Discovering cysteine protease covalent inhibitors using deep learning

Author: Carla Feliu Farré, Tutor: Jordi Villa Freixa

[a] *Universitat de Vic - Universitat Central de Catalunya (UVic-UCC),*
[b] *Department of Biociences, UVic-UCC,*

Abstract

Cysteine proteases, are vital enzymes to biological processes, but they can contribute to diseases when dysregulated. They are characterised by having a catalytic diad of cysteine (active site) and histidine. The search for inhibitors of cysteine proteases has become a goal in medical and pharmacological research. Covalent bonds are formed by the interaction of the nucleophilic cysteine and reactive functional group of the ligand. These bonds are more effective than non-covalent ones, resulting in complete inhibition of the target due to their long permanence compared to non-covalent. Machine Learning (ML) has proven highly effective in drug discovery, using various algorithms to predicting novel compound properties. This study explores a Deep Learning (DL), a ML technique. DL techniques include an Artificial Neural Networks (ANNs), which are artificial neuron units capable of transmitting signals to each other through multiple layers to make predictions. This study works with an algorithm based on a directed message passing neural network (D-MPNN) module for molecular extraction and a feed-forward neural network (FNN) for property prediction. First, we select appropriate compound databases, one with known interactions with the target and two others without such interactions, for the purpose of making predictions abut potential cysteine protease inhibitors. The DL model is trained using the prepared datasets containing known interaction activities, employing various training methodologies to explore and enhance the stability and predictive accuracy when applied to other datasets for generating predictions. These results are meticulously analyzed to assess the reliability of the predictions in the context of inhibitor discovery. This research highlights the potential of machine learning, particularly D-MPNN, as a powerful tool in streamlining the drug discovery process by facilitating the identification of promising compounds for covalent inhibition.

Keywords: Cysteine protease, Deep learning, Artificial Neural Networks, Covalent inhibitors, Directed message passing neural network (D-MPNN)

1. Introduction

Proteases, also known as peptid hydrolases, are enzymes capable of catalysing hydrolytic reactions that degrade protein molecules to peptides, and finally to free amino acids. They regulate various enzymatic cascades that form part of metabolic cycles [1]. The role they play is key in carrying out vital biological processes, such as the regulation of various cellular processes as well as differentiation, gene expression and cell death. Proteases constitute an extensive group of enzymes, classified based on various factors, including the specific amino acid residue present at their catalytic site. Commonly recognized subgroups include serine proteases, cysteine proteases, aspartic proteases, and metalloproteases, among others. This diversity is crucial for proteases to fulfill fundamental roles in numerous biological processes [1].

Cysteine proteases (also known as thiol proteases), are characterised by catalysing the breakdown of proteins by cleavage of peptide bonds using a cysteine nucleophile thiol. These proteases are classified into two types according to their location within the organism; Cathepsins, located in the lysosome, and Calpains, located in the cy-

tosol [2]. Having a catalytic dyad or triad, which includes a cysteine (active site), proximately of an histidine. Depending on the type of cysteine proteases, can be Cys-His (Cysteine - Histidine) or Cys-His-Asp (Cysteine - Histidine - Aspartic acid)[3]. (Figure 1)

Although these proteases have a key role in vital processes, when there is overexposure or dysregulation in pathological conditions, they can contribute to the development of many diseases. Search for inhibitors of cystein proteases has become an objective in medical and pharmacological research, in order to develop new therapeutic opportunities [4]. In several studies, it has been shown that cathepsins, are related to tumour progression [5] [6]. These cysteine proteases allow cancer cells to attack nearby tissues, blood and lymphatic vessels and metastasise to peripheral tissues [2] [7].

They represents interesting pharmacological targets due to the reactivity of the cysteine in the active site. The high nucleophilicity of the cysteine thiol under physiological conditions provides an ideal anchoring site for small electrophilic molecules. [8]

The inhibition process is defined by the effectiveness of the binding of a protein to the selected targets. This depends on the molecular reactions between the functional groups of the drug and the active site residues of the protein. The final interactions can be either covalent or non-covalent. Covalent bonds are usually formed by the interaction of certain amino acids, including the nucleophilic cysteine and a reactive functional group of the ligand. They are more effective than non covalent inhibitors due to their long permanence in the active site and high binding affinity resulting in complete inhibition of the target compared to the reversible effect of non-covalent drugs. [9]

The drug development process is a very elaborate process, which can involve a lot of time and resources. It begins with the search for and characterisation of a possible biological target for a specific disease or enzyme, and then the process of creating the most suitable therapeutic compound begins. Before reaching drug development, many molecular properties must be optimised to identify and validate the target, as this is one of the most important steps in drug development. [10] Historically, this search was guided by the calculations and intuition of expert scientists, but with the great advances in new technologies and the increase in resources, it has been possible to apply new methods that are much faster and more reliable. [11]

Machine Learning (ML), particularly Deep Learning (DL), has proven highly effective in diverse chemistry applications. Unlike traditional physical models relying on specific equations, ML uses various algorithms to establish patterns for predicting novel compound properties in chemistry, biology, and physics. This approach has thrived with the increasing quantity and quality of available data, making it an optimal choice for solving problems with abundant data and variables but no known model or formula to relate these variables among themselves with the expected outcome. In drug discovery, the shift toward to large datasets led to the evolution of ML into DL. These is built upon artificial neural networks (ANN), allowing computer systems to make predictions or decisions based on past experiences without explicit programming [12].

Artificial neural networks, are created to simulate a network of model neurons on a computer. Applying algorithms that mimic the processes of real neurons, models are generated that learn to solve many types of problems. The networks are composed of artificial neuron units, capable of transmitting signals to each other. Each reception of a signal in the neurons is associated with a weight that varies as learning progresses, which can increase or decrease the strength of the signal that will then be sent to the other neurons. These signals pass the multilayers of artificial neurons until they reach the final layer, where the result of these connections is the prediction of the algorithm (Figure 2) [13] [12].

The chemprop Architecture is a neural network built for the prediction of molecular properties that includes a directed message passing neural network (D-MPNN) module for molecular feature extraction and a feed-forward neural network (FNN) for property prediction. The Directed Message Passing Neural Network (D-MPNN) model is a variation of the Message Passing Neural Network (MPNN) architecture built upon an existing Graph Convolutional Network (GCN). The primary distinction between D-MPNN and conventional MPNN lies in the nature of the messages transmitted throughout the molecule during message passing [14]. A feed-forward neural network, operates without loops in its node connections, ensuring that information flows strictly in a forward direction without any feedback loops [15]. This model takes

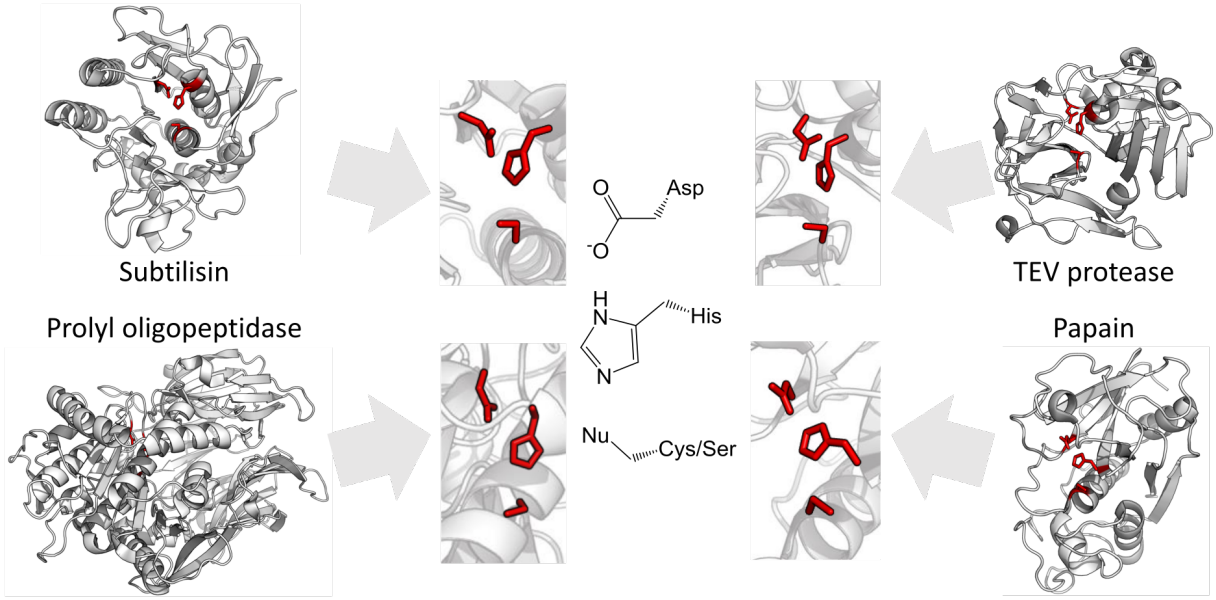


Figure 1: In the figure, can be observed the organization of catalytic involving systeine and serine residues. It illustrates how these triads interact with the active site of enzimes such as subtilisin, prolyl oligopeptidase, TEV protease and papain. These triads have converged to nearly identical arrangements, driven by the mechanistic similarities between cysteine and serine proteolysis mechanisms.

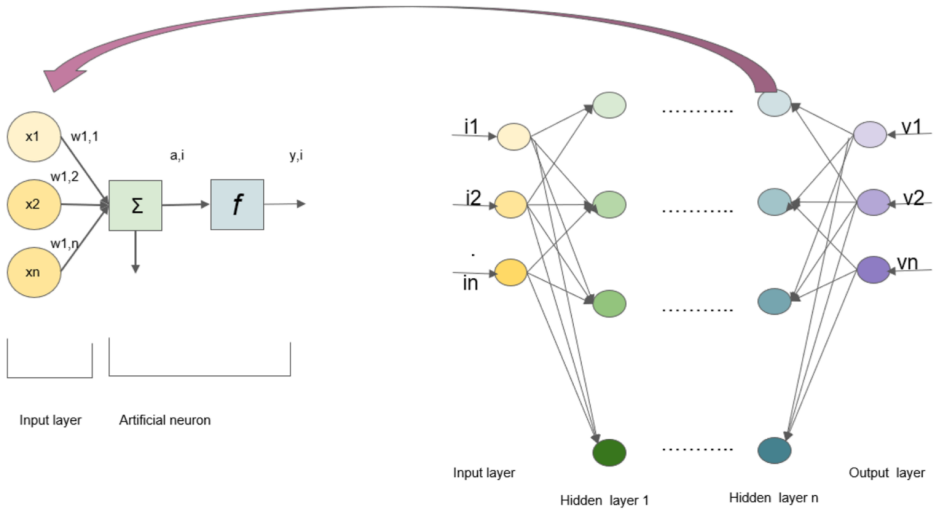


Figure 2: Schematic diagram that illustrates how a single neuron functions within an artificial neural network (ANN). The performance of the ANN depends on factors such as the number of layers it has, the number of neurons, how they communicate with each other, the bias, and the mathematical operations they perform to transform information. The figure shows various input data (i1,i2..) and output data (v1,v2...) which are depicted as small neurons connected to each other.

molecular SMILES as input and extracts all the atomic and bonding characteristics of the molecules to generate a single vector of characteristics which, by entering it into the FNN, can predict the activity of the candidate molecules. [16]

Given the growing interest in cysteine proteases as targets for therapeutic treatments and the availabikity of new machine learning tools for these drug discovery processes, one question is raised in this work.

- Can ANN be used to identify potential covalent targets and molecular patterns for cysteine protease inhibitors in diverse databases?

Specific objectives:

1. Identify a database with known targets of cysteine proteases that includes inhibitory activity, which will serve as the training data of our ML model.
2. Download the database and extract the (Simplified Molecular Input Line Entry Sistem)SMILES information. Prepare the required dataframe structure for working with the ML algorithm, obtaining various features for each molecule and its inhibitory activity (pIC50).
3. Train the chemprop model using various approaches, evaluating which one makes the model more stable for our predictions.
4. Identify suitable databases to obtain the molecules for our predictions. Similar to the training dataframe, construct the structure necessary to make the predictions with our training model.
5. Analyze the data generated from the predictions made by the ML model, checking the correlation of the data and their similarity to understand if the predictions obtained are suitable for our questions raised.
6. Detect patterns of molecules that could be optimal for working in the discovery of cystein protease inhibitors.

2. Methods

2.1. Datasets

In this study, we worked with three different databases to extract the necessary information for our research. Data extraction has been performed depending on each database.

2.1.1. ChEMBL

ChEMBL is a database of bioactive small molecules similar to drugs, containing calculated properties and bioactivities. [17]

Our interest in this database is to obtain information about molecules that can interact with cysteine proteases in order to establish a reference point for interaction activity, allowing us to train our ML model. To access the relevant information to our study, we filtered the ChEMBL database with the term ”cysteine protease” and extract the resulting targets. Once we obtained the cysteine protease targets, we used the python ”requests” module to retrieve information about all the molecules interacting with each of the targets, along with their interaction activity, represented by the pIC50 value. The value was used to train our ML model.

2.1.2. ChemDiv

ChemDiv is a globally recognized research organization in drug discovery solutions. They have identified specific libraries aimed at addressing varius targets, protein domains, cellular processes and more. [18]

From this database, we obtained the ”Cysteine Targeted Covalent Library”, which consists of 39,301 compounts with specific warheads designed to react with cysteine. The objective of using this library is, by applying our trained model, to find possible compounds that exhibit significant interaction activity with cysteine. This will help us identify patterns that can be used to understand what type of molecule could be a good inhibitor.

2.1.3. Zinc

ZINC is a database that houses an extensive collection of commercially available chemical compounds used in virtual screening and drug design research. [19]

Given that this database contains a vast amount of information, with over 230 million compounds, we have randomly selected a portion of this data to conduct our analysis. From the database, we extracted a .uri document containing .url links to different segments of information within the zinc database. Using a Python function, we collected information from three of these files to conduct an analysis of a portion of the zinc data. This has allowed us to work with a more manageable amount of data and focus our objective on applying the algorithm to a comprehensive database.

2.2. Machine learning model

To carry out the study and work with the Machine Learning algorithm, we have used the the Visual Studio Code editor to generate all the necessary code in the python programming lenguage.

Machine Learning is an expansive field where various algorithms and models can be used to achieve multiple objectives. In this study, we developed code that uses a deep learning model incorporating a directed message passing neural network (D-MPNN) to obtain predictions.

We employed the Chemprop architecture, an open-source machine learning framework. This framework includes a D-MPNN module for molecular feature extraction and a feed-forward neural network (FNN) for property prediction. To work with this algorithm, input in SMILES format (representing chemical formulas) is required. The algorithm then transforms this input into a molecular graph structure, where atoms serve as nodes, and bonds as edges.

The D-MPNN module is a critical component responsible for extracting distinctive characteristics from individual atoms, effectively constructing a comprehensive dataset that represents the entire molecule. This process involves a one-way flow of information, starting from the input layer and culminating in the final output layer. The resulting vector, which encapsulates the molecule's essential features, is subsequently utilized in the FNN module to make predictions concerning the activity of the studied molecules. This sequential data flow ensures that relevant information is systematically incorporated into the predictive modeling process, enhancing its accuracy and effectiveness. [16]

A deep learning artificial neural network can solve classification or regression problems. Classification problems involve the use of labels, where data must be categorized into multiple categories. On the other hand, regression problems require the specification of a numerical quantity to classify variables. In this work, we have focused on regression problems, using the IC50 value variable extracted from the ChEMBL database, which classifies molecules on their interaction activity.

2.3. pIC50

The inhibition constant IC50 value is a quantitative measure of a substance's potency in inhibiting a specific biological or chemical function. It signifies the amount of a particular inhibitory substance required to inhibit a specific biological component in 50%. pIC50 is a logarithmic transformation of IC50 values:

$$pIC_{50} = -\log_{10}(IC_{50}) \tag{1}$$

Higher values of pIC50 indicate exponentially more potent inhibitors [20].

2.4. Features

The features are the characteristics to describe data. The following features have been introduced to be constructed the dataframes for inputting into the ANN model:

- Molecular weight (MW): The total sum of the atomic weights.
- LogP: The lipophilicity or hydrophobicity of a molecule.
- Number of hydrogen donors: Number of hydrogen atoms in a molecule that are capable of forming hydrogen bonds as donors.
- Number of hydrogen acceptors: Number of sites in a molecule that can form hydrogen bonds as acceptors.
- Topological Polar Surface Area (TPSA): Quantifies the surface area of a molecule that is polar or can potentially participate in hydrogen bonding.
- Number of rotatable bonds: Number of single bonds in a molecule that are not part of a rigid ring structure.
- Ring counts: Number of distinct ring structures within a molecule.
- Heavy atom counts: Number of non-hydrogen atoms in a molecule.
- Fraction of sp3 hybridized carbons (CSP3): Fraction of carbons in a molecule that have sp3 hybridization.
- Balaban J Index: Information about the degree of branching in a molecule's structure.

2.5. Data preparation and curation

To carry out the process of training machine learning models effectively, it is crucial to have suitable and properly prepared dataset. Therefore, prior preparation is required.

This data structure should have a first column containing the SMILES of the molecules, features describing the molecules, and finally, in a last row, the interaction activity of these molecules, the pIC50 value. Data preparation also includes conducting a thorough analysis of the data structure, which is essential for a deep understanding of the data and, consequently, for conducting a more accurate result analysis.

2.5.1. ANN training

To prepare the training data, we started with a list of all the targets that interact with cysteine proteases and their interactions, from the ChEMBL database. These initial data will provide us a lot of information, but there is an important data preprocessing step before introducing the dataframe into our model.

From this dataframe containing the ligands of the cysteine protease targets, we extracted the list of all the ChEMBL IDs of the molecules that interact with them, along with their pIC50 activity values. Using a ChEMBL API and Python function, we obtained the SMILES for each of the registered molecules.

The data cleaning process is very important, which includes the removal of null values, as they represent a lack of information that may not be accepted or could influence the outcome of the algorithm, and the elimination of duplicated SMILES to avoid duplicated information.

To prepare the training data, we used the chemoinformatics and machine learning software RDKit [21]. From this software, we employed a Python function to iterate over each SMILE, resulting in a total of 10 different features extracted from each compound. Once we had the data prepared, we used a .sh scripts with the necessary commands to work with the chemprop algorithm to perform the data training. In order to ensure and assess the quality of the trained model, we conducted four trainings using four slightly different feature dataframes. These dataframes were analyzed, and ultimately, we determined which one is the most suitable for analyzind our data.

- Train 1: No curated and no normalized data.
- Train 2: Curated and no normalized data.
- Train 3: Curated and z-score normalized data.
- Train 4: Curated and MinMaxScaler normalized data.

2.5.2. Ligand prediction

The first prediction was carried out using data from ChemDiv. To extract this data, we specifically selected the ChemDiv Cysteine Targeted Covalent Library to obtain the initial results.

We downloaded the ChemDiv library from the database using a Python script. The script provided an SDF file containing comprehensive information for each molecule. Utilizing a function that interfaces with RDKit module, we constructed a dataframe. This dataframe included the first column containing SMILES representations, followed by 10 different features for each molecule. Importantly, the pIC50 activity was omitted from the dataframe, as it will be acquired through the ML model.

To prepare the data for input into the algorithm, we performed a curation similar to the one mentioned in the training data preparation. Additionally, we added 200 molecules extracted from the dataframe used to train the model, as these molecules have been entered into the system with real and validated values, providing a reference. Finally, to ensure that the data have the same distribution to the trained model that will be used, we performed a normalization by rescaling all the data to a range of [0-1].

The second prediction was carried out with data from the general Zinc database. It was used to address the goal of finding patterns within a general database, without especificity to cysteine proteases.

From Zinc, we retrieved a file with a .uri extension containing the complete database distributed across various URLs. Utilizing Python's wget module, we programmatically accessed these URLs and downloaded the data, yielding multiple files that collectively constitute the Zinc database. Following this, a Python script was employed to download three of these files. The resulting dataframe from this download comprises SMILES paired with their corresponding ZINC IDs.

Through a function that uses RDKit, we filtered out invalid SMILES to minimize potential errors before preparing the dataframe for the algorithm. Similarly to what

we mentioned earlier for the other dataframes, we used a function with RDKit to generate the appropriate matrix for inputting into the algorithm, with the first column containing the SMILES, followed by 10 features for each one.

Finally, as we did with the first dataframe prediction, we added the same 200 molecules extracted from the training dataframe and normalized data by rescaling all the data to a range of [0-1].

2.6. Results data analysis

First, to analyse the training results, we will evaluate the metrics obtained from each of the training sessions:

- Mean Absolute Error (MAE): MAE is a measure of the average difference between the model's predictions and the actual values. The closer MAE is to 0, the better the model's performance, as it indicates that the predictions are closer to the actual values.
- Mean Squared Error (MSE): MSE takes the square of the differences between predictions and actual values and calculates an average. The lower the value, the better the model's performance.
- Coefficient of Determination (R^2): It is a measure that assesses how well the predictions fit the actual values. This value varies from 0 to 1, where 1 indicates a perfect fit to the model, and 0 indicates no fit at all.
- Root Mean Squared Error (RMSE): RMSE is the square root of MSE and is used to provide an average of the model's error. The lower it is, the better the model's performance.

Once the result of the prediction from the used model was obtained, an analysis was conducted to assess the residual error of the resulting prediction. The 200 reference molecules, which were used both for training and obtaining the predictions, were used to generate a histogram and a regression line. This analysis allows us to visualize the model's error and the reliability of our data.

To analyze and draw conclusions from the results, correlation graphs of the variables were generated for the two predictions made, in order to observe the relationship between all the calculated features. Additionally, a correlation graph of all the features with the pIC50 activity value calculated by the model was created.

Finally, based on the results of the two predictions from the different databases, a selection of the molecules with a higher pIC50 value was made, generating a dataframe for each prediction of the 10 molecules showed stronger interactions. To conclude, a similarity analysis was conducted. To perform the similarity analysis, the Tanimoto coefficient, or Jaccard-Tanimoto index was used, which compares the similarity between two sets of samples. The results it provides are measured on a scale from 0 to 1, with a result closer to 1 indicating a higher degree of similarity among the molecules [22]. This will enable us to observe and analyze if there is a pattern among the molecules with higher activity, providing an indication of whether our model aligns with the objective or not.

3. Results

The different dataframes prepared for training the model originated from a dataset extracted from ChEMBL containing 19,262 molecules. In the case of train 1, a less elaborate data curation was performed, resulting in a dataframe with 19,172 molecules. For train 2, an enhanced data curation was undertaken, beginning with the same database. This process involved the removal of duplicate SMILES, resulting in a reduced dataset containing 11,703 unique molecules. Train 3 and Train 4 were then derived from the dataset generated in Train 2, with data normalization applied using the z-score and the MinMaxScaler methods, respectively. (Table 1)

Train	Initial	Curated	Norm.
Train 1	19,262	19,172	No
Train 2	19,262	11,703	No
Train 3	19,262	11,703	Z-Score
Train 4	19,262	11,703	MinMaxScale

Table 1: Training sets characteristics

In both Table 2 and Figures 3 and 4, the evaluation metrics of the training models are presented, where we can see the different results:

- Mean Absolute Error (MAE): As can be seen in the results of the mentioned figures, training model 4 shows the best performance in this aspect, while training model 1 performs the worst.

- Mean Squared Error (MSE): It can be observed that training model 4 continues to have the best performance, while training model 1 performs the worst.
- Coefficient of Determination (R^2): That the model that approaches 1 the closest and therefore fits better is training model 2, followed by training models 4 and 3, which have the same R^2 value with very little difference from 2. Finally, training model 1 is the furthest from a perfect fit.
- Root Mean Squared Error (RMSE): Training model 1 stands out by a significant margin as having the best performance.

The results obtained from the 4 training runs conducted with the ML model algorithm show that the model with the highest possibility of error is training model 1, prepared with less elaborate data curation and with non-normalized data, while the one with the most reliable prediction is training model 4, with stricter data curation and a MinMaxScaler normalization. With the training results analysis, it is decided to use Train 4 to obtain the predictions for our study.

Train	MSE	RMSE	R2	MAE
Train 1	0.77	0.88	0.52	0.67
Train 2	0.53	0.73	0.68	0.55
Train 3	0.35	0.59	0.67	0.44
Train 4	0.01	0.11	0.67	0.08

Table 2: Results of the evaluation metrics for the training sessions conducted with the model

Once the model predictions have been made, the results of the 200 molecules used for training, for which we already had information about inhibitory activity, were compared with the same 200 molecules resulting from the prediction. In the case of these ones, the algorithm generates a prediction of the activity because when inputting the dataset into the model for prediction, this information was not included.

When analyzing the results of the histogram (Figure 5), it is noted that the distribution of the data is not perfect. Although the figure could suggest a pattern that indicate a good distribution, there are many intermediate values that generate multiple peaks that don't follow the same

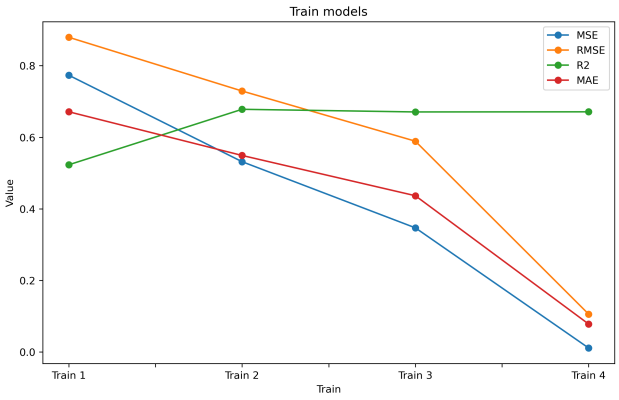


Figure 3: Line chart of training set results metrics

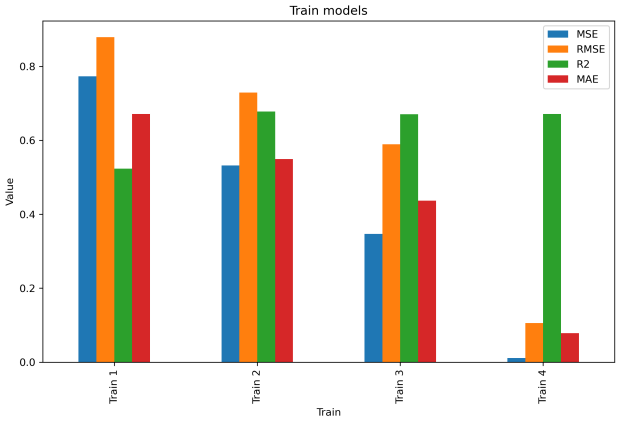


Figure 4: Barplot of training set results metrics

trend. This can also be observed with the regression line (Figure 6), which does not fit the central line of the actual values, indicating that it does not fully adapt to the real values. Furthermore, it can be observed how this dispersion in the distribution increases with higher values.

To analyze the results obtained from the prediction, correlation graphs between variables have been generated, as well as correlation graphs for each of the calculated variables with the inhibitory activity value. In order to compare these results and draw a conclusion based on real data, the same analysis has also been performed on the data used to train the model used for the predictions. These graphs will provide us information about patterns of data correlation, considering that a value of 1 will indi-

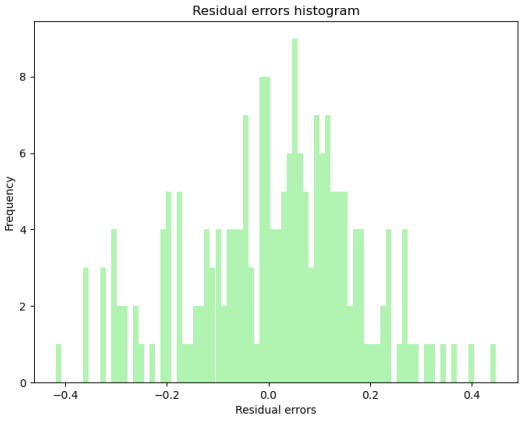


Figure 5: Histogram of predicted residual errors, comparing 200 molecules from real values against predicted values

Figure 6: Linear regression of predicted residual errors, comparing 200 molecules from real values against predicted values

cate a perfect positive correlation between variables, and a value -1 a perfect negative correlation, meaning that when one variable increases, the other decreases.

First of all, to analyze the dataset used to train the model (ChEMBL database), we can observe in Figure 7 that the variables most correlated with each other are Molecular Weight (MW) with Heavy Atom Counts and Hydrogen Bond Donor (H-bond donor) with Topological Polar Surface Area (TPSA) with correlation values of 0,99 and 0,82 respectively. Looking at Figure 8, we see that the maximum correlation value is 0,13. This indicates that, in general, the correlation of the variables with what interests us is not very high. The variables most correlated with the value of inhibitory activity within the given values, are Rotatable Bonds, with a value of 0,13, and Balaban J Index, with a value of 0,11.

The results for the dataset of the first prediction (ChemDiv database), show that the most correlated variables among them, as see in Figure 9, are MW with Heavy Atom Counts, H-bond Donor with Rotatable Bonds and H-bond Acceptor with TPSA, with correlation values of 0,97 , 0,85 and 0,85 respectively. Regarding the relationship of each variable with the activity value (Figure 10), the ones with the highest correlation are H-bond acceptor and Rotatable Bonds with correlation values of 0,34 and 0,3.

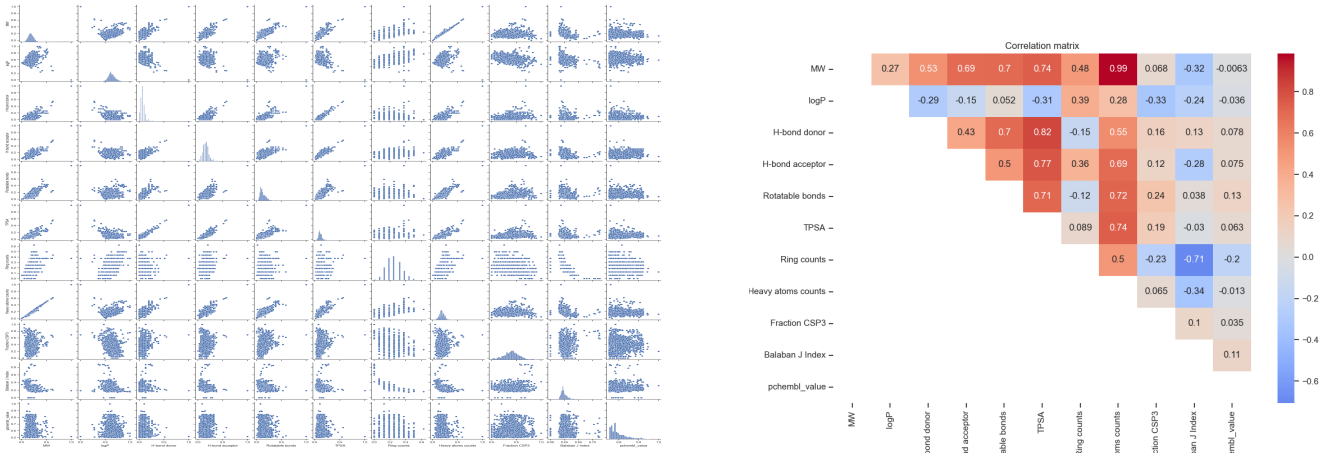
For the dataset of the second train (Zinc database), the variables most related to each other, as seen in Figure 13, are MW with Heavy Atom Counts and TPSA with Heavy Atom Counts, with correlation values of 0,98 and 0,9 respectively. The variables that are most related to the activity value (Figure 11) are H-bond acceptor and TPSA with values 0,41 and 0,34.

We conclude the analysis with a similarity test of the top 10 molecules selected from each of the datasets resulting from the predictions (Table 16). Using the Tanimoto coefficient, a score of 1 exactly similarity, while 0 indicates no similarity. It is important to note that in all cases, very low similarity values are observed, with a maximum value of 0.2.

Upon examining the images of the top 10 molecules from the first set of predictions (Figure14), the top 10 molecules from the second set of predictions (Figure 15), and referring to Table 16, we observe that the molecules with the highest similarity values within the results are M1 from Figure 14 with M5 and M7 from Figure 15, both having a value of 0.207. The next best similarity results at value 0.2 , are for M4 on Figure 14 also with M5 and M7 on Figure 15.

4. Discussion

When analyzing the results of the different Chemprop trainings, we have used training 4 to make predictions for



(a) (b)

Figure 7: Correlation analysis from the dataset used to train the model (ChEMBL). (a) Pariplot correlation matrix. (b)Correlation matrix heatmap

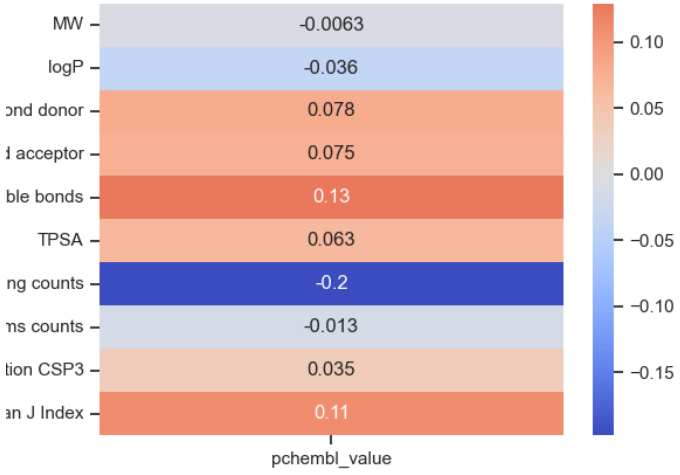
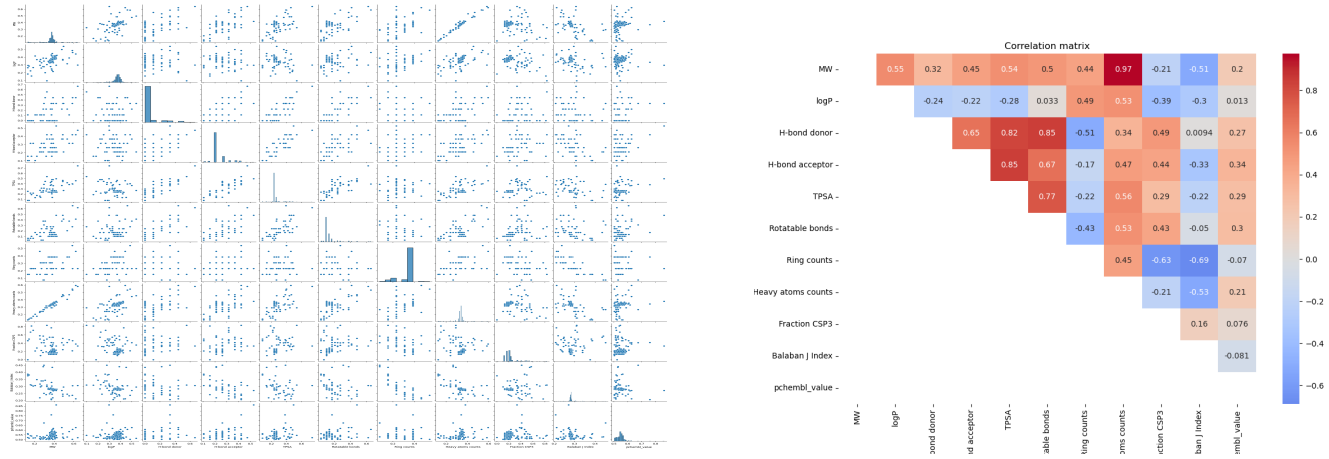


Figure 8: Correlation heatmap of variables with the pIC50 value from the dataset used to train the model



(a) (b)

Figure 9: Correlation analysis of the features to determine if there is any correlation pattern among them. Results of the first predicted model (ChemDiv). (a)Pariplot correlation matrix (b)Correlation matrix heatmap

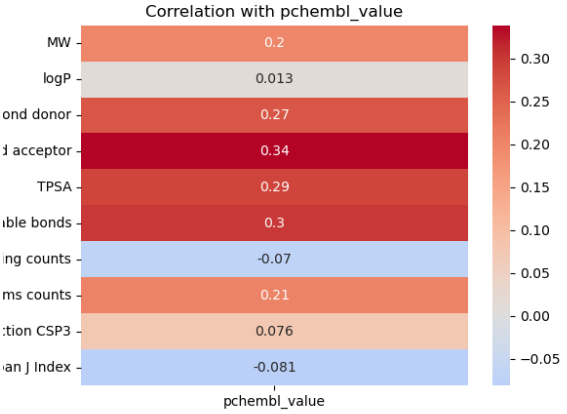


Figure 10: Correlation heatmap of variables with the pIC50 value from the dataset used to the first predict model (ChemDiv)

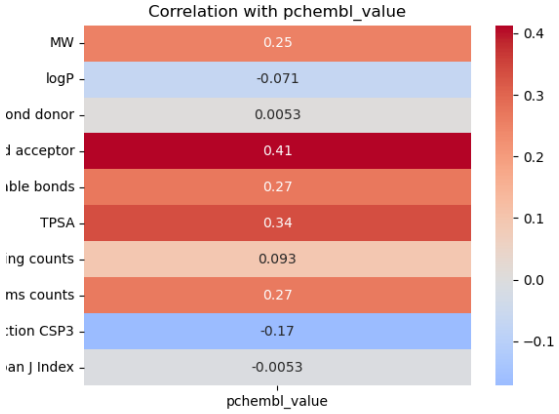
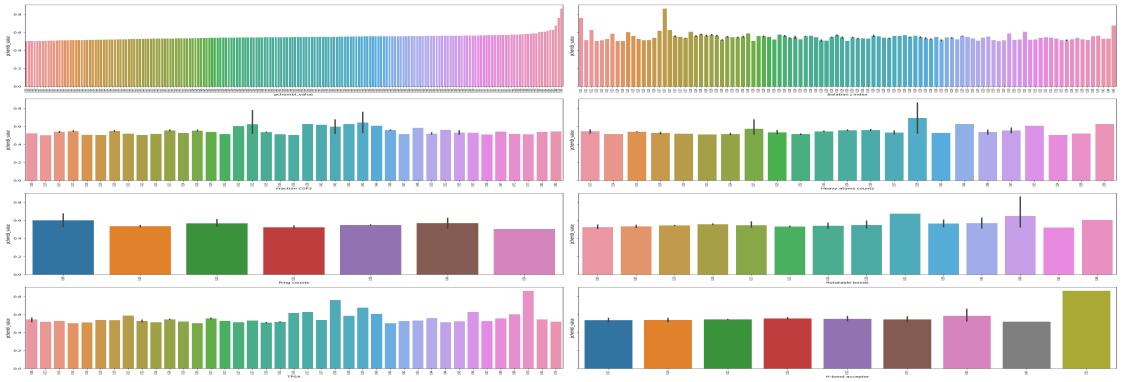


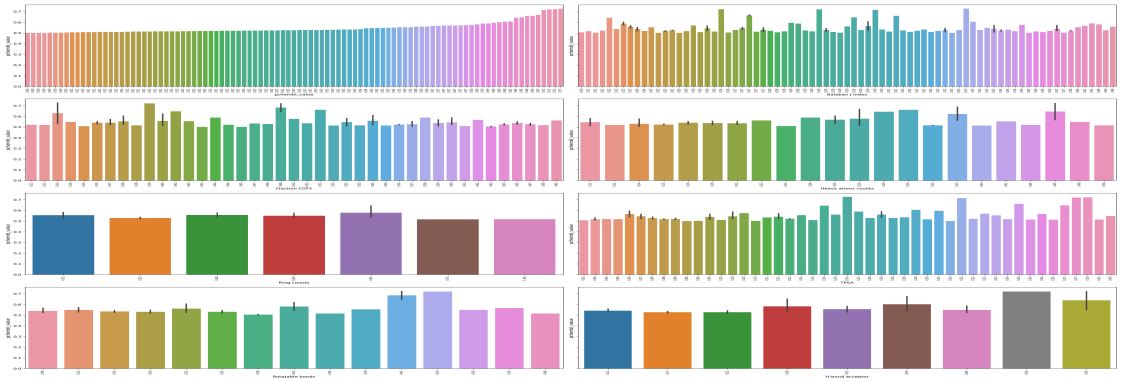
Figure 11: Correlation heatmap of variables with the pIC50 value from the dataset used to the second predict model (Zinc)

our study. We considered it to be the most appropriate and well-fitted, as both the MSE, RMSE, and MAE values are the most favorable in this training. Although the highest R^2 value was obtained in training 2 with a value of 0.68, the difference between this result and training 4 with an R^2 of 0.67 is negligible, so we have decided to proceed with training 4 as the model to predict our results.

The most desviated result from a well-fitted model was observed in training 1, which can be attributed to the fact that this training was performed with a dataframe containing very raw data, with many repeated SMILES, which could have affected the machine learning model. Therefore, we can see how the importance of good data cleaning is reflected in the training results.



(a)



(b)

Figure 12: Barplot matrix analysis from predicted results of each feature among pIC50. Each subgraph bar in figures A and B, displays the dataset for each feacture according to its pIC50. It provides information on how each feature relates to pIC50. (a)Fist predict, Chemdiv (b)Second predict, Zinc

(a) Pairplot correlation matrix (b) Correlation matrix

Figure 13: Correlation analysis of the features to determine if there is any correlation pattern among them. Results of the second predicted model (Zinc). (a)Pariplot correlation matrix (b)Correlation matrix heatmap

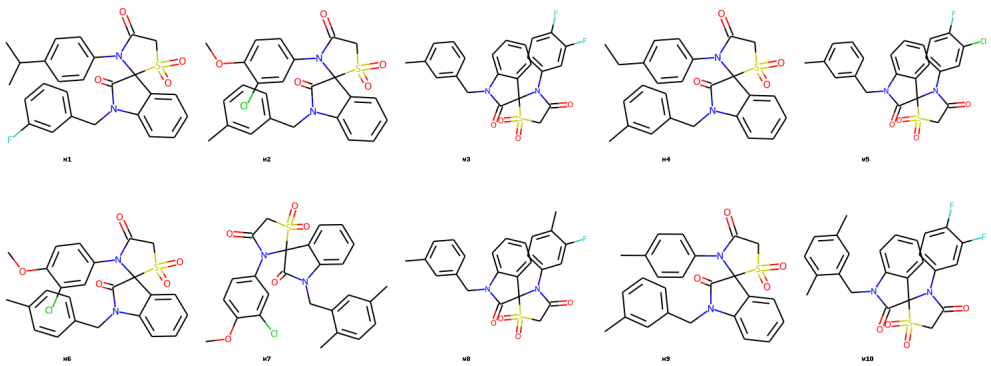


Figure 14: The 10 molecules from the first predicted results (ChemDiv) with a higher activity value (pIC50)

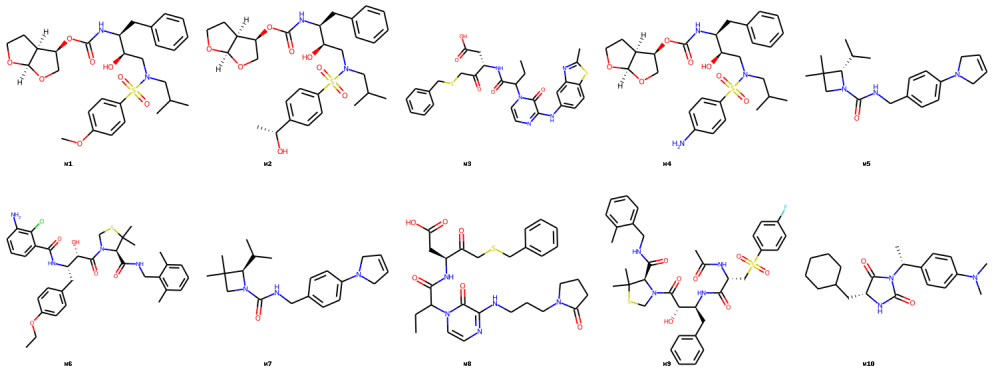


Figure 15: The 10 molecules from the second predicted results (Zinc) with a higher activity value (pIC50)

Upon analyzing the prediction model versus the trained model, we observe that for the data of greatest interest to us, those with higher pIC50 values, the regression line shows an increase in dispersion (see Figure 6. This indicates that our model is not entirely reliable, especially for results where pIC50 is crucial for inhibitor discovery.

Regarding the features analyzed in the prediction results based on the correlation of variables, in all three cases analyzed (see Figures 7, 9, 13), we can see a strong correlation between the MW and Heavy Atom Count features. This may be due to the fact that heavier molecules tend to have more heavy atoms.

As for the relationship between variables and the pIC50 value, the analysis suggests that in none of the three datasets is there a strong correlation with this variable, particularly in the training dataset. This indicates that a strong interaction activity is not significantly correlated with these chemical characteristics. Nonetheless, in Figure 12 of the predictions, we can observe how it supports the results of the correlation heatmap, where it is evident that the variables H-bond acceptor, TPSA, Heavy Atom Counts, and Rotatable Bonds are the ones with the highest values.

Finally, regarding the similarity analysis, we have been unable to obtain conclusive results because the maximum similarity values are very low and, therefore, unreliable. However, we have observed that molecules M5 and M7 from the second training set, based on zinc data, are highly similar. We have confirmed this initially through Figure X and secondly by noticing that they share the same similarity values with a specific molecule from the first Chemdiv training set. This could serve as a starting point for further research to understand the characteristics of these molecules.

We have also noticed some similarity among all molecules in the Chemdiv database, as they interact with very little difference in similarity compared to molecules in the zinc database. This is attributed to the fact that all these molecules originate from a cisteine protease warheads database, hence sharing similar features.

	index	SMILES_CHEMDIV	SMILES_ZINC	Similarity	Mol_ChemDiv	Mol_Zinc
0	4	CC(C)c1ccc(N2C(=O)CS(=O)(=O)C23C(=O)N(Cc2cccc(F)c2)c2ccccc23)cc1	CC(C)[C@H]1N(C(=O)Nc2ccc(N3CC=CC3)cc2)CC1(C)C	0.2073170731707317	M1	M5
1	6	CC(C)c1ccc(N2C(=O)CS(=O)(=O)C23C(=O)N(Cc2cccc(F)c2)c2ccccc23)cc1	CC(C)[C@H]1N(C(=O)NCc2ccc(N3CC=CC3)cc2)CC1(C)C	0.2073170731707317	M1	M7
2	34	CCc1ccc(N2C(=O)CS(=O)(=O)C23C(=O)N(Cc2ccccc(C)c2)c2ccccc23)cc1	CC(C)[C@H]1N(C(=O)Nc2ccc(N3CC=CC3)cc2)CC1(C)C	0.2	M4	M5
3	36	CCc1ccc(N2C(=O)CS(=O)(=O)C23C(=O)N(Cc2ccccc(C)c2)c2ccccc23)cc1	CC(C)[C@H]1N(C(=O)NCc2ccc(N3CC=CC3)cc2)CC1(C)C	0.2	M4	M7
4	84	Cc1ccc(N2C(=O)CS(=O)(=O)C23C(=O)N(Cc2ccccc(C)c2)c2ccccc23)cc1	CC(C)[C@H]1N(C(=O)Nc2ccc(N3CC=CC3)cc2)CC1(C)C	0.19230769230769232	M9	M5
5	86	Cc1ccc(N2C(=O)CS(=O)(=O)C23C(=O)N(Cc2ccccc(C)c2)c2ccccc23)cc1	CC(C)[C@H]1N(C(=O)NCc2ccc(N3CC=CC3)cc2)CC1(C)C	0.19230769230769232	M9	M7
6	54	COc1ccc(N2C(=O)CS(=O)(=O)C23C(=O)N(Cc2ccccc(C)cc2)c2ccccc23)cc1Cl	CC(C)[C@H]1N(C(=O)Nc2ccc(N3CC=CC3)cc2)CC1(C)C	0.19047619047619047	M6	M5
7	56	COc1ccc(N2C(=O)CS(=O)(=O)C23C(=O)N(Cc2ccccc(C)cc2)c2ccccc23)cc1Cl	CC(C)[C@H]1N(C(=O)NCc2ccc(N3CC=CC3)cc2)CC1(C)C	0.19047619047619047	M6	M7
8	24	Cc1cccc(CN2C(=O)C3(c4ccccc42)N(c2ccc(F)c(F)c2)C(=O)CS3(=O)=O)c1	CC(C)[C@H]1N(C(=O)Nc2ccc(N3CC=CC3)cc2)CC1(C)C	0.18072289156626506	M3	M5
9	26	Cc1cccc(CN2C(=O)C3(c4ccccc42)N(c2ccc(F)c(F)c2)C(=O)CS3(=O)=O)c1	CC(C)[C@H]1N(C(=O)NCc2ccc(N3CC=CC3)cc2)CC1(C)C	0.18072289156626506	M3	M7
10	74	Cc1cccc(CN2C(=O)C3(c4ccccc42)N(c2ccc(C)c(F)c2)C(=O)CS3(=O)=O)c1	CC(C)[C@@H]1N(C(=O)Nc2ccc(N3CC=CC3)cc2)CC1(C)C	0.17857142857142858	M8	M5
11	76	Cc1cccc(CN2C(=O)C3(c4ccccc42)N(c2ccc(C)c(F)c2)C(=O)CS3(=O)=O)c1	CC(C)[C@H]1N(C(=O)NCc2ccc(N3CC=CC3)cc2)CC1(C)C	0.17857142857142858	M8	M7
12	44	Cc1cccc(CN2C(=O)C3(c4ccccc42)N(c2ccc(F)c(Cl)c2)C(=O)CS3(=O)=O)c1	CC(C)[C@H]1N(C(=O)Nc2ccc(N3CC=CC3)cc2)CC1(C)C	0.1744186046511628	M5	M5
13	46	Cc1cccc(CN2C(=O)C3(c4ccccc42)N(c2ccc(F)c(Cl)c2)C(=O)CS3(=O)=O)c1	CC(C)[C@H]1N(C(=O)NCc2ccc(N3CC=CC3)cc2)CC1(C)C	0.1744186046511628	M5	M7
14	14	COc1ccc(N2C(=O)CS(=O)(=O)C23C(=O)N(Cc2ccccc(C)c2)c2ccccc23)cc1Cl	CC(C)[C@H]1N(C(=O)Nc2ccc(N3CC=CC3)cc2)CC1(C)C	0.17045454545454544	M2	M5
15	16	COc1ccc(N2C(=O)CS(=O)(=O)C23C(=O)N(Cc2ccccc(C)c2)c2ccccc23)cc1Cl	CC(C)[C@H]1N(C(=O)NCc2ccc(N3CC=CC3)cc2)CC1(C)C	0.17045454545454544	M2	M7
16	8	CC(C)c1ccc(N2C(=O)CS(=O)(=O)C23C(=O)N(Cc2cccc(F)c2)c2ccccc23)cc1	CC(=O)N[C@H](CS(=O)(=O)c1ccc(F)cc1)C(=O)N[C@@H](Cc1ccccc1)[C@H](O)C(=O)N1CSC(C)(C)[C@H]1C(=O)NCc1ccccc1C	0.16521739130434782	M1	M9
17	78	Cc1cccc(CN2C(=O)C3(c4ccccc42)N(c2ccc(C)c(F)c2)C(=O)CS3(=O)=O)c1	CC(=O)N[C@H](CS(=O)(=O)c1ccc(F)cc1)C(=O)N[C@@H](Cc1ccccc1)[C@H](O)C(=O)N1CSC(C)(C)[C@H]1C(=O)NCc1ccccc1C	0.16521739130434782	M8	M9
18	98	Cc1ccc(C)c(CN2C(=O)C3(c4ccccc42)N(c2ccc(F)c(F)c2)C(=O)CS3(=O)=O)c1	CC(=O)N[C@H](CS(=O)(=O)c1ccc(F)cc1)C(=O)N[C@@H](Cc1ccccc1)[C@H](O)C(=O)N1CSC(C)(C)[C@H]1C(=O)NCc1ccccc1C	0.16521739130434782	M10	M9
19	94	Cc1ccc(C)c(CN2C(=O)C3(c4ccccc42)N(c2ccc(F)c(F)c2)C(=O)CS3(=O)=O)c1	CC(C)[C@H]1N(C(=O)Nc2ccc(N3CC=CC3)cc2)CC1(C)C	0.16470588235294117	M10	M5

Figure 16: This table displays the results of the similarity analysis between the top molecules from each predicted dataset

5. Conclusions

After conducting an analysis and training algorithms of ANN that allowed us to identify patterns and improve the stability of the prediction model. However, even though we have made progress in this regard, the results have not yet reached the necessary optimization to ensure full data reliability. This suggests the importance of conducting deeper and more comprehensive training to achieve a more precise adaptation of the model and enhance its predictive capacity.

Regarding the search for patterns through the correlation between molecular features, our results indicate that we have not found a strong enough correlation between the analyzed variables and inhibitory activity. This finding leads us to consider the need to expand our research, exploring a broader set of molecular features to delve deeper into the relationship between these features and inhibitory activity, which could shed light on new markers and predictors.

Finally, despite the overall unfruitful search for similarities, we highlight the interesting discovery of two highly similar molecules in the zinc database, both with elevated pIC50 values. This discovery could serve as a starting point for further investigations exploring the relationship between these molecules and their ability to inhibit cysteine proteases, thus opening new avenues of study in this field.

In summary, this study highlights the potential of machine learning, particularly artificial neural networks (ANNs), as a powerful tool for streamlining the drug discovery process by facilitating the identification of promising compounds for covalent inhibition. Although the results may not have yielded effective inhibitors in this specific context, this research underscores the value of ML technology in exploring new avenues and approaches in the quest for innovative therapies.

Appendix A. Acknowledgments

I would like to express my gratitude to my project supervisor, Jordi Villa, for guiding me throughout the project, and to my partner and family for their unwavering support during the entire process.

Appendix B. Data availability

The code used for this study is available on the following GitHub repository: https://github.com/carlafeliu/TFM

Appendix C. References

References

[1] Oscar L. Ramos and F. Xavier Malcata. *Food-grade enzymes*. Elsevier, 1 2019.

[2] Satya P. Gupta and Sayan Dutta Gupta. *Cancer-leading proteases: An introduction*. Elsevier, 1 2020.

[3] D. W. Nicholson and G. Melino. *Caspases and Cell Death*, pages 388–396. Elsevier Inc., 2 2013.

[4] Johanna A Joyce, Amos Baruch, Kareem Chehade, Nicole Meyer-Morse, Enrico Giraudo, Fong-Ying Tsai, Doron C Greenbaum, Jeffrey H Hager, Matthew Bogyo, and Douglas Hanahan. Cathepsin cysteine proteases are effectors of invasive growth and angiogenesis during multistage tumorigenesis.

[5] Izabela Berdowska. Cysteine proteases as disease markers, 4 2004.

[6] Mona Mostafa Mohamed and Bonnie F. Sloane. Cysteine cathepsins: Multifunctional enzymes in cancer. volume 6, pages 764–775, 10 2006.

[7] Vasilena Gocheva, Wei Zeng, Danxia Ke, David Klimstra, Thomas Reinheckel, Christoph Peters, Douglas Hanahan, and Johanna A. Joyce. Distinct roles for cysteine cathepsin genes in multistage tumorigenesis. *Genes and Development*, 20:543–556, 3 2006.

[8] Aaron J. Maurais and Eranthie Weerapana. Reactive-cysteine profiling for drug discovery, 6 2019.

[9] Aimen Aljoundi, Imane Bjij, Ahmed El Rashedy, and Mahmoud E.S. Soliman. Covalent versus non-covalent enzyme inhibition: Which route should we take? a justification of the good and bad from molecular modelling perspective, 4 2020.

[10] J. P. Hughes, S. S. Rees, S. B. Kalindjian, and K. L. Philpott. Principles of early drug discovery, 3 2011.

[11] Patrick Hop, Brandon Allgood, and Jessen Yu. Geometric deep learning autonomously learns chemical features that outperform those engineered by domain experts. *Molecular Pharmaceutics*, 15:4371–4377, 10 2018.

[12] Sagorika Nag, Anurag T.K. Baidya, Abhimanyu Mandal, Alen T. Mathew, Bhanuranjan Das, Bharti Devi, and Rajnish Kumar. Deep learning tools for advancing drug discovery and development, 5 2022.

[13] Alan Talevi, Juan Francisco Morales, Gregory Hather, Jagdeep T. Podichetty, Sarah Kim, Peter C. Bloomingdale, Samuel Kim, Jackson Burton, Joshua D. Brown, Almut G. Winterstein, Stephan Schmidt, Jensen Kael White, and Daniela J. Conrado. Machine learning in drug discovery and development part 1: A primer. *CPT: Pharmacometrics and Systems Pharmacology*, 9:129–142, 3 2020.

[14] Kevin Yang, Kyle Swanson, Wengong Jin, Connor Coley, Philipp Eiden, Hua Gao, Angel Guzman-Perez, Timothy Hopper, Brian Kelley, Miriam Mathea, Andrew Palmer, Volker Settels, Tommi Jaakkola, Klavs Jensen, and Regina Barzilay. Analyzing learned molecular representations for property prediction. *Journal of Chemical Information and Modeling*, 59:3370–3388, 8 2019.

[15] Pascal Wallisch, Michael Lusignan, Marc Benayoun, Tanya I. Baker, Adam S. Dickey, and Nicholas G. Hatsopoulos. Neural networks part i: Unsupervised learning. *Matlab for Neuroscientists*, pages 307–317, 1 2009.

[16] Liying Wang, Zhongtian Yu, Shiwei Wang, Zheng Guo, Qi Sun, and Luhua Lai. Discovery of novel sars-cov-2 3cl protease covalent inhibitors using deep learning-based screen. *European Journal of Medicinal Chemistry*, 244, 12 2022.

[17] About - chembl interface documentation.

[18] Focused and targeted libraries - chemdiv inc.

[19] Teague Sterling and John J. Irwin. Zinc 15 - ligand discovery for everyone, 11 2015.

[20] Aman Thakur, Ajay Kumar, Vivek Sharma, and Vineet Mehta. Pic50: An open source tool for interconversion of pic 50 values and ic 50 for efficient data representation and analysis.

[21] Rdkit - development infrastructure for the rdkit software provided by github and sourceforge.

[22] Dávid Bajusz, Anita Rácz, and Károly Héberger. Why is tanimoto index an appropriate choice for fingerprint-based similarity calculations? *Journal of Cheminformatics*, 7, 12 2015.