to Medin and Lactadherin are shown as orange and pink dashed lines, respectively. p-values are given for each molecule as the probability of having a protein with higher binding affinity to the molecule than Medin.

Molecular Descriptors Driving Medin Binding

Our next goal was to recover some underlying information from this black-box model. We examine whether any molecular descriptors could explain the predicted binding affinity to Medin. The PaDEL molecular descriptors software was used to calculate over 1900 molecular descriptors for a subset of the molecules screened (see Methods). A random sample of 100,000 molecules from the ranked list of affinities was obtained with systematic sampling. Descriptors for this subset were then used as predictors in a Random Forest Regressor. Sampling was also done for the 50,000 best and worst binders to Medin and descriptors were used as input to a Random Forest Classifier to classify molecules into "low" or "high" affinity.

We found the regression model was able to predict binding affinity to Medin with an $R^2 = 0.59$, Spearman's $\rho = 0.71$, Pearson's $r = 0.77$, $MAE = 0.28$, $RMSE = 0.36$. The most important regression features were $SpDiam_D$ and Eccentric connectivity index. The classification model was able to predict affinity class ("low" or "high") with a ROC-AUC = 0.99 and accuracy = 0.99. The most important classification features were Eccentric connectivity index and SpMax8. It should be noted that although the predictors found in the regression and classification tasks were different, they are all highly correlated. All the features relate to the topological distance matrix of the molecule. For instance, the Eccentric Connectivity Index [55] is given by

$$\xi = \sum_{i=1}^{n} E(i)V(i) \tag{3.1}$$

where E(i) is the eccentricity, the maximum topological distance between a pair atoms i and j in a molecule, and V(i) is the degree of atom i. In other words, the binding to Medin is associated with the parameters of the topological matrix of a molecule.

We used the top 5 predictors in each model to perform dimensionality reduction. Figure 3.17 shows PCA, t-SNE and UMAP for both the regression and classification models. In both cases, the non-linear techniques (t-SNE and UMAP) show better clustering of molecules according to their affinity.

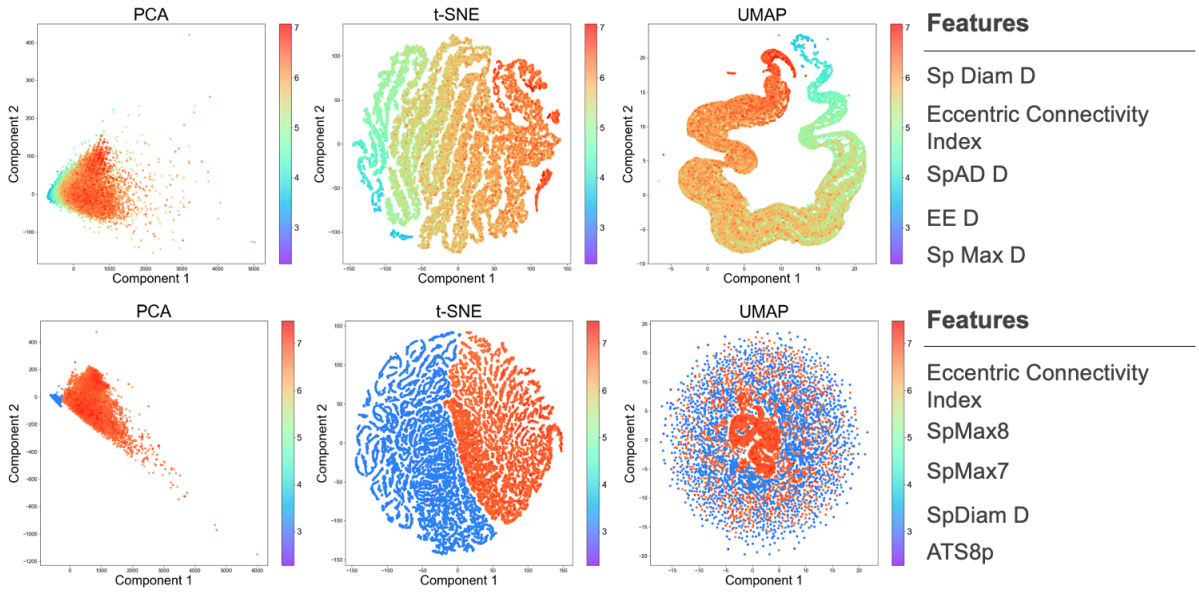

Figure 3.17: **Clustering of Random Forest's Top Predictors for Medin binding affinity.** (top) Molecule sampling was done systematically to sample 100,000 molecules from ranked list of affinities to Medin. Clustering was done based on the top 5 predictors in Random Forest Regressor. (bottom) Molecule samples were taken from the 50,000 best and worst binding affinities to Medin. Clustering was done based on the top 5 predictors from Random Forest Classifier.

Doing further investigation, we found a linear relationship between binding affinity to Medin and the log_{10} of these molecular descriptors (Figure 3.18), explaining why the non-linear dimensionality reduction techniques performed better. This is once again consistent with the finding of the binding pocket formed in Medin, which requires long molecules to bind across the β sheets. In other words, longer molecules, with more inter-connected atoms, are likely to fit well in the Medin binding pocket.

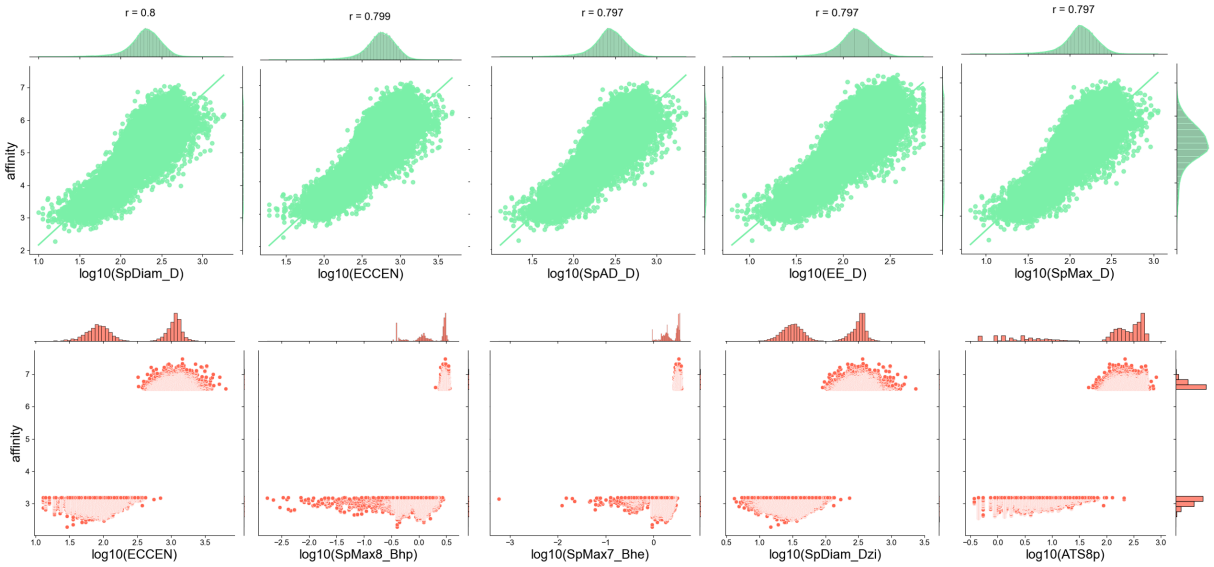

Figure 3.18: **Correlation between binding affinity and** $log10$ **of topological descriptors.** Scatter plots demonstrating the relationship between affinity (pK_D) and the $log10$) of top 5 features for regression (top) and classification (bottom) found with Random Forest models.

Overall, we have found a list of candidate compounds with high binding affinity to Medin to be evaluated *in vitro* in aggregation assays. We went further to demonstrate it is likely these compounds will bind across the β hairpin structure, but is likely to bind to other proteins in the proteome *in vivo*, since few molecules show specific binding to Medin. Finally, we found that longer, more inter-connected molecules are more likely to bind to Medin.

Chapter 4

Discussion

Our goals were to extend the Ligand-Transformer to predict binding (pK_{on}) and unbinding (pK_{off}) rates for protein-ligand complexes, and to use the model to characterise binding to an Alzheimer's-linked, amyloid-aggregating protein, Medin. To predict binding kinetics, we added a Kinetics Head to the original Ligand-Transformer architecture, following a similar approach to the Affinity Head. To characterise Medin binding, we used the model to select 30 molecules to be experimentally validated in aggregation assays with Medin. We also characterised the binding sites, specificity and common molecular features found in these compounds.

Our model's prediction of binding kinetics is comparable in accuracy to the best performing models in the literature [1, 39, 40, 56]. We also go beyond these studies, which used a "warm-setting" to evaluate the model, having some proteins or ligands in the training set also present in the validation set (see Methods), to show that proteins and ligands unseen in the training of the Kinetics Head, had relatively high accuracy (Pearson's r = 0.75).

Predictions from our model are also as accurate as other experimental techniques in determining the value of binding kinetic parameters. Georgi et al. (2020) [57] show a comparison of binding kinetics parameters measured using a range of experimental methods. They show k_D experimental measures have a correlation of r = 0.74; k_{on} measure show r = 0.62 and k_{off} shows a correlation of r = 0.75. Thus, given the "ground truth" values of experimental methods have little reliability, it is impossible to achieve a perfect prediction of these parameters. In light of this, when comparing the correlations achieved here of r = 0.79 for pK_D (Figure 3.7D) and r = 0.74 for the unseen ratio of pK_{off}/pK_{on} (Figure 3.7E), we can conclude our model performs as well as experimental procedures in determining the binding kinetics of protein-ligand complexes.

Our results should, however, be considered in light of their limitations. Due to the lack of available experimental data, it is difficult to generalise the model's predictions beyond the training set. Additionally, we currently do not consider the error in the experimental

measurements from which the data is derived from. In some cases, the errors reach 20-50% of the K_D, K_{on} and K_{off} values [58]. By considering experimental measurements as ground truth values, we disregard the inherit uncertainty in these experiments and cannot accurately evaluate the model's performance. In the next steps of our work, we plan to consider these experimental errors in the evaluation of our own model. Therefore, we propose adding another error function to measure the performance of the model. By transforming the error using the function

$$f(x) = \begin{cases} 0 & \text{for } -0.5 \leq x \leq 0.5 \\ x^2 & \text{otherwise} \end{cases} \tag{4.1}$$

we allow predictions within 0.5 deviation to be considered an accurate prediction and add a quadratic error if the prediction is outside this range. Thus, in the future, we aim to add more experimental data to train the model and employ this error function to improve its evaluation.

When considering the Medin case study, the main indicator of the pipeline performance will be experimental validation. Nonetheless, we present interesting theoretical results. For instance, contrary to previous literature [45, 44, 59], we show Medin is largely an ordered protein. The distance maps from both AlphaFold2 and MD simulations reveal the presence of β-sheet hairpins, which are ordered motifs. The prediction of disorder revealed the majority of the sequence has a low probability of disorder, except for the termini, which show a probability of disorder of up to 80% (Figure 3.9). The protein's termini also correspond to the regions with the lowest confidence from AlphaFold2 (Figure 3.8C). The analysis shows all residues also have a low probability of being disordered binding regions. These results will be taken into account for the experimental assays.

Considering the predicted binding sites with the highest probability of binding are located within residues 11-37 of the protein, we can conclude Medin would undergo ordered binding. Therefore, the main goal of binding is to stabilise protein aggregation rather than the protein's conformational fluctuations. The predicted binding sites are located at a crucial point in the protein. It is known that when Medin aggregates, they tend to form a cross-β structure, where the β hairpins from different proteins are parallel to each other [45]. Additionally, Medin-Aβ aggregates in Alzheimer's disease are known to be β-sheet rich [59]. Our predicted binding site crucially lies across the protein's β-hairpin, indicating molecules will likely interfere with the aggregation process. Whether this interference promotes or delays aggregation can only be determined experimentally. Furthermore, 2d Nuclear Magnetic Resonance studies have shown cross-peaks between Trp^{11} and Ile^{35} in Medin [44], indicating these residues are coupled through magnetization transfer and thus, probably lie in close distance in the structure. These residues coincide with two of our most likely binding sites, providing some experimental confirmation of the predicted

structures and likelihood of forming a binding pocket.

The β hairpin is a fairly common motif in the human proteome, with over 50,000 structures documented in the PDB [54]. The fact that most of our selected molecules seem to bind to this region of the protein raised the question of how specific this binding would be. The specificity analysis revealed that among the selected compounds some, but few, bind specifically to Medin, with this protein being in the top 10% of strongest binders. This is a good indication not only for *in vitro* assays but also *in vivo* assays, where competitive binding will play a role in determining how well these compounds bind to Medin. At present, specificity analyses are computationally expensive and therefore impossible to conduct as part of the drug discovery pipeline. A proxy measure for specificity is needed to include this measure in the drug discovery pipeline while limiting computational expense. Going forward, once there is experimental validation of basic functioning of the model, we aim to incorporate a compound's specificity score into the selection pipeline, rather than verifying its specificity after selection.

We also found that molecular features related to the topological distance matrix of a molecule, such as the Eccentric Connectivity Index [55], can predict binding to Medin. This can be explained by the predicted binding site of the model, likely stretching across the hairpin, indicating longer molecules will be needed to achieve higher binding strength. Specifically, topological descriptors are linearly related to K_D, the experimentally-measured binding affinity metric (Equation 1.1).

Looking ahead, we aim to incorporate the steps taken in the case study into a more robust lead selection pipeline. Using binding kinetic parameters to select drug leads is important to refine drug efficacy, select for specific kinetic patterns according to the target disorder, or explore different kinetic patterns if the nature of the target disorder is unknown [1]. The molecular clustering allows compound leads to be selected to explore the chemical space and potentially more than one mechanism of action [51]. Finally, a compound's specificity to its target is important when selecting lead compounds to ensure these drugs will bind *in vivo*. All the steps taken here are important to computationally screen large libraries to select the best potential compounds to bind to different proteins.

We present a model to reliably predict binding kinetic parameters pK_D, pK_{on} and pK_{off}. The model's performance is comparable to the accuracy of other experimental methods, given similar correlations between different experimental measures, and our "ground truth" and predicted values. We aim to improve the model's performance with the inclusion of more experimental data in training and incorporating the error associated with experimental measures of these parameters into the model's predictions. For Medin, we find the protein is ordered contrary to previous literature. The selection of molecules likely to bind to the protein uses a combination of molecular clustering techniques and binding affinity measures, which we hope to formally incorporate into a candidate compound

selection pipeline. For that selection, we characterise binding patterns and find a common binding site, indicating a binding pocket in the protein. This binding pocket forms across the β hairpin, tentatively indicating interference with the amyloid aggregation process. Finally, we characterise the binding specificity of these molecules to the target protein by comparing the binding to the whole proteome and our results indicate that the selection of drug leads will benefit from incorporating this measure into the pipeline. Through experimental validation of the model with aggregation assays for Medin and the selected molecules, we hope to confirm the model's ability to infer binding patterns between protein-ligand complexes and use the Ligand-Transformer for Kinetics in the drug discovery pipeline.

Bibliography

[1] See Hong Chiu and Lei Xie. Toward High-Throughput Predictive Modeling of Protein Binding/Unbinding Kinetics. *Journal of chemical information and modeling*, 56 (6):1164–1174, 6 2016. ISSN 1549-960X. doi: 10.1021/ACS.JCIM.5B00632. URL https://pubmed.ncbi.nlm.nih.gov/27159844/.

[2] Shengyu Zhang, Donghui Huo, Robert I Horne, Yumeng Qi, Sebastian Pujalte Ojeda, Aixia Yan, and Michele Vendruscolo. Sequence-based drug design using transformers. doi: 10.1101/2023.11.27.568880. URL https://doi.org/10.1101/2023.11.27.568880.

[3] Jessica Wagner, Karoline Degenhardt, Marleen Veit, Nikolaos Louros, Katerina Konstantoulea, Angelos Skodras, Katleen Wild, Ping Liu, Ulrike Obermüller, Vikas Bansal, Anupriya Dalmia, Lisa M. Häsler, Marius Lambert, Matthias De Vleeschouwer, Hannah A. Davies, Jillian Madine, Deborah Kronenberg-Versteeg, Regina Feederle, Domenico Del Turco, K. Peter R. Nilsson, Tammaryn Lashley, Thomas Deller, Marla Gearing, Lary C. Walker, Peter Heutink, Frederic Rousseau, Joost Schymkowitz, Mathias Jucker, and Jonas J. Neher. Medin co-aggregates with vascular amyloid-β in Alzheimer's disease. *Nature 2022 612:7938*, 612(7938):123–131, 11 2022. ISSN 1476-4687. doi: 10.1038/s41586-022-05440-3. URL https://www.nature.com/articles/s41586-022-05440-3.

[4] John Jumper, Richard Evans, Alexander Pritzel, Tim Green, Michael Figurnov, Olaf Ronneberger, Kathryn Tunyasuvunakool, Russ Bates, Augustin Žídek, Anna Potapenko, Alex Bridgland, Clemens Meyer, Simon A A Kohl, Andrew J Ballard, Andrew Cowie, Bernardino Romera-Paredes, Stanislav Nikolov, Rishub Jain, Jonas Adler, Trevor Back, Stig Petersen, David Reiman, Ellen Clancy, Michal Zielinski, Martin Steinegger, Michalina Pacholska, Tamas Berghammer, Sebastian Bodenstein, David Silver, Oriol Vinyals, Andrew W Senior, Koray Kavukcuoglu, Pushmeet Kohli, Demis Hassabis, and Demis Hassabis. Highly accurate protein structure prediction with AlphaFold. *Nature*, 596:583, 2021. doi: 10.1038/s41586-021-03819-2. URL https://doi.org/10.1038/s41586-021-03819-2.

[5] Liya Hu, Wilhelm Salmen, Banumathi Sankaran, Yi Lasanajak, David F. Smith, Sue E. Crawford, Mary K. Estes, and B. V.Venkataram Prasad. Novel fold of rotavirus glycan-binding domain predicted by AlphaFold2 and determined by X-ray crystallography. *Communications biology*, 5(1), 12 2022. ISSN 2399-3642. doi: 10.1038/S42003-022-03357-1. URL https://pubmed.ncbi.nlm.nih.gov/35513489/ https://pubmed.ncbi.nlm.nih.gov/35513489/?dopt=Abstract.

[6] Feng Ren, Xiao Ding, Min Zheng, Mikhail Korzinkin, Xin Cai, Wei Zhu, Alexey Mantsyzov, Alex Aliper, Vladimir Aladinskiy, Zhongying Cao, Shanshan Kong, Xi Long, Bonnie Hei Man Liu, Yingtao Liu, Vladimir Naumov, Anastasia Shneyderman, Ivan V. Ozerov, Ju Wang, Frank W. Pun, Daniil A. Polykovskiy, Chong Sun, Michael Levitt, Alán Aspuru-Guzik, and Alex Zhavoronkov. AlphaFold accelerates artificial intelligence powered drug discovery: efficient discovery of a novel CDK20 small molecule inhibitor. *Chemical science*, 14(6):1443–1452, 1 2023. ISSN 2041-6520. doi: 10.1039/D2SC05709C. URL https://pubmed.ncbi.nlm.nih.gov/36794205/https://pubmed.ncbi.nlm.nih.gov/36794205/?dopt=Abstract.

[7] Casper Goverde, Benedict Wolf, Hamed Khakzad, Stéphane Rosset, and Bruno E. Correia. De novo protein design by inversion of the AlphaFold structure prediction network. *bioRxiv*, page 2022.12.13.520346, 12 2022. doi: 10.1101/2022.12.13.520346. URL https://www.biorxiv.org/content/10.1101/2022.12.13.520346v1https://www.biorxiv.org/content/10.1101/2022.12.13.520346v1.abstract.

[8] Christian B. Anfinsen. Principles that govern the folding of protein chains. *Science (New York, N.Y.)*, 181(4096):223–230, 1973. ISSN 0036-8075. doi: 10.1126/SCIENCE.181.4096.223. URL https://pubmed.ncbi.nlm.nih.gov/4124164/.

[9] Amos Bairoch, Rolf Apweiler, Cathy H. Wu, Winona C. Barker, Brigitte Boeckmann, Serenella Ferro, Elisabeth Gasteiger, Hongzhan Huang, Rodrigo Lopez, Michele Magrane, Maria J. Martin, Darren A. Natale, Claire O'Donovan, Nicole Redaschi, and Lai Su L. Yeh. The Universal Protein Resource (UniProt). *Nucleic acids research*, 33(Database issue), 1 2005. ISSN 1362-4962. doi: 10.1093/NAR/GKI070. URL https://pubmed.ncbi.nlm.nih.gov/15608167/.

[10] Robin Pearce and Yang Zhang. Toward the solution of the protein structure prediction problem. *Journal of Biological Chemistry*, 297:100870, 2021. doi: 10.1016/j.jbc.2021.100870. URL https://doi.org/10.1016/j.jbc.2021.100870.

[11] Peter Güntert. Automated NMR structure calculation with CYANA. *Methods in molecular biology (Clifton, N.J.)*, 278, 2004. ISSN 10643745. doi: 10.1385/1-59259-809-9:353.

[12] Helen M. Berman, John Westbrook, Zukang Feng, Gary Gilliland, T. N. Bhat, Helge Weissig, Ilya N. Shindyalov, and Philip E. Bourne. The Protein Data Bank. *Nucleic acids research*, 28(1):235–242, 1 2000. ISSN 0305-1048. doi: 10.1093/NAR/28.1.235. URL https://pubmed.ncbi.nlm.nih.gov/10592235/.

[13] Jeffrey Skolnick and Hongyi Zhou. Why is There a Glass Ceiling for Threading Based Protein Structure Prediction Methods? *Journal of Physical Chemistry B*, 121 (15):3546–3554, 4 2017. ISSN 15205207. doi: 10.1021/ACS.JPCB.6B09517/ASSET/IMAGES/LARGE/JP-2016-09517P{_}0008.JPEG. URL https://pubs.acs.org/doi/full/10.1021/acs.jpcb.6b09517.

[14] James U. Bowie and David Eisenberg. An evolutionary approach to folding small alpha-helical proteins that uses sequence information and an empirical guiding fitness function. *Proceedings of the National Academy of Sciences*, 91(10):4436–4440, 5 1994. ISSN 00278424. doi: 10.1073/PNAS.91.10.4436. URL https://www.pnas.org/doi/abs/10.1073/pnas.91.10.4436.

[15] Carol A. Rohl, Charlie E.M. Strauss, Kira M.S. Misura, and David Baker. Protein Structure Prediction Using Rosetta. *Methods in Enzymology*, 383, 2004. ISSN 00766879. doi: 10.1016/S0076-6879(04)83004-0.

[16] Rakesh Trivedi and Hampapathalu Adimurthy Nagarajaram. Intrinsically Disordered Proteins: An Overview, 2022. ISSN 14220067.

[17] Gang Hu, Kui Wang, Jiangning Song, Vladimir N Uversky, and Lukasz Kurgan. Dark Proteomes www.proteomics-journal.com Taxonomic Landscape of the Dark Proteomes: Whole-Proteome Scale Interplay Between Structural Darkness, Intrinsic Disorder, and Crystallization Propensity. doi: 10.1002/pmic.201800243. URL https://doi.org/10.1002/pmic.201800243.

[18] Alex S. Holehouse and Birthe B. Kragelund. The molecular basis for cellular function of intrinsically disordered protein regions, 2024. ISSN 14710080.

[19] Massimiliano Bonomi, Gabriella T. Heller, Carlo Camilloni, and Michele Vendruscolo. Principles of protein structural ensemble determination, 2017. ISSN 1879033X.

[20] Antonio Deiana, Sergio Forcelloni, Alessandro Porrello, and Andrea Giansanti. Intrinsically disordered proteins and structured proteins with intrinsically disordered regions have different functional roles in the cell. *PloS one*, 14(8), 8 2019. ISSN 1932-6203. doi: 10.1371/JOURNAL.PONE.0217889. URL https://pubmed.ncbi.nlm.nih.gov/31425549/.

[21] Ryan J. Emenecker, Karina Guadalupe, Nora M. Shamoon, Shahar Sukenik, and Alex S. Holehouse. Sequence-ensemble-function relationships for disordered proteins in live cells. *bioRxiv*, page 2023.10.29.564547, 11 2023. doi: 10.1101/2023.10.29.564547. URL `https://www.biorxiv.org/content/10.1101/2023.10.29.564547v1https://www.biorxiv.org/content/10.1101/2023.10.29.564547v1.abstract`.

[22] Vladimir N. Uversky, Christopher J. Oldfield, and A. Keith Dunker. Intrinsically disordered proteins in human diseases: introducing the D2 concept. *Annual review of biophysics*, 37:215–246, 2008. ISSN 1936-122X. doi: 10.1146/ANNUREV.BIOPHYS.37.032807.125924. URL `https://pubmed.ncbi.nlm.nih.gov/18573080/`.

[23] Vladimir N. Uversky. Intrinsically disordered proteins and their (disordered) proteomes in neurodegenerative disorders. *Frontiers in Aging Neuroscience*, 7(MAR), 2015. ISSN 16634365. doi: 10.3389/fnagi.2015.00018.

[24] Lyubomir T. Vassilev, Binh T. Vu, Bradford Graves, Daisy Carvajal, Frank Podlaski, Zoran Filipovic, Norman Kong, Ursula Kammlott, Christine Lukacs, Christian Klein, Nader Fotouhi, and Emily A. Liu. In vivo activation of the p53 pathway by small-molecule antagonists of MDM2. *Science (New York, N.Y.)*, 303(5659):844–848, 2 2004. ISSN 1095-9203. doi: 10.1126/SCIENCE.1092472. URL `https://pubmed.ncbi.nlm.nih.gov/14704432/`.

[25] Pietrobono D, Giacomelli C, Trincavelli ML, Daniele S, and Martini C. Inhibitors of protein aggregates as novel drugs in neurodegenerative diseases. *Global Drugs and Therapeutics*, 2(3), 2017. doi: 10.15761/GDT.1000119.

[26] Neil J. Bruce, Gaurav K. Ganotra, Daria B. Kokh, S. Kashif Sadiq, and Rebecca C. Wade. New approaches for computing ligand–receptor binding kinetics. *Current Opinion in Structural Biology*, 49:1–10, 4 2018. ISSN 0959-440X. doi: 10.1016/J.SBI.2017.10.001.

[27] David C. Swinney. Molecular Mechanism of Action (MMoA) in Drug Discovery. *Annual Reports in Medicinal Chemistry*, 46:301–317, 1 2011. ISSN 0065-7743. doi: 10.1016/B978-0-12-386009-5.00009-6.

[28] Wade Borcherds, François Xavier Theillet, Andrea Katzer, Ana Finzel, Katie M. Mishall, Anne T. Powell, Hongwei Wu, Wanda Manieri, Christoph Dieterich, Philipp Selenko, Alexander Loewer, and Gary W. Daughdrill. Disorder and residual helicity alter p53-Mdm2 binding affinity and signaling in cells. *Nature chemical biology*, 10 (12), 2014. ISSN 15524469. doi: 10.1038/nchembio.1668.

[29] Basile I.M. Wicky, Sarah L. Shammas, and Jane Clarke. Affinity of IDPs to their targets is modulated by ion-specific changes in kinetics and residual structure. *Proceedings of the National Academy of Sciences of the United States of America*, 114 (37), 2017. ISSN 10916490. doi: 10.1073/pnas.1705105114.

[30] T. Reid Alderson, Iva Pritišanac, Desika Kolaric, Alan M. Moses, and Julie D. Forman-Kay. Systematic identification of conditionally folded intrinsically disordered regions by AlphaFold2. *Proceedings of the National Academy of Sciences of the United States of America*, 120(44):e2304302120, 10 2023. ISSN 10916490. doi: 10.1073/PNAS.2304302120/SUPPL{_}FILE/PNAS.2304302120.SAPP.PDF. URL https://www.pnas.org/doi/abs/10.1073/pnas.2304302120.

[31] Shengchao Liu, Hanchen Wang, Weiyang Liu, Joan Lasenby, Hongyu Guo, and Jian Tang. Pre-training Molecular Graph Representation with 3D Geometry. *ICLR 2022 - 10th International Conference on Learning Representations*, 10 2021. URL https://arxiv.org/abs/2110.07728v2.

[32] B. O. Häggqvist, Jan Näslund, Knut Sletten, Gunilla T. Westermark, Gerd Mucchiano, Lars O. Tjernberg, Christer Nordstedt, Ulla Engström, and Per Westermark. Medin: an integral fragment of aortic smooth muscle cell-produced lactadherin forms the most common human amyloid. *Proceedings of the National Academy of Sciences of the United States of America*, 96(15):8669–8674, 7 1999. ISSN 0027-8424. doi: 10.1073/PNAS.96.15.8669. URL https://pubmed.ncbi.nlm.nih.gov/10411933/.

[33] Dong Guo, Julia M. Hillger, Adriaan P. Ijzerman, and Laura H. Heitman. Drug-target residence time–a case for G protein-coupled receptors. *Medicinal research reviews*, 34(4):856–892, 2014. ISSN 1098-1128. doi: 10.1002/MED.21307. URL https://pubmed.ncbi.nlm.nih.gov/24549504/.

[34] David A Sykes, Holly Moore, Lisa Stott, Nicholas Holliday, Jonathan A Javitch, J Robert Lane, and Steven J Charlton. Extrapyramidal side effects of antipsychotics are linked to their association kinetics at dopamine D 2 receptors. doi: 10.1038/s41467-017-00716-z. URL www.nature.com/naturecommunications.

[35] Kin Sing Stephen Lee, Jun Yang, Jun Niu, Connie J. Ng, Karen M. Wagner, Hua Dong, Sean D. Kodani, Debin Wan, Christophe Morisseau, and Bruce D. Hammock. Drug-Target Residence Time Affects in Vivo Target Occupancy through Multiple Pathways. *ACS Central Science*, 5(9):1614–1624, 9 2019. ISSN 23747951. doi: 10.1021/ACSCENTSCI.9B00770/ASSET/IMAGES/MEDIUM/OC9B00770{_}0005.GIF. URL https://pubs.acs.org/doi/full/10.1021/acscentsci.9b00770.

[36] Gilbert J. Kersh, Ellen N. Kersh, Daved H. Fremont, and Paul M. Allen. High- and low-potency ligands with similar affinities for the TCR: the importance of kinetics in TCR signaling. *Immunity*, 9(6):817–826, 1998. ISSN 1074-7613. doi: 10.1016/S1074-7613(00)80647-0. URL https://pubmed.ncbi.nlm.nih.gov/9881972/.

[37] Ming Kuang, Jingwei Zhou, Laiyou Wang, Zhihong Liu, Jiao Guo, and Ruibo Wu. Binding Kinetics versus Affinities in BRD4 Inhibition. *Journal of Chemical Information and Modeling*, 55(9), 2015. ISSN 1549960X. doi: 10.1021/acs.jcim.5b00265.

[38] Dong Guo, Laura H. Heitman, and Adriaan P. Ijzerman. The Role of Target Binding Kinetics in Drug Discovery. *ChemMedChem*, 10(11):1793–1796, 11 2015. ISSN 1860-7187. doi: 10.1002/CMDC.201500310. URL https://onlinelibrary.wiley.com/doi/full/10.1002/cmdc.201500310https://onlinelibrary.wiley.com/doi/abs/10.1002/cmdc.201500310https://chemistry-europe.onlinelibrary.wiley.com/doi/10.1002/cmdc.201500310.

[39] Daria B. Kokh, Marta Amaral, Joerg Bomke, Ulrich Grädler, Djordje Musil, Hans Peter Buchstaller, Matthias K. Dreyer, Matthias Frech, Maryse Lowinski, Francois Vallee, Marc Bianciotto, Alexey Rak, and Rebecca C. Wade. Estimation of Drug-Target Residence Times by τ-Random Acceleration Molecular Dynamics Simulations. *Journal of chemical theory and computation*, 14(7):3859–3869, 7 2018. ISSN 1549-9626. doi: 10.1021/ACS.JCTC.8B00230. URL https://pubmed.ncbi.nlm.nih.gov/29768913/.

[40] Nurlybek Amangeldiuly, Dmitry Karlov, and Maxim V. Fedorov. Baseline Model for Predicting Protein-Ligand Unbinding Kinetics through Machine Learning. *Journal of chemical information and modeling*, 60(12):5946–5956, 12 2020. ISSN 1549-960X. doi: 10.1021/ACS.JCIM.0C00450. URL https://pubmed.ncbi.nlm.nih.gov/33183000/.

[41] Doris A. Schuetz, Lars Richter, Riccardo Martini, and Gerhard F. Ecker. A structure–kinetic relationship study using matched molecular pair analysis. *RSC Medicinal Chemistry*, 11(11):1285, 11 2020. doi: 10.1039/D0MD00178C. URL /pmc/articles/PMC8126976//pmc/articles/PMC8126976/?report=abstracthttps://www.ncbi.nlm.nih.gov/pmc/articles/PMC8126976/.

[42] Karoline Degenhardt, Jessica Wagner, Angelos Skodras, Michael Candlish, Anna Julia Koppelmann, Katleen Wild, Rusheka Maxwell, Carola Rotermund, Felix Von Zweydorf, Christian Johannes Gloeckner, Hannah A. Davies, Jillian Madine, Domenico Del Turco, Regina Feederle, Tammaryn Lashley, Thomas Deller, Philipp Kahle, Jasmin K. Hefendehl, Mathias Jucker, and Jonas J. Neher. Medin aggregation causes

cerebrovascular dysfunction in aging wild-type mice. *Proceedings of the National Academy of Sciences of the United States of America*, 117(38), 2020. ISSN 10916490. doi: 10.1073/pnas.2011133117.

[43] O. Sumner Makin, Edward Atkins, Pawel Sikorski, Jan Johansson, and Louise C. Serpell. Molecular basis for amyloid fibril formation and stability. *Proceedings of the National Academy of Sciences of the United States of America*, 102(2), 2005. ISSN 00278424. doi: 10.1073/pnas.0406847102.

[44] Hannah A. Davies, Jillian Madine, and David A. Middleton. Comparisons with Amyloid-β Reveal an Aspartate Residue That Stabilizes Fibrils of the Aortic Amyloid Peptide Medin. *Journal of Biological Chemistry*, 290(12), 2015. ISSN 1083351X. doi: 10.1074/jbc.M114.602177.

[45] Jillian Madine, Alastair Copland, Louise C. Serpell, and David A. Middleton. Cross-β Spine architecture of fibrils formed by the amyloidogenic segment NFGSVQFV of medin from solid-state NMR and X-ray fiber diffraction measurements. *Biochemistry*, 48(14), 2009. ISSN 00062960. doi: 10.1021/bi802164e.

[46] Hannah A. Davies, Daniel J. Rigden, Marie M. Phelan, and Jillian Madine. Probing Medin Monomer Structure and its Amyloid Nucleation Using 13C-Direct Detection NMR in Combination with Structural Bioinformatics. *Scientific Reports 2017 7:1*, 7(1):1–10, 3 2017. ISSN 2045-2322. doi: 10.1038/srep45224. URL `https://www.nature.com/articles/srep45224`.

[47] Milot Mirdita, Konstantin Schütze, Yoshitaka Moriwaki, Lim Heo, Sergey Ovchinnikov, and Martin Steinegger. ColabFold: making protein folding accessible to all. *Nature Methods 2022 19:6*, 19(6):679–682, 5 2022. ISSN 1548-7105. doi: 10.1038/s41592-022-01488-1. URL `https://www.nature.com/articles/s41592-022-01488-1`.

[48] Betsabeh Tanoori and Mansoor Zolghadri Jahromi. Using drug-drug and protein-protein similarities as feature vector for drug-target binding prediction. *Chemometrics and Intelligent Laboratory Systems*, 217, 2021. ISSN 18733239. doi: 10.1016/j.chemolab.2021.104405.

[49] Bálint Mészáros, Gábor Erdös, and Zsuzsanna Dosztányi. IUPred2A: context-dependent prediction of protein disorder as a function of redox state and protein binding. *Nucleic Acids Research*, 46(W1):W329–W337, 7 2018. ISSN 0305-1048. doi: 10.1093/NAR/GKY384. URL `https://dx.doi.org/10.1093/nar/gky384`.

[50] Andreas Bender and Robert C. Glen. Molecular similarity: a key technique in molecular informatics. *Organic & Biomolecular Chemistry*, 2(22): 3204–3218, 11 2004. ISSN 1477-0539. doi: 10.1039/B409813G. URL https://pubs.rsc.org/en/content/articlehtml/2004/ob/b409813ghttps://pubs.rsc.org/en/content/articlelanding/2004/ob/b409813g.

[51] Darko Butina. Unsupervised data base clustering based on daylight's fingerprint and Tanimoto similarity: A fast and automated way to cluster small and large data sets. *Journal of Chemical Information and Computer Sciences*, 39(4):747–750, 1999. ISSN 00952338. doi: 10.1021/CI9803381/ASSET/IMAGES/MEDIUM/CI9803381E00016.GIF. URL https://pubs.acs.org/doi/full/10.1021/ci9803381.

[52] Chun Wei Yap. PaDEL-descriptor: an open source software to calculate molecular descriptors and fingerprints. *Journal of computational chemistry*, 32(7):1466–1474, 5 2011. ISSN 1096-987X. doi: 10.1002/JCC.21707. URL https://pubmed.ncbi.nlm.nih.gov/21425294/.

[53] Zsuzsanna Dosztányi, Veronika Csizmók, Péter Tompa, and István Simon. The pairwise energy content estimated from amino acid composition discriminates between folded and intrinsically unstructured proteins. *Journal of Molecular Biology*, 347(4), 2005. ISSN 00222836. doi: 10.1016/j.jmb.2005.01.071.

[54] Cory D. DuPai, Bryan W. Davies, and Claus O. Wilke. A systematic analysis of the beta hairpin motif in the Protein Data Bank. *Protein Science*, 30(3), 2021. ISSN 1469896X. doi: 10.1002/pro.4020.

[55] Vikas Sharma, Reena Goswami, and A. K. Madan. Eccentric connectivity index: A novel highly discriminating topological descriptor for structure-property and structure-activity studies. *Journal of Chemical Information and Computer Sciences*, 37(2):273–282, 1997. ISSN 00952338. doi: 10.1021/CI960049H/ASSET/IMAGES/LARGE/CI960049HF00014.JPEG. URL https://pubs.acs.org/doi/full/10.1021/ci960049h.

[56] Yujing Zhao, Qilei Liu, Jian Du, — Qingwei Meng, and Lei Zhang. Machine learning methods for developments of binding kinetic models in predicting protein-ligand dissociation rate constants. 2023. doi: 10.1002/smo.20230012. URL https://onlinelibrary.wiley.com/doi/10.1002/smo.20230012.

[57] Victoria Georgi, Felix Schiele, Benedict-Tilman Berger, Andreas Steffen, Paula A Marin Zapata, Hans Briem, Stephan Menz, Cornelia Preusse, James D Vasta,

Matthew B Robers, Michael Brands, Stefan Knapp, and Amaury Ferna. Binding Kinetics Survey of the Drugged Kinome. 2018. doi: 10.1021/jacs.8b08048. URL `https://pubs.acs.org/sharingguidelines`.

[58] Benedict Tilman Berger, Marta Amaral, Daria B. Kokh, Ariane Nunes-Alves, Djordje Musil, Timo Heinrich, Martin Schröder, Rebecca Neil, Jing Wang, Iva Navratilova, Joerg Bomke, Jonathan M. Elkins, Susanne Müller, Matthias Frech, Rebecca C. Wade, and Stefan Knapp. Structure-kinetic relationship reveals the mechanism of selectivity of FAK inhibitors over PYK2. *Cell Chemical Biology*, 28(5), 2021. ISSN 24519448. doi: 10.1016/j.chembiol.2021.01.003.

[59] Fengjuan Huang, Xinjie Fan, Ying Wang, Yu Zou, Jiangfang Lian, Chuang Wang, Feng Ding, and Yunxiang Sun. Computational insights into the cross-talk between medin and Aβ: implications for age-related vascular risk factors in Alzheimer's disease. *Briefings in Bioinformatics*, 25(2), 2024. ISSN 14774054. doi: 10.1093/bib/bbad526.

Appendix A

Supplementary figures

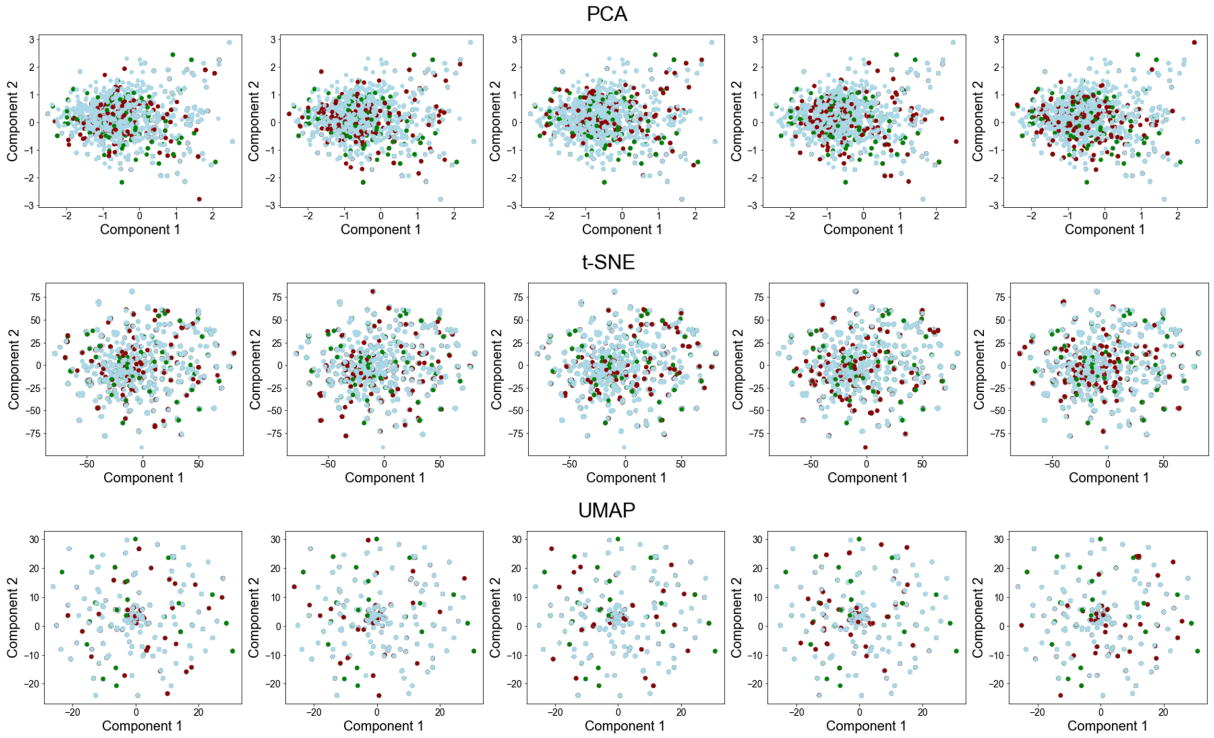

Figure A.1: **Chemical space coverage for folds in cross-validation** Morgan fingerprints for each molecule were computed (radius = 2; bits = 124) and dimensionality reduction techniques used to visualise the space. Each row represents a dimensionality reduction technique and each column represents a split in the cross-validation. The plots show good chemical space coverage for each fold in the set with test points shown in green, validation points shown in red and training points shown in blue.

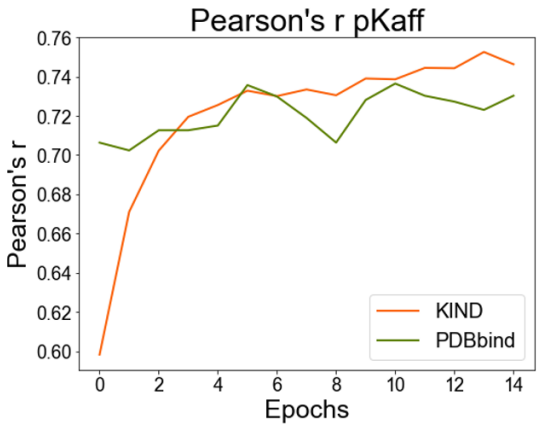

Figure A.2: **Kinetics Head for pK_D prediction.** During the external validation, the Kinetics Head was used to calculate predicted pK_D values for the KIND and PDBbind. We show the model performs similarly on both datasets.

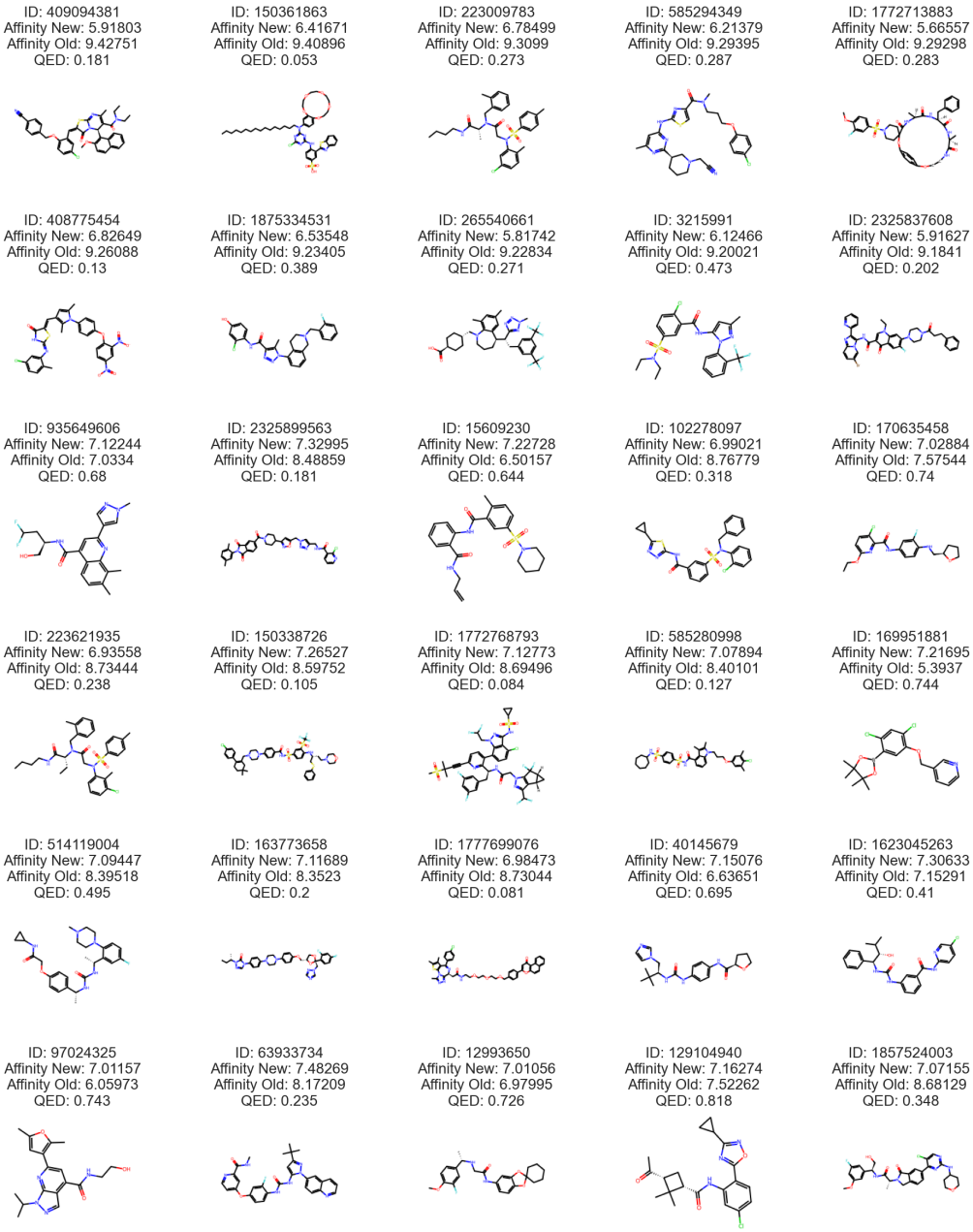

Figure A.3: **Structures of selected molecules.** The figure shows the drawings of structures from the selected molecules

Table A.1: Compound selection full information.

Index	ZINC ID	SMILES	Affinity Model 1	Affinity Model 2	QED
M1	409094381	CCN(CC)C(=O)C1=C(C)N=c2s/c(=C/c3cc(Cl)ccc3OCc3ccc(C#N)cc3)c(=O)n2[C@@H]1c1c(OC)ccc2ccccc12	9.427	5.918	0.180
M2	150361863	CCCCCCCCCCCCCCCCCCN(c1ccc2c(c1)OCCOCCOCCOCCO2)c1nc(Cl)nc(Nc2ccc(S(=O)(=O)O)cc2-c2nc3ccccc3s2)n1	9.408	6.416	0.052
M3	223009783	CCCCNC(=O)[C@@H](C)N(Cc1ccccc1C)C(=O)CN(c1cc(Cl)ccc1C)S(=O)(=O)c1ccc(C)cc1	9.30	6.784	0.273
M4	585294349	Cc1cc(Nc2nc(C(=O)N(C)CCCOc3ccc(Cl)cc3)cs2)nc([C@@H]2CCCN(CC#N)C2)n1	9.293	6.213	0.286
M5	1772713883	COc1ccc(S(=O)(=O)N2CCC3(CC2)Oc2ccc(cc2)OCCNC(=O)[C@H](C)NC(=O)[C@H](Cc2ccccc2)NC(=O)[C@H](C)NC3=O)cc1F	9.292	5.665	0.283
M6	408775454	Cc1ccc(Cl)cc1N=C1NC(=O)/C(=C/c2cc(C)n(-c3ccc(Oc4ccc([N+](=O)[O-])cc4[N+](=O)[O-])cc3)c2C)S1	9.260	6.826	0.130
M7	1875334531	Cc1c(C(=O)Nc2ccc(O)cc2Cl)nnn1-c1cccc2c1CCN(Cc1ccccc1F)C2	9.234	6.535	0.388
M8	265540661	Cc1cc(C)c2c(c1)[C@@H]([C@@H](Cc1cc(C(F)(F)F)cc(C(F)(F)F)c1)c1nnn(C)n1)CCCN2C[C@H]1CC[C@H](C(=O)O)CC1	9.228	5.817	0.270
M9	3215991	CCN(CC)S(=O)(=O)c1ccc(Cl)c(C(=O)Nc2cc(C)nn2-c2ccccc2C(F)(F)F)c1	9.200	6.124	0.473
M10	2325837608	CCn1cc(C(=O)Nc2c(-c3ccccn3)nc3ccc(Br)cn23)c(=O)c2cc(F)c(N3CCN(C(=O)CCc4ccccc4)CC3)cc21	9.18	5.916	0.201
M11	935649606	Cc1ccc2c(C(=O)N[C@@H](CO)CC(F)F)cc(-c3cnn(C)c3)nc2c1C	7.03	7.122	0.680
M12	2325899563	Cc1ccc(C)c(N2C(=O)c3ccc(C(=O)N4CCC(c5cc(Cn6cc(CNC(=O)c7cccnc7Cl)nn6)on5)CC4)cc3C2=O)c1	8.488	7.329	0.181
M13	15609230	C=CCNC(=O)c1ccccc1NC(=O)c1cc(S(=O)(=O)N2CCCCC2)ccc1C	6.501	7.227	0.6445
M14	102278097	O=C(Nc1nnc(C2CC2)s1)c1cccc(S(=O)(=O)N(Cc2ccccc2)c2ccccc2Cl)c1	8.767	6.9902	0.318
M15	170635458	CCOc1ccc(Cl)c(C(=O)Nc2ccc(NC[C@H]3CCCO3)c(F)c2)n1	7.575	7.028	0.74
M16	223621935	CCCCNC(=O)[C@@H](CC)N(Cc1ccccc1C)C(=O)CN(c1cccc(Cl)c1C)S(=O)(=O)c1ccc(C)cc1	8.734	6.935	0.238
M17	150338726	CC1(C)CCC(c2ccc(Cl)cc2)=C(CN2CCN(c3ccc(C(=O)NS(=O)(=O)c4ccc(N[C@H](CCN5CCOCC5)CSc5ccccc5)c(S(=O)(=O)C(F)(F)F)c4)cc3)CC2)C1	8.597	7.265	0.104
M18	1772768793	CC(C)(C#Cc1ccc(-c2ccc(Cl)c3c(NS(=O)(=O)C4CC4)nn(CC(F)F)c23)c([C@H](Cc2cc(F)cc(F)c2)NC(=O)Cn2nc(C(F)F)c3c2C(F)(F)[C@@H]2C[C@H]32)n1)S(C)(=O)=O	8.694	7.127	0.047
M19	23349660	COc1cc(OC)c2c(c1)c(CN1CCN(c3cccc(Cl)c3)CC1)c(=O)c1c(cccc21)C(=O)N(C)c1ccccc1C	9.240	6.032	0.251
M20	539026178	CC[C@@H](C)n1ncn(-c2ccc(N3CCN(c4ccc(OCc5ccc(C(=O)O)cc5)cc4)CC3)cc2)c1=O	8.598	7.136	0.253
M21	155816748	COC(=O)[C@@H]1CC[C@@H](C(=O)NCCOCCOCCOCc2ccc(-c3cc(=O)c4ccccc4o3)cc2)CC1	8.395	7.094	0.494
M22	163773658	CC[C@H](C)n1ncn(-c2ccc(N3CCN(c4ccc(OC[C@@H]5CO[C@](Cn6cncn6)(c6cc(F)cc6F)O5)cc4)CC3)cc2)c1=O	8.35	7.116	0.2
M23	1777699076	Cc1sc2c(c1C)C(c1ccc(Cl)cc1)=N[C@@H](CC(=O)NCCOCCOCCOCc1ccc(-c3cc(=O)c4c(ccc5ccccc54)o3)cc1)c1nnc(C)n1-2	8.730	6.984	0.081
M24	40145679	CC(C)(C)[C@@H](Cn1ccnc1)NC(=O)Nc1ccc(NC(=O)[C@H]2CCCO2)cc1	6.636	7.150	0.694
M25	1623045263	CC(C)[C@H](O)[C@@H](NC(=O)Nc1cccc(C(=O)Nc2ccc(Cl)cn2)c1)c1ccccc1	7.152	7.306	0.411
M26	97024325	Cc1cc(-c2cc(C(=O)NCCO)c3cnn(C(C)C)c3n2)c(C)o1	6.059	7.011	0.743
M27	63933734	CNC(=O)c1cc(Oc2ccc(NC(=O)Nc3cc(C(C)(C)C)nn3-c3ccc4ncccc4c3)c(F)c2)ccn1	8.172	7.482	0.235
M28	12993650	COc1ccc([C@H](C)NCC(=O)Nc2ccc3c(c2)OC2(CCCCC2)O3)cc1F	6.979	7.010	0.725
M29	129104940	CC(=O)[C@@H]1C[C@H](C(=O)Nc2cc(Cl)ccc2-c2nc(C3CC3)no2)C1(C)C	7.522	7.162	0.8189
M30	1857524003	COc1cc(F)cc([C@@H](CO)NC(=O)[C@@H](C)N2Cc3ccc(-c4nc(NC5CCOCC5)ncc4Cl)cc3C2=O)c1	8.681	7.071	0.347