Linking QSAR-based drug-target prediction with the AlphaFold dataset

George Iniatis

School of Computing Science
Sir Alwyn Williams Building
University of Glasgow
G12 8QQ

Interim Report

December 16, 2022

Contents

1 Introduction

This section will introduce the project on a high level and examine its motivations and objectives.

1.1 Motivation

Drug-target interactions (DTIs) refer to the interactions of chemical compounds and biological targets, proteins in our case, inside the human body (Sachdev and Gupta 2019). Given that both proteins and drugs are chemically active molecules in the bloodstream, it would make sense that they interact in some way (Yartsev 2022). These interactions usually form an ever-changing, benign and reversible binding where both molecules move through the bloodstream interlocked together (Yartsev 2022). The vast majority of drugs administered take this into account and use this process (Yartsev 2022). A protein-bound drug is usually too big to pass through a biological membrane like that of a cell. Therefore only the unbound drug, usually in equilibrium with the bound drug, can pass through and produce the desired pharmacological effect, like the treatment of a disease, or the targeting of a tumour (Davis 2018).

Consequently, the degree of how much a drug binds to a protein can enhance or diminish the drug's effectiveness and performance. For example, minimally protein-bound drugs tend to penetrate tissue better and are excreted much faster from the body than those highly bound (Scheife 1989). In contrast, highly protein-bound drugs, usually meaning that the protein binding is so impactful that we have to pay attention to it, tend to last much longer. This is because the protein acts as a drug "depot" that slowly releases the drug into the bloodstream, again keeping the bound and free drug in equilibrium (Yartsev 2022; Davis 2018).

DTIs play a crucial role in drug discovery and pharmacology. However, the experimental determination of these interactions with methods, such as fluorescence assays, is time-consuming and limited due to funding and the difficulty of purifying proteins (Shar et al. 2016; Wang et al. 2020). Past quantitative structure-relationship activity (QSAR) studies discussing protein-drug binding focused on testing thousands of drugs with just a single protein that they deemed important enough (Colmenarejo 2003; Estrada et al. 2006; Vallianatou et al. 2013). These studies would often not even consider the protein's sequence or structural information, concentrating their efforts on the drug molecules and their descriptors. However, this is not what we aim to do in this study. Unwanted or unexpected DTIs could cause severe side effects, therefore, the creation of in silico machine learning models with high throughput that can quickly and confidently predict whether thousands of drugs and proteins bind together and how much could be crucial for medicinal chemistry and drug development, acting as a supplement to biological experiments (Shar et al. 2016; Wang et al. 2020).

1.2 Objectives

The project aims to gather publicly available data on known drug-target interactions and place them into a new curated dataset. Then, using this new dataset, train multiple machine learning models using simple QSAR descriptors derived from a drug's chemical properties and a protein's sequence and 3D structural information to predict whether they bind together. Each protein's 3D structure will be extracted from the AlphaFold protein structure database (Jumper et al. 2021; Varadi et al. 2022) and one of the main challenges of the project will be in creating a simple embedding, which efficiently encodes the structural information of the protein, that can then be used in our training process. The models created should then be evaluated on robustness and performance, and a rudimentary system using these models should be constructed.

2 Background Survey

This section will cover some essential background information without going too in-depth, especially in biomedical concepts. We will also explore how past studies have tackled variations of the problem, or parts of it, and any valuable techniques, strategies, and knowledge that could be extracted from their experiments.

The following background research was critical in better understanding this previously unknown problem which included complex chemical and medical concepts, while also introducing us to key machine learning concepts and best practices.

Given the important nature of the problem, as discussed in Section 1.1, there have been numerous attempts to construct machine learning models using various methods, techniques, and biochemical properties (Shar et al. 2016; Wang et al. 2020; Jiang et al. 2020).

Classification models try to predict whether a particular drug will bind with a selected protein (Active) or not (Inactive). However, their accuracy will be influenced by the threshold used to separate the two classes as suggested by Shar et al. (2016), and the definition of a successful binding may vary substantially for different proteins. Regression models can address these problems by trying to predict the drug-binding affinity, which can take multiple forms, with the dissociation constant (K_d), the inhibition constant (K_i), and the 50% inhibitory concentration (IC_{50}) being the most common amongst them (Jiang et al. 2020). These values are usually represented by their logarithmic versions.

2.1 Proteins Outline

Proteins are large complex molecules essential to all biological processes in every living thing (*AlphaFold Blog* 2020; O'Connor et al. 2010). They help with digestion, blood circulation, and muscle movement, provide structures, and defend our bodies from diseases. They are made from a combination of amino acids, and the interactions between these chains of amino acids make the protein fold. Each amino acids sequence usually maps 1-to-1 to a 3D structure, and that 3D structure defines what the protein does and how it works (Fridman 2020). If a protein is misfolded, it can lead to diseases such as Alzheimer's and Parkinson's (Chaudhuri and Paul 2006). There are about 10^{300} ways to fold a protein given its amino acid sequence (Levinthal 1969). An almost impossible-to-solve problem known as 'Levinthal's paradox' or more simply as the 'Protein Folding Problem', a problem that scientists worldwide have been trying to solve for the past 50 years, and a problem that nature solves effortlessly in milliseconds (*AlphaFold Blog* 2020; Torrisi et al. 2020).

2.2 AlphaFold Breakthrough

We are aware of billions of proteins, and the number keeps increasing, but we only know the exact 3D shape of a small minority of these, roughly 170,000 (Jumper et al. 2021).

Mapping these proteins using state-of-the-art methods such as X-ray crystallography and nuclear magnetic resonance is costly, time-consuming, and relies on extensive trial and error, making them highly inefficient and unsuitable for high-throughput screening (HTS). So naturally, scientists worldwide wanted to create a system that could predict a protein's 3D structure just by its amino acid makeup. This is precisely what DeepMind tried to achieve by creating an AI system called AlphaFold, trained on the known sequences and structures of the manually mapped-out proteins.

DeepMind's latest AlphaFold AI system, AlphaFold2, provided the first highly accurate and novel computational solution to this problem, not solving it in its entirety but arguably taking a large step forward (Jumper et al. 2021). This was demonstrated at the 14th Critical Assessment of protein Structure Prediction (CASP), where AlphaFold outperformed all the other entries in the competition and achieved accuracy similar to that of experimental methods. CASP is organised every two years and uses recently discovered protein 3D structures as a blind test for the prediction systems submitted. It serves as the gold-standard assessment for the prediction accuracy of protein structures.

This breakthrough allowed DeepMind to release protein structure predictions covering almost the entire human proteome (98.5%). This mapping out of previously unknown human protein structures can provide us with highly beneficial information, allowing science to understand biological processes better and create more targeted, and therefore more effective, interventions (Tunyasuvunakool et al. 2021).

Discussing AlphaFold in detail is out of the scope of this project. However, it is an incredibly complex state-of-the-art system that directly predicts the 3D coordinates of all heavy atoms for a given protein just by using its amino acid sequence. We will not pretend that we understand completely how it works under the hood, but what we can all appreciate is its achievement and significance to the scientific community and the world as a whole.

2.3 Chemoinformatics Overview

Chemoinformatics is a well-established discipline, embracing chemistry and computer science, as its name suggests, focused on extracting, processing and extrapolating valuable data from chemical structures (Lo et al. 2018).

The rapid rise of big data worldwide and in every field of science has caused machine learning to become an invaluable and rapidly evolving tool for computer-aided drug discovery, contrary to physical mathematical models like molecular dynamics simulations or quantum chemistry (Lo et al. 2018). Machine learning models are much more effective and scalable to larger datasets. They use pattern recognition algorithms to discover mathematical connections between observed small molecules and extrapolate them to predict novel compounds' physical, chemical and biological properties.

For example, we might wish to optimise the chemical structure of a novel drug to improve its biological responses or binding affinity. Fifty years ago, this problem would require multiple

expensive, labour-intensive and time-consuming medicinal chemistry synthesis and analysis cycles. However, today machine learning techniques can be used to accurately predict, in silico, how specific chemical modifications to the drug could influence its biological behaviour. Then, physical experiments could confirm their findings (Lo et al. 2018).

2.4 Molecular Descriptors

Molecular descriptors are numerical features extracted from chemical structures that can be one-dimensional (0D or 1D), 2D, 3D or 4D (Lo et al. 2018). These descriptors are nicely summarised in Table 1.

Although very simple to compute or extract, one-dimensional descriptors contain little contextual information on their own. Instead, they describe aggregate information such as atom counts, bond counts, fragment counts and molecular weight. In addition, multiple chemical structures can have the same value for a common descriptor, making the usage of just a single 1D descriptor nearly meaningless. Therefore, they are usually expressed as feature vectors of multiple 1D descriptors or used together with descriptors of higher dimensionality.

Two-dimensional descriptors are the most common type found in literature and include molecular profiles, topological indices and 2D auto-correlation descriptors. Graph invariance differentiates them from 1D descriptors and makes them more valuable, meaning that descriptor values are not affected by the renumbering of graph nodes.

Three-dimensional descriptors extract chemical features and information from 3D coordinate representations and are considered to be the most sensitive to chemical structural differences. They include auto-correlation descriptors, quantum-chemical descriptors, substituent constants and surface:volume descriptors. They help identify distinct chemical scaffolds with similar binding activities, also known as 'scaffold hops'. However, one of their fundamental limitations is the computational complexity of structure alignments added when performing QSAR analysis.

Four-dimensional descriptors are an extension of 3D descriptors with the addition that they simultaneously consider multiple structural confrontations.

Commonly used software tools that can compute these types of descriptors include Dragon (Mauri et al. 2006) used by Shar et al. (2016), Mold2 (Hong et al. 2008), and PaDEL (Yap 2011).

2.4.1 Molecular Fingerprints

The molecular fingerprints of sub-structures can effectively capture the molecular information of drugs by converting them into a bit vector containing 0s and 1s (Wang et al. 2020). Each molecular sub-structure is mapped to a position in the bit vector. If a molecule

contains a molecular sub-structure, a value of 1 is assigned to the corresponding bit in the vector or a 0 otherwise (*PubChem Fingerprints* 2022). Even though this method splits the molecule into individual fragments, it still retains the entire structure of the drug molecule (Wang et al. 2020).

One of the most common molecular fingerprint sources is PubChem (Kim et al. 2021) which contains 881 molecular sub-structures (*PubChem Fingerprints* 2022). It should be noted that when retrieving molecular fingerprints from PubChem (Kim et al. 2021) they include some padding in the prefix and suffix, which need to be removed before making use of them.

2.4.2 Important Molecular Descriptors

Shar et al. (2016), after analysing their trained random forest model, found that 2D auto-correlation, topological charge indices and 3D-MoRSE descriptors of compounds were the most essential chemical descriptors in predicting K_i. However, this set of descriptors might not generalise well in other datasets with different labels.

Table 1: Table taken from Lo et al. (2018) showcasing the common chemical descriptors used in QSAR analysis.

Common chemical descriptors for QSAR/QSPR analysis		
Chemical descriptors	**Based on**	**Examples**
Theoretical descriptors		
0D	Molecular formula	Molecular weights, atom counts, bond counts
1D	Chemical graph	Fragment counts, functional group counts
2D	Structural topology	Weiner index, Balaban index, Randic index, BCUTS
3D	Structural geometry	WHIM, autocorrelation, 3D-MORSE, GETAWAY
4D	Chemical conformation	Volsurf, GRID, Raptor
Experimental descriptors		
Hydrophobic parameters	Hydrophobicity	Partition coefficents (logP), hydrohobic substituent constant (π)
Electronic parameters	Electronic properties	Acid dissociation constant, Hammett constant
Steric parameters	Steric properties	Taft steric constant, Charton's constant

2.5 Protein Sequence Descriptors

Structural and physiochemical descriptors calculated from amino acid sequences are widely used in protein-related machine learning research approaches such as the prediction of structural and functional classes and protein-protein interactions (Xiao et al. 2015). The type of descriptors chosen, the numerical representation that encodes the amino acids sequence, is a critical step and can significantly affect the predictive performance of the models a study is trying to train.

Past web servers and stand-alone programs like PROFEAT (Zhang et al. 2017), which currently seems inactive, and PseAAC (Shen and Chou 2008) that tried to calculate these descriptors were often limited in the number of descriptors they were providing, not flexible enough and difficult to integrate into the machine learning pipeline (Xiao et al. 2015).

Protr (Xiao et al. 2015), on the other hand, is a comprehensive package, written in R, that generates various numerical representations of proteins and peptides from amino acid sequences, calculating 8 descriptor groups composed of 22 types of commonly used descriptors, showcased in Table 3, that include roughly 22,700 descriptor values. In addition, this package also allows users to create custom descriptors, calculate similarity scores between pairs of proteins and provides helper functions like loading amino acid sequences from FASTA and PDB files, batch downloading from UniProt (Bateman et al. 2021) and amino acid type sanity check.

It also provides a user-friendly web page (*Protr Web Page* 2022) for people without any programming knowledge, allowing them to calculate the most common descriptors for an amino acids sequence and save them to a CSV file.

2.5.1 Position-Specific Scoring Matrix

Position-Specific Scoring Matrix (PSSM). also known as Position Weight Matrix (PSW) and Position-Specific Weight Matrix (PSWM), is a typical representation of motifs, patterns in biological sequences, and therefore proteins (Jiang et al. 2020; Ranganathan et al. 2019). Motifs are represented as a vector of values, often probabilities, although different representations can also be found for every possible amino acid or residue at each sequence position.

2.5.2 Important Protein Sequence Descriptors

Shar et al. (2016), after analysing their trained random forest model, found that autocorrelation descriptors, amphiphilic pseudo-amino acid composition, and quasi-sequence-order descriptors of protein sequences were found to be the most effective in predicting Ki. Again, this set of descriptors might not generalise well in other datasets with different labels.

Ong et al. (2007) also evaluated the performance of different standard protein sequence descriptor sets, as showcased in Table 2, individually but also in various combinations and concluded that every set on its own is beneficial. However, the predictive performance of models can potentially be enhanced by utilising different combinations of them. This selection could most likely be made through standard feature selection processes.

Table 2: Table taken from Ong et al. (2007) showcasing common protein sequence descriptor sets used in predicting functional protein families.

Sets	Descriptor-sets	No. of descriptors (properties)	No. of components	Type	Physicochemical properties
D1	Amino acid composition	1	20	Sequence composition	
D2	Dipeptide composition	1	400	Sequence composition	
D3	Normalized Moreau – Broto autocorrelation	8	240	Correlation of physicochemical properties	Hydrophobicity scale, average flexibility index, polarizability parameter, free energy of amino acid solution in water, residue accessible surface area, amino acid residue volume, steric parameters, relative mutability
D4	Moran autocorrelation	8	240	Correlation of physicochemical properties	Hydrophobicity scale, average flexibility index, polarizability parameter, free energy of amino acid solution in water, residue accessible surface area, amino acid residue volume, steric parameters, relative mutability
D5	Geary autocorrelation	8	240	Square correlation of physicochemical properties	Hydrophobicity scale, average flexibility index, polarizability parameter, free energy of amino acid solution in water, residue accessible surface area, amino acid residue volume, steric parameters, relative mutability
D6	Descriptors of composition, transition and distribution	21	147	Distribution and variation of physicochemical properties	Hydrophobicity, Van der Waals volume, polarity, polarizability, charge, secondary structures, solvent accessibility
D7	Quasi sequence order	4	160	Combination of sequence composition and correlation of physicochemical	Hydrophobicity, hydrophilicity, polarity, side-chain volume
D8	Pseudo amino acid composition	3	298	Combination of sequence composition and square correlation of physicochemical	Hydrophobicity, hydrophilicity, side chain mass
D9	Combination of amino acid and dipeptide composition	2	420	Combination of sequence compositions	
D10	Combination of all eight sets of descriptors	54	1745	Combination of all sets	

Table 3: Table taken from Xiao et al. (2015) showcasing common protein sequence descriptor sets.

Descriptor groups	Descriptor	Number
Amino acid composition	Amino acid composition	20
	Dipeptide composition	400
	Tripeptide composition	8000
	Normalized Moreau-Broto	240[a]
Autocorrelation	Moran	240[a]
	Geary	240[a]
	Composition	21
CTD	Transition	21
	Distribution	105
Conjoint Triad	Conjoint Triad	343
Quasi-sequence-order	Sequence-order-coupling number	60[a]
	Quasi-sequence-order descriptors	100[a]
Pseudo-amino acid composition	Type I	50[a]
	Type II	80[a]
Proteochemometric descriptors	Principal components analysis (amino acid properties based)	175[b]
	Principal components analysis (2D and 3D molecular descriptors based)	4025[b]
	Factor analysis (amino acid properties based)	175[b]
	Factor analysis (2D and 3D molecular descriptors based)	4025[b]
	Multidimensional scaling (amino acid properties based)	175[b]
	Multidimensional scaling (2D and 3D molecular descriptors based)	4025[b]
	BLOSUM and PAM matrix-derived descriptors	175[b]
PSSM	PSSM profile	–

[a]The number of descriptor values depends on the choice of the number of properties of amino acid and the choice of the parameter
[b]The number of descriptor values depends on the choice of the number of components and the choice of the lag parameter

2.6 QSAR Modelling Process

The general QSAR supervised modelling process suggested by Lo et al. (2018), Shar et al. (2016), and Wang et al. (2020) mention the following modular steps:

- Extracting chemical descriptors from chemical structures, normalising them, and labelling them.
- Performing feature selection to retrieve the descriptors holding the most predictive power and reducing the dimensionality of the feature vector to make model training faster.
- Splitting the dataset into a training and a holdout test set, having in mind data contamination, training data spilling into the test set, allowing the model to achieve better predictive performance not representative of that achieved with real-world data.
- Deciding what evaluation metrics to use. These depend on the type of problem we are working on, classification or regression. Nevertheless, it is considered good practice to report multiple as a single metric can easily lead to wrong conclusions.
- Training machine learning model on the training set while using cross-validation (CV) to optimise the model's hyperparameters and reduce model overfitting. 5-Fold CV seems to be the norm.

- Evaluating the model's predictive performance on the holdout test set.
- Trying to make the model's decisions more interpretable. For example, explaining the 'thought process' behind a prediction can help experts and non-experts alike gain a deeper understanding and confidence in the model's outcomes.
- Testing the robustness of the model using methods like dummy models or permutation tests. This ensures that the model has recognised patterns in the data and makes better than arbitrary decisions.
- Testing the model's predictive performance against that of other state-of-the-art models.

2.7 Ligand-Based Approaches

Ligand-based methods are the most widely used and include QSAR and similarity search-based approaches (Shar et al. 2016). They make use of a drug's chemical and a protein's sequence descriptors without considering the protein's 3D structure (Aparoy et al. 2012).

Such approaches include the classification study conducted by Wang et al. (2020) and the regression study by Shar et al. (2016). In both studies, machine-learning models were trained, optimised and evaluated to predict drug-target interactions, with their performances showcased in Tables 4 and 5 respectively. However, their methodologies varied from one another, using different databases, datasets and tools, clearly expressing the myriad of distinct approaches one can use to solve this problem.

Shar et al. (2016) utilised the *Ki Database* (2022) from the Psychoactive Drug Screening Program (PDSP) (Roth et al. 2016) to retrieve DTIs. Then for each drug and protein combination used PubChem (Kim et al. 2021), ChemSpider (Pence and Williams 2010) and DrugBank (Wishart et al. 2018) to retrieve each drug's structure and UniProt (Bateman et al. 2021) to retrieve each protein's sequence. Molecular descriptors were then calculated using Dragon (Mauri et al. 2006) and protein sequence descriptors using PROFEAT (Zhang et al. 2017). These descriptors were then fed into two models, one based on a support vector machine and another based on a random forest. Their approach is nicely showcased in Figure 1.

Wang et al. (2020) utilised the DTIs datasets of Yamanishi et al. (2008), split into Enzymes, Ion Channels, GPCRs and Nuclear Receptors. Then for each drug and protein combination used a PSSM, as mentioned in Subsection 2.5.1, to convert the protein sequence into numerical descriptors containing biological evolutionary information and then a discrete cosine transform (DCT) algorithm to extract the hidden features and integrate them with the molecular fingerprints extracted from PubChem (Kim et al. 2021). These features were then passed to a rotation forest model. Their approach is showcased in Figure 2.

Both datasets used were relatively small and had less than 10,000 DTI entries, but that did not stop the models trained from achieving excellent predictive performances and even

outperforming state-of-the-art models, possibly highlighting the dataset quality and the processes used.

2.7.1 F-Test

An F-test is a statistical hypothesis test with the null hypothesis stating that there is no statistically significant difference between the predictions of two models, trained on the same dataset, and an alternative hypothesis stating that there is indeed a statistically significant difference between the two. It returns a p-value which represents the probability of the observed predictions occurring due to simply random chance, therefore confirming the null hypothesis. Generally if this p-value is less than or equal to 0.05, we assume that we have a statistically significant result and reject the null hypothesis.

The trained models by Shar et al. (2016) had very similar performance, showcased in Table 4, and the subsequent F-test returned a p-value greater than 0.05, indicating that there was no significant difference between the two models in predicting K_i for the study's particular dataset.

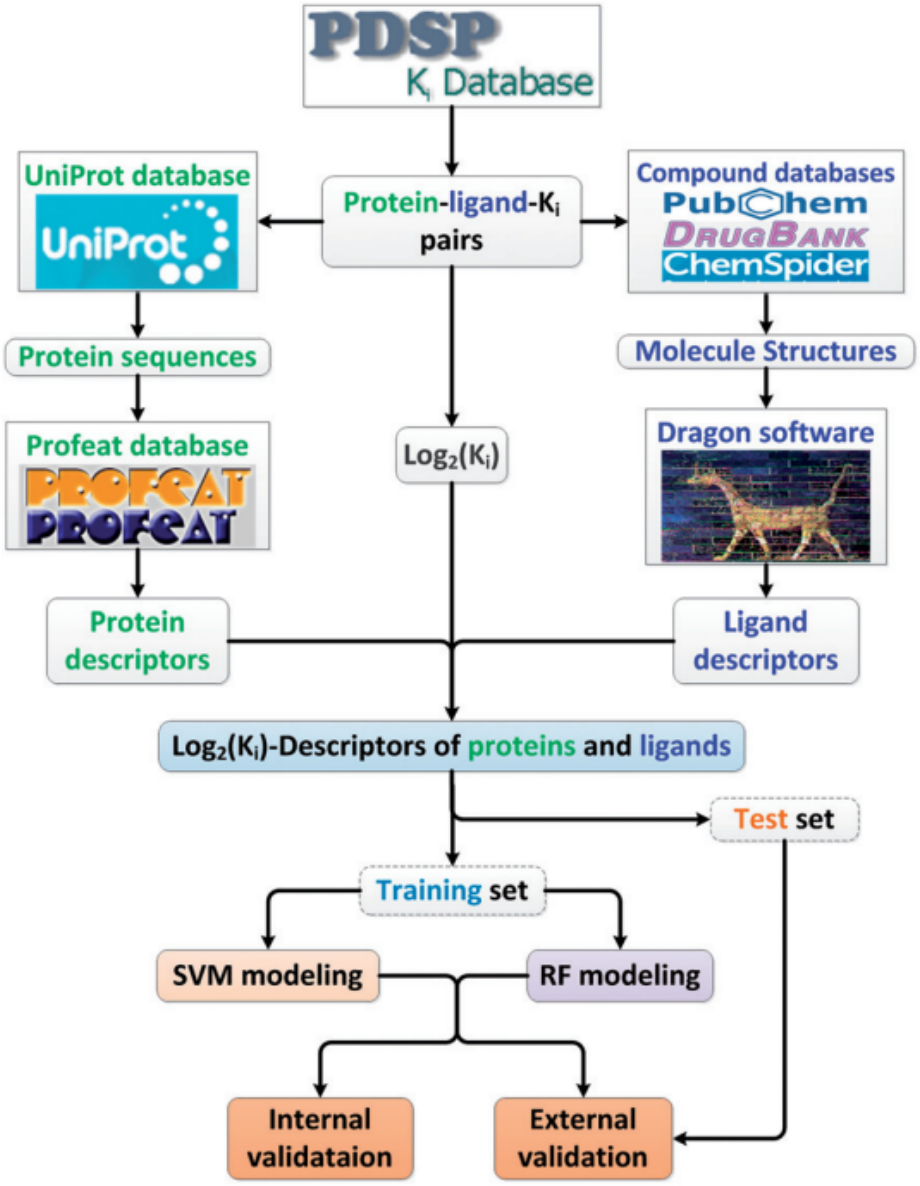


Figure 1: Figure taken from Shar et al. (2016) showcasing their process.

Table 4: Table taken from Shar et al. (2016) showcasing their trained models' performances on their training and test sets.

	R2		MSE	
Model	Internal	External	Internal	External
SVM	0.8596 ± 0.0043	0.6079 ± 0.0117	2.4591 ± 0.0729	7.0487 ± 0.2619
RF	0.8802 ± 0.0025	0.6267 ± 0.0238	2.2855 ± 0.0465	6.5828 ± 0.4075

The numbers of samples in training and test sets are approximately 7958 and 1990, respectively.

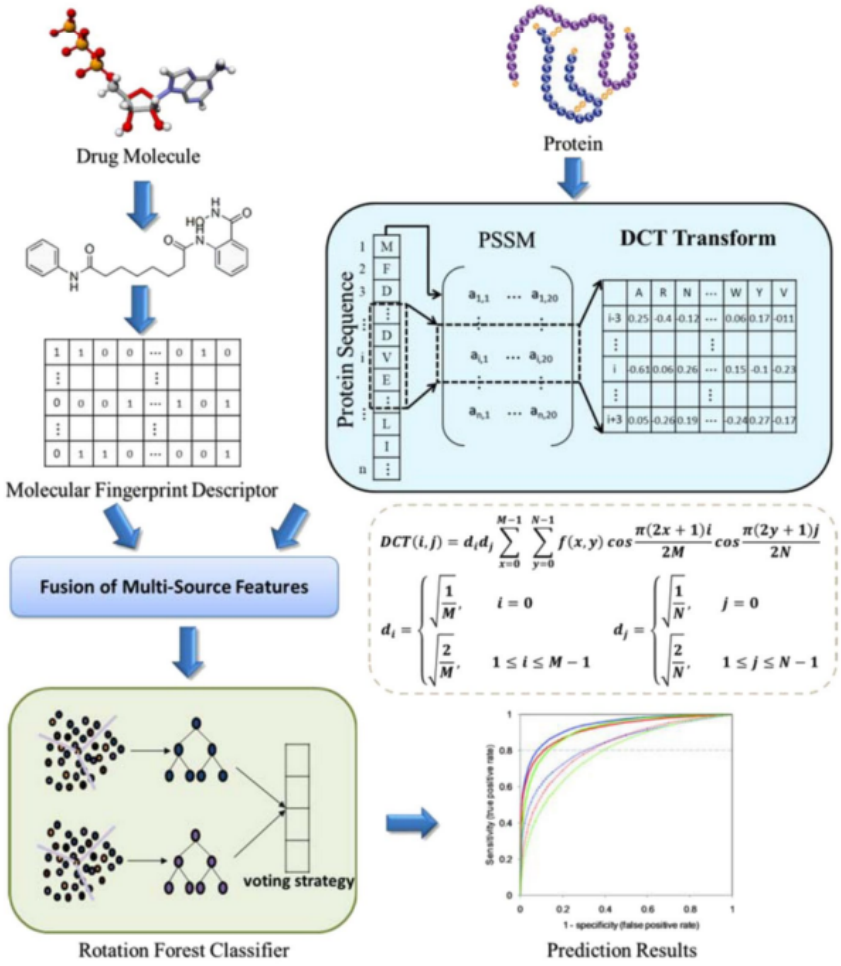


Figure 2: Figure taken from Wang et al. (2020) showcasing their process.

Table 5: Table taken from Wang et al. (2020) showcasing their trained model's performance on their training and test sets.

Dataset	Evaluation Criteria	Accu.	Prec.	Sen.	MCC	AUC
Enzymes	Average	0.9140	0.9202	0.9070	0.8428	0.9088
	Standard Deviation	0.0075	0.0139	0.0225	0.0125	0.0116
Ion Channels	Average	0.8919	0.8928	0.8899	0.7836	0.8925
	Standard Deviation	0.0107	0.0188	0.0166	0.0237	0.0140
GPCRs	Average	0.8724	0.8799	0.8632	0.7454	0.8673
	Standard Deviation	0.0066	0.0337	0.0272	0.0134	0.0181
Nuclear Receptors	Average	0.8111	0.8040	0.8346	0.6328	0.7993
	Standard Deviation	0.0412	0.0944	0.1160	0.0817	0.0593

2.8 Receptor-Based Approaches

Receptor-based approaches such as reverse docking try to predict the preferred conformation and binding strength of a compound to a protein pocket (Shar et al. 2016). They are used when the 3D structure of a protein is mapped and large quantities of data are present. However, such methods are only accurate if the 3D structure of a protein is known, but this could be overcome with predicted 3D protein structures.

One such approach was the study conducted by Jiang et al. (2020), where the structural information of molecules and the predicted structural information of proteins were used, creating two different graphs that were then fed into two graph neural networks (GNN) to obtain their representations. These representations were then concatenated and used to make DTI predictions. Their approach is showcased in Figure 3.

Jiang et al. (2020) utilised the Davis (Davis et al. 2011) and KIBA (He et al. 2017; Tang et al. 2014) datasets, with Davis containing selected entries from the kinase protein family, quantified with K_d values, and KIBA containing entries quantified by a combination of kinase inhibitor bioactivities, Ki, K_d and IC_{50}, called KIBA score.

Graph neural networks have been widely used in various research fields to solve different problems. A graph made of nodes and edges, irrespective of its size, is passed as the input to the GNN, providing a flexible format to extract in-depth information (Jiang et al. 2020).

The drug graph was constructed using its SMILE notation, which describes its unique chemical structure, taking the atom as nodes and the bonds between them as edges. Then the related adjacency matrix was created. A selection of node features based on atoms was also used, shown in Table 6.

The protein graph was constructed by predicting the protein's contact map, with a threshold of 8Å, from its sequence, using a tool called Pconsc4 (Michel et al. n.d.). A contact map is a 2D representation, usually a matrix, of a protein's 3D structure and can be passed directly to a GNN as an adjacency matrix.

More formally, the contact map of a protein sequence with length L is a 'matrix M with L rows and L columns where each element M_{ij} indicates whether the corresponding residue pair, residue i and residue j, are in contact or not', i.e. have a euclidean distance less than a set threshold, usually 6, 8 or 10 Å.

After getting the protein adjacency matrix, the node features were extracted for further processing. Since the graph was constructed with the residues as the nodes, the features should be selected around them. These properties are shown in Table 7, with PSSM being especially important.

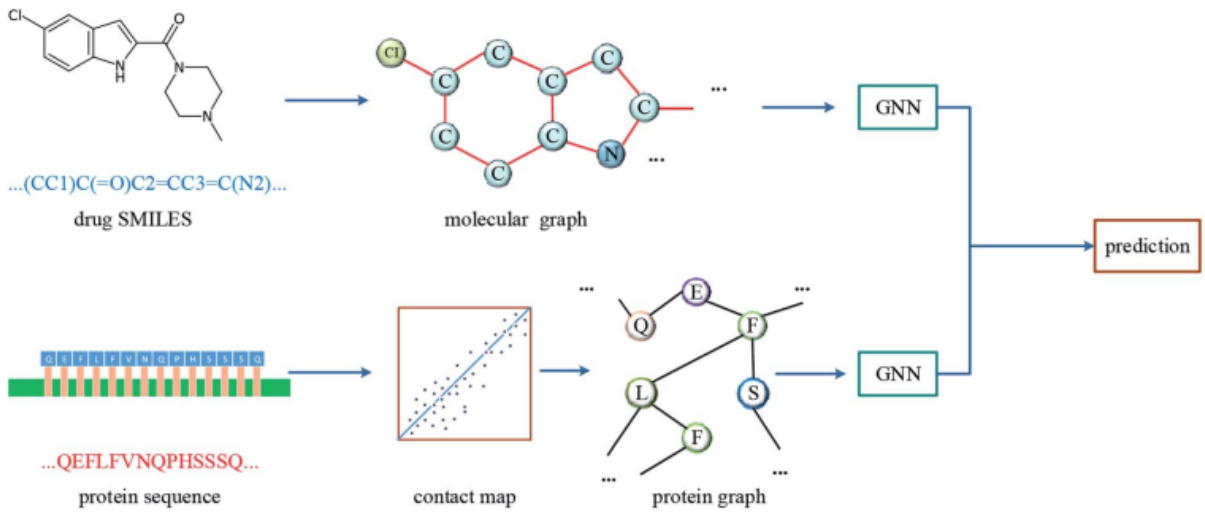


Figure 3: Figure taken from Jiang et al. (2020) showcasing their process.

Table 6: Table taken from Jiang et al. (2020) showcasing the atom node features used.

Number	Feature	Dimension
1	One-hot encoding of the atom element	44
2	One-hot encoding of the degree of the atom in the molecule, which is the number of directly-bonded neighbors (atoms)	11
3	One-hot encoding of the total number of H bound to the atom	11
4	One-hot encoding of the number of implicit H bound to the atom	11
5	Whether the atom is aromatic	1
	All	78

Table 7: Table taken from Jiang et al. (2020) showcasing the residue node features used.

Number	Feature	Dimension
1	One-hot encoding of the residue symbol	21
2	Position-specific scoring matrix (PSSM)	21
3	Whether the residue is aliphatic	1
4	Whether the residue is aromatic	1
5	Whether the residue is polar neutral	1
6	Whether the residue is acidic charged	1
7	Whether the residue is basic charged	1
8	Residue weight	1
9	The negative of the logarithm of the dissociation constant for the –COOH group[64]	1
10	The negative of the logarithm of the dissociation constant for the $-NH_3$ group[64]	1
11	The negative of the logarithm of the dissociation constant for any other group in the molecule[64]	1
12	The pH at the isoelectric point[64]	1
13	Hydrophobicity of residue (pH = 2)[65]	1
14	Hydrophobicity of residue (pH = 7)[66]	1
	All	54

Another interesting study, not for DTI prediction, but for protein function prediction, was that of Gligorijević et al. (2021) where protein sequences and structures were fed into a two-stage architecture model involving a task-agnostic language model and a graph convolutional network (GCN).

The language model was used to extract residue-level features from PDB sequences, and then these together with contact maps with a threshold of 10Å, constructed from the

protein structures, were fed into the GCN. Their approach is showcased in Figure 4, and as we can see, even if it is trying to solve a different problem, it uses a very similar procedure to process a protein's 3D structure.

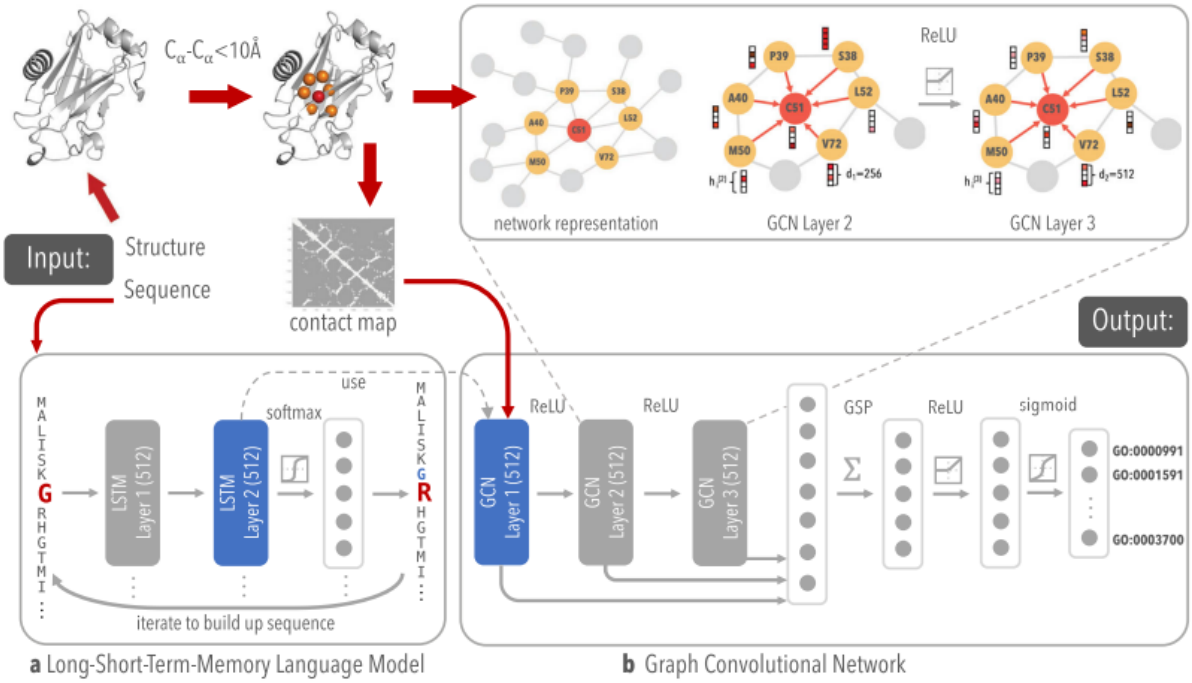


Figure 4: Figure taken from Gligorijević et al. (2021) showcasing their process.

2.9 Valuable Strategies & Concepts Discovered

All studies highlighted that the complicated structure of proteins and molecules make the creation of accurate representations, the features that will be passed into the models, one of the hardest parts of the whole process. This is an active area of research in it of itself in computer-aided medicine. (Jiang et al. 2020)

Both Jiang et al. (2020) and Gligorijević et al. (2021) agree that the most efficient way to process a protein 3D structure is with a GCN as it generalises convolutional operations on efficient graph-like molecular representations. GCNs have also shown vast success in problems such as the prediction of biochemical activity of drugs and prediction of interfaces between pairs of proteins (Gligorijević et al. 2021).

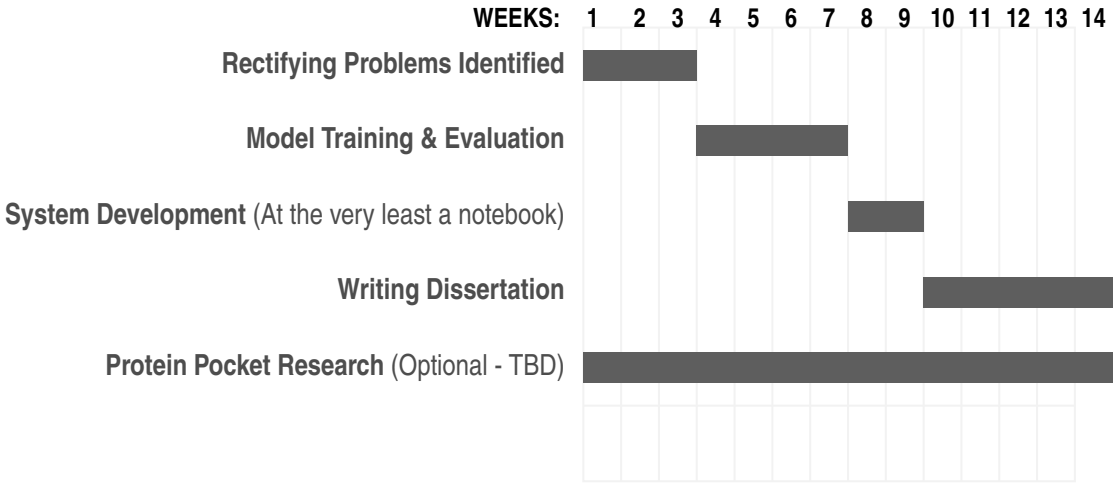
3 Progress

- Research questions laid out
  - Can accurate machine learning models be trained using drug QSAR descriptors and protein sequence QSAR descriptors to predict whether a protein and a drug will bind together? How accurate are such models?
  - Does the addition of protein structure embeddings improve the predictive performance of our models?
  - Are the structure embeddings created from the whole protein structure good enough, or should a more targeted approach involving the structure of protein pockets be used?
  - Can the trained models predict previously unknown DTIs?
- Dataset populated
  - This dataset will be used to train our baseline models and later augmented with the protein structure embeddings.
  - Contains 165,155 DTIs after data cleaning. DTIs were retrieved from PubChem (Kim et al. 2021).
  - Class imbalance 3:1 favouring active DTIs.
  - Contains drug descriptors and protein sequence descriptors. Drug descriptors were retrieved from PubChem (Kim et al. 2021) while protein descriptors were calculated using Protr (Xiao et al. 2015).
  - Feature selection reduced the number of features from 6607 to 1044. Accomplished using Scikit-Learn's recursive feature elimination with 5-fold cross-validation (*Scikit-Learn RFECV* 2022) and a random forest classifier model.
  - Optimised memory usage to 2.02 GBs.
- Baseline Models trained
  - Dataset was split into training and test sets.
  - Optimised using BayesSearchCV (*Scikit-Optimize BayesSearchCV* 2022) using 5-fold cross-validation
  - Achieved very high, questionable, performance on the test set, which led to the identification of some problems, discussed in Subsection 3.1, that will be rectified during the break and the 2nd semester.
- Neural Network ready to be trained
  - Will be used to create protein structure embeddings.
  - Contact map, PSSM and amino acid descriptors were calculated for each unique protein.
  - Amino acid embeddings extracted from UniProt for each unique protein.

3.1 Problems Identified

- Baseline models achieved a very high test set performance. Most likely due to data contamination between the two sets.
  - We are currently capping DTIs at 100 for each protein. This process could add some bias to the dataset if those drugs are sorted in any way. Therefore the DTI retrieval process will be run again with the drugs shuffled to remove any bias.
  - The training and test set will be re-split, ensuring that no drug or protein present in the training set is present in the test set as well.
- Checking the robustness of the models using permutation testing is unfeasible, given their training times.
  - Bootstrap confidence intervals could be used instead.

4 Work Plan



References

AlphaFold Blog (2020). Last accessed: 28-11-2022.
URL: *https://www.deepmind.com/blog/alphafold-a-solution-to-a-50-year-old-grand-challenge-in-biology*

Aparoy, P., Reddy, K. K. and Reddanna, P. (2012), 'Structure and ligand based drug design strategies in the development of novel 5-lox inhibitors', *Current Medicinal Chemistry* 19, 3763.
URL: *https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3480706/*

Bateman, A., Martin, M. J., Orchard, S., Magrane, M., Agivetova, R., Ahmad, S., Alpi, E., Bowler-Barnett, E. H., Britto, R., Bursteinas, B., Bye-A-Jee, H., Coetzee, R., Cukura, A., da Silva, A., Denny, P., Dogan, T., Ebenezer, T. G., Fan, J., Castro, L. G., Garmiri, P., Georghiou, G., Gonzales, L., Hatton-Ellis, E., Hussein, A., Ignatchenko, A., Insana, G., Ishtiaq, R., Jokinen, P., Joshi, V., Jyothi, D., Lock, A., Lopez, R., Luciani, A., Luo, J., Lussi, Y., MacDougall, A., Madeira, F., Mahmoudy, M., Menchi, M., Mishra, A., Moulang, K., Nightingale, A., Oliveira, C. S., Pundir, S., Qi, G., Raj, S., Rice, D., Lopez, M. R., Saidi, R., Sampson, J., Sawford, T., Speretta, E., Turner, E., Tyagi, N., Vasudev, P., Volynkin, V., Warner, K., Watkins, X., Zaru, R., Zellner, H., Bridge, A., Poux, S., Redaschi, N., Aimo, L., Argoud-Puy, G., Auchincloss, A., Axelsen, K., Bansal, P., Baratin, D., Blatter, M. C., Bolleman, J., Boutet, E., Breuza, L., Casals-Casas, C., de Castro, E., Echioukh, K. C., Coudert, E., Cuche, B., Doche, M., Dornevil, D., Estreicher, A., Famiglietti, M. L., Feuermann, M., Gasteiger, E., Gehant, S., Gerritsen, V., Gos, A., Gruaz-Gumowski, N., Hinz, U., Hulo, C., Hyka-Nouspikel, N., Jungo, F., Keller, G., Kerhornou, A., Lara, V., Mercier, P. L., Lieberherr, D., Lombardot, T., Martin, X., Masson, P., Morgat, A., Neto, T. B., Paesano, S., Pedruzzi, I., Pilbout, S., Pourcel, L., Pozzato, M., Pruess, M., Rivoire, C., Sigrist, C., Sonesson, K., Stutz, A., Sundaram, S., Tognolli, M., Verbregue, L., Wu, C. H., Arighi, C. N., Arminski, L., Chen, C., Chen, Y., Garavelli, J. S., Huang, H., Laiho, K., McGarvey, P., Natale, D. A., Ross, K., Vinayaka, C. R., Wang, Q., Wang, Y., Yeh, L. S., Zhang, J., Ruch, P. and Teodoro, D. (2021), 'Uniprot: the universal protein knowledgebase in 2021', *Nucleic Acids Research* 49, D480–D489.
URL: *https://academic.oup.com/nar/article/49/D1/D480/6006196*

Chaudhuri, T. K. and Paul, S. (2006), 'Protein-misfolding diseases and chaperone-based therapeutic approaches', *The FEBS journal* 273, 1331–1349.
URL: *https://pubmed.ncbi.nlm.nih.gov/16689923/*

Colmenarejo, G. (2003), 'In silico prediction of drug-binding strengths to human serum albumin', *Medicinal research reviews* 23, 275–301.
URL: *https://pubmed.ncbi.nlm.nih.gov/12647311/*

Davis, J. L. (2018), 'Pharmacologic principles', *Equine Internal Medicine: Fourth Edition* pp. 79–137.

Davis, M. I., Hunt, J. P., Herrgard, S., Ciceri, P., Wodicka, L. M., Pallares, G., Hocker, M., Treiber, D. K. and Zarrinkar, P. P. (2011), 'Comprehensive analysis of kinase inhibitor

selectivity', *Nature biotechnology* 29, 1046–1051.
URL: *https://pubmed.ncbi.nlm.nih.gov/22037378/*

Estrada, E., Uriarte, E., Molina, E., Simón-Manso, Y. and Milne, G. W. (2006), 'An integrated in silico analysis of drug-binding to human serum albumin', *Journal of Chemical Information and Modeling* 46, 2709–2724.
URL: *https://pubs.acs.org/doi/abs/10.1021/ci600274f*

Fridman, L. (2020), 'Deepmind solves protein folding'. Last accessed: 28-11-2022.
URL: *https://youtu.be/W7wJDJ56c88*

Gligorijević, V., Renfrew, P. D., Kosciolek, T., Leman, J. K., Berenberg, D., Vatanen, T., Chandler, C., Taylor, B. C., Fisk, I. M., Vlamakis, H., Xavier, R. J., Knight, R., Cho, K. and Bonneau, R. (2021), 'Structure-based protein function prediction using graph convolutional networks', *Nature Communications 2021 12:1* 12, 1–14.
URL: *https://www.nature.com/articles/s41467-021-23303-9*

He, T., Heidemeyer, M., Ban, F., Cherkasov, A. and Ester, M. (2017), 'Simboost: a read-across approach for predicting drug-target binding affinities using gradient boosting machines', *Journal of cheminformatics* 9.
URL: *https://pubmed.ncbi.nlm.nih.gov/29086119/*

Hong, H., Xie, Q., Ge, W., Qian, F., Fang, H., Shi, L., Su, Z., Perkins, R. and Tong, W. (2008), 'Mold2, molecular descriptors from 2d structures for chemoinformatics and toxicoinformatics', *Journal of Chemical Information and Modeling* 48, 1337–1344.
URL: *https://pubs.acs.org/doi/abs/10.1021/ci800038f*

Jiang, M., Li, Z., Zhang, S., Wang, S., Wang, X., Yuan, Q. and Wei, Z. (2020), 'Drug–target affinity prediction using graph neural network and contact maps', *RSC Advances* 10, 20701–20712.
URL: *https://pubs.rsc.org/en/content/articlelanding/2020/ra/d0ra02297g*

Jumper, J., Evans, R., Pritzel, A., Green, T., Figurnov, M., Ronneberger, O., Tunyasuvunakool, K., Bates, R., Žídek, A., Potapenko, A., Bridgland, A., Meyer, C., Kohl, S. A., Ballard, A. J., Cowie, A., Romera-Paredes, B., Nikolov, S., Jain, R., Adler, J., Back, T., Petersen, S., Reiman, D., Clancy, E., Zielinski, M., Steinegger, M., Pacholska, M., Berghammer, T., Bodenstein, S., Silver, D., Vinyals, O., Senior, A. W., Kavukcuoglu, K., Kohli, P. and Hassabis, D. (2021), 'Highly accurate protein structure prediction with alphafold', *Nature 2021 596:7873* 596, 583–589.
URL: *https://www.nature.com/articles/s41586-021-03819-2*

Ki Database (2022). Last accessed: 05-12-2022.
URL: *https://pdsp.unc.edu/databases/kidb.php*

Kim, S., Chen, J., Cheng, T., Gindulyte, A., He, J., He, S., Li, Q., Shoemaker, B. A., Thiessen, P. A., Yu, B., Zaslavsky, L., Zhang, J. and Bolton, E. E. (2021), 'Pubchem in 2021: new data content and improved web interfaces', *Nucleic Acids Research* 49, D1388–D1395.
URL: *https://doi.org/10.1093/nar/gkaa971*

Levinthal, C. (1969), 'Levinthal's paradox'.
URL: *https://web.archive.org/web/20110523080407/http://www-miller.ch.cam.ac.uk/levinthal/levinthal.html*

Lo, Y. C., Rensi, S. E., Torng, W. and Altman, R. B. (2018), 'Machine learning in chemoinformatics and drug discovery', *Drug Discovery Today* 23, 1538–1546.

Mauri, A., Consonni, V., Pavan, M. and Todeschini, R. (2006), 'Dragon software: An easy approach to molecular descriptor calculations'.
URL: *shorturl.at/oDEFH*

Michel, M., Hurtado, D. M. N. and Elofsson, A. (n.d.), 'Pconsc4: fast, accurate and hassle-free contact predictions'.
URL: *https://academic.oup.com/bioinformatics/article/35/15/2677/5259184*

Ong, S. A., Lin, H. H., Chen, Y. Z., Li, Z. R. and Cao, Z. (2007), 'Efficacy of different protein descriptors in predicting protein functional families', *BMC Bioinformatics* 8, 1–14.
URL: *https://bmcbioinformatics.biomedcentral.com/articles/10.1186/1471-2105-8-300*

O'Connor, C., Adams, J. U. and Fairman, J. (2010), 'Proteins are responsible for a diverse range of structural and catalytic functions in cells'.
URL: *https://www.nature.com/scitable/ebooks/cell-biology-for-seminars-14760004/122995633/*

Pence, H. E. and Williams, A. (2010), 'Chemspider: An online chemical information resource', *Journal of Chemical Education* 87, 1123–1124.
URL: *https://pubs.acs.org/doi/full/10.1021/ed100697w*

Protr Web Page (2022). Last accessed: 07-12-2022.
URL: *https://nanx.app/protr/*

PubChem Fingerprints (2022). Last accessed: 05-12-2022.
URL: *shorturl.at/msBCH*

Ranganathan, S., Gribskov, M., Nakai, K. and Schönbach, C. (2019), 'Applications, volume 3', *Encyclopedia of Bioinformatics and Computational Biology* 3, 938–952.
URL: *http://www.sciencedirect.com:5070/referencework/9780128114322/encyclopedia-of-bioinformatics-and-computational-biology*

Roth, B. L., Lopez, E., Patel, S. and Kroeze, W. K. (2016), 'The multiplicity of serotonin receptors: Uselessly diverse molecules or an embarrassment of riches?', *http://dx.doi.org/10.1177/107385840000600408* 6, 252–262.
URL: *https://doi.org/10.1177/107385840000600408*

Sachdev, K. and Gupta, M. K. (2019), 'A comprehensive review of feature based methods for drug target interaction prediction', *Journal of Biomedical Informatics* 93, 103159.

Scheife, R. T. (1989), 'Protein binding: what does it mean?', *DICP : the annals of pharmacotherapy* 23.
URL: *https://pubmed.ncbi.nlm.nih.gov/2669380/*

Scikit-Learn RFECV (2022). Last accessed: 12-12-2022.
URL: *https://scikit-learn.org/stable/modules/generated/sklearn.feature_selection.RFECV.html*

Scikit-Optimize BayesSearchCV (2022). Last accessed: 12-12-2022.
URL: *https://scikit-optimize.github.io/stable/modules/generated/skopt.BayesSearchCV.html*

Shar, P. A., Tao, W., Gao, S., Huang, C., Li, B., Zhang, W., Shahen, M., Zheng, C., Bai, Y. and Wang, Y. (2016), 'Pred-binding: large-scale protein–ligand binding affinity prediction', *Journal of Enzyme Inhibition and Medicinal Chemistry* 31, 1443–1450.
URL: *https://pubmed.ncbi.nlm.nih.gov/26888050/*

Shen, H. B. and Chou, K. C. (2008), 'Pseaac: a flexible web server for generating various kinds of protein pseudo amino acid composition', *Analytical biochemistry* 373, 386–388.
URL: *https://pubmed.ncbi.nlm.nih.gov/17976365/*

Tang, J., Szwajda, A., Shakyawar, S., Xu, T., Hintsanen, P., Wennerberg, K. and Aittokallio, T. (2014), 'Making sense of large-scale kinase inhibitor bioactivity data sets: a comparative and integrative analysis', *Journal of chemical information and modeling* 54, 735–743.
URL: *https://pubmed.ncbi.nlm.nih.gov/24521231/*

Torrisi, M., Pollastri, G. and Le, Q. (2020), 'Deep learning methods in protein structure prediction', *Computational and Structural Biotechnology Journal* 18, 1301–1310.

Tunyasuvunakool, K., Adler, J., Wu, Z., Green, T., Zielinski, M., Žídek, A., Bridgland, A., Cowie, A., Meyer, C., Laydon, A., Velankar, S., Kleywegt, G. J., Bateman, A., Evans, R., Pritzel, A., Figurnov, M., Ronneberger, O., Bates, R., Kohl, S. A., Potapenko, A., Ballard, A. J., Romera-Paredes, B., Nikolov, S., Jain, R., Clancy, E., Reiman, D., Petersen, S., Senior, A. W., Kavukcuoglu, K., Birney, E., Kohli, P., Jumper, J. and Hassabis, D. (2021), 'Highly accurate protein structure prediction for the human proteome', *Nature 2021 596:7873* 596, 590–596.
URL: *https://www.nature.com/articles/s41586-021-03828-1*

Vallianatou, T., Lambrinidis, G. and Tsantili-Kakoulidou, A. (2013), 'In silico prediction of human serum albumin binding for drug leads.', *Expert Opinion on Drug Discovery* 8, 583–595.
URL: *https://europepmc.org/article/med/23461733*

Varadi, M., Anyango, S., Deshpande, M., Nair, S., Natassia, C., Yordanova, G., Yuan, D., Stroe, O., Wood, G., Laydon, A., Zídek, A., Green, T., Tunyasuvunakool, K., Petersen, S., Jumper, J., Clancy, E., Green, R., Vora, A., Lutfi, M., Figurnov, M., Cowie, A., Hobbs, N., Kohli, P., Kleywegt, G., Birney, E., Hassabis, D. and Velankar, S. (2022), 'Alphafold protein structure database: massively expanding the structural

coverage of protein-sequence space with high-accuracy models', *Nucleic Acids Research* 50, D439–D444.
URL: *https://doi.org/10.1093/nar/gkab1061*

Wang, L., You, Z. H., Li, L. P., Yan, X. and Zhang, W. (2020), 'Incorporating chemical substructures and protein evolutionary information for inferring drug-target interactions', *Scientific Reports 2020 10:1* 10, 1–11.
URL: *https://www.nature.com/articles/s41598-020-62891-2*

Wishart, D. S., Feunang, Y. D., Guo, A. C., Lo, E. J., Marcu, A., Grant, J. R., Sajed, T., Johnson, D., Li, C., Sayeeda, Z., Assempour, N., Iynkkaran, I., Liu, Y., Maclejewski, A., Gale, N., Wilson, A., Chin, L., Cummings, R., Le, D., Pon, A., Knox, C. and Wilson, M. (2018), 'Drugbank 5.0: A major update to the drugbank database for 2018', *Nucleic Acids Research* 46, D1074–D1082.

Xiao, N., Cao, D.-S., Zhu, M.-F. and Xu, Q.-S. (2015), 'protr/ProtrWeb: R package and web server for generating various numerical representation schemes of protein sequences', *Bioinformatics* 31(11), 1857–1859.
URL: *https://doi.org/10.1093/bioinformatics/btv042*

Yamanishi, Y., Araki, M., Gutteridge, A., Honda, W. and Kanehisa, M. (2008), 'Prediction of drug–target interaction networks from the integration of chemical and genomic spaces', *Bioinformatics* 24, i232–i240.
URL: *https://doi.org/10.1093/bioinformatics/btn162*

Yap, C. W. (2011), 'Padel-descriptor: An open source software to calculate molecular descriptors and fingerprints', *Journal of Computational Chemistry* 32, 1466–1474.
URL: *https://onlinelibrary.wiley.com/doi/full/10.1002/jcc.21707*

Yartsev, A. (2022), 'Protein binding of drugs'. Last accessed: 09-11-2022.
URL: *https://derangedphysiology.com/main/cicm-primary-exam/required-reading/pharmacokinetics/Chapter%20334/protein-binding-drugs*

Zhang, P., Tao, L., Zeng, X., Qin, C., Chen, S. Y., Zhu, F., Yang, S. Y., Li, Z. R., Chen, W. P. and Chen, Y. Z. (2017), 'Profeat update: A protein features web server with added facility to compute network descriptors for studying omics-derived networks', *Journal of Molecular Biology* 429, 416–425.