CHALMERS
UNIVERSITY OF TECHNOLOGY

UNIVERSITY OF GOTHENBURG

Machine Learning for PROTAC Decomposition and Enhanced Degradation Prediction

Master's thesis in Computer Science and Engineering

Ranxuan Zhang

Department of Computer Science and Engineering
CHALMERS UNIVERSITY OF TECHNOLOGY

UNIVERSITY OF GOTHENBURG
Gothenburg, Sweden 2025

Master's thesis 2025

Machine Learning for PROTAC Decomposition and Enhanced Degradation Prediction

Ranxuan Zhang

UNIVERSITY OF
GOTHENBURG

CHALMERS
UNIVERSITY OF TECHNOLOGY

Department of Computer Science and Engineering
Chalmers University of Technology
University of Gothenburg
Gothenburg, Sweden 2025

Machine Learning for PROTAC Decomposition and Enhanced Degradation Prediction

Ranxuan Zhang

Supervisor: Stefano Ribes, Department of Computer Science and Engineering
Eva Nittinger, AstraZeneca
Christian Tyrchan, AstraZeneca
Examiner: Rocío Mercado, Department of Computer Science and Engineering

Master's Thesis 2025
Department of Computer Science and Engineering
Chalmers University of Technology and University of Gothenburg
SE-412 96 Gothenburg
Telephone +46 31 772 1000

Cover: Description of the picture on the cover page (if applicable)

Typeset in LaTeX
Gothenburg, Sweden 2025

Machine Learning for PROTAC Decomposition and Enhanced Degradation Prediction

Ranxuan Zhang
Department of Computer Science and Engineering
Chalmers University of Technology and University of Gothenburg

Abstract

PROTACs (Proteolysis Targeting Chimeras) are bifunctional molecules composed of three components that mediate the degradation of target proteins, and are widely used in drug discovery. This project explores the application of machine learning in two key aspects: splitting PROTAC molecules into their three components (E3 ligase, linker, and POI), and predicting the degradation potential of PROTACs on target proteins. We evaluated an existing splitting model using internal data from AstraZeneca. Given recent updates to the public PROTAC dataset, we retrained the degradation prediction model on the expanded data. Additionally, we are transitioning the model from a binary classification task to a regression approach to directly predict degradation-related values such as DC_{50} and $D_{\max}$. We also investigated whether the solvent-accessible surface area (SASA) of lysine residues on the target protein influences degradation outcomes, though no clear relationship was observed.

Keywords: Deep Learning, Machine Learning, PRTOACs, Cheminformatics, Data analysis, CADD

Acknowledgements

I am truly grateful for the opportunity to carry out my thesis project at both Chalmers and AstraZeneca. The working atmosphere at both institutions has been exceptionally supportive - everyone is not only highly professional in their respective fields but also always open to communication and collaboration.

I would like to sincerely thank all my supervisors and my examiner for their invaluable guidance throughout this journey. Each meeting was like reaching a save point in a game - after our discussions, I always felt re-energized, reassured, and ready to move forward, knowing that I had made progress and could confidently continue my work.

Stefano, who has always been incredibly helpful. Communication with you was always easy and encouraging. You often motivated me by reminding me that we are making progress, and I very much appreciate your patience in explaining things from the foundational level whenever I had questions. You are truly a coding master.

Eva, thank you for always being proactive in communication and for starting each meeting with a warm smile. You encouraged me to ask more questions and to try new approaches. Your advice was always professional and practical, and your feedback was always detailed and timely.

Christian, your extensive knowledge of machine learning and computational chemistry is truly impressive. Your introductions to new papers during our group meetings not only broadened my perspective but also gave me valuable insight into the industry.

Rocío, you always got straight to the point and identified problems efficiently. Despite your busy schedule, you always brought energy and approachability to our discussions.

It has been a true pleasure working with all of you. I have learned so much - not only professionally, but also in a very enjoyable way. Everyone demonstrated great expertise in their own areas and was incredibly supportive to me as well.

I am grateful that you all repeatedly reminded me that there are no stupid questions. Although I still did not ask as many as I wanted, knowing this really made a difference for me. It helped me realize the importance of communication and that asking questions is actually a highly efficient way to make progress.

Thank you for providing such a healthy and exemplary working environment - it has set a high standard that I hope to encounter again in the future. Looking back, there are many things I could have done better, but these experiences have become my valuable lessons for the future. Finally, I want to thank my fellow thesis students for sharing this journey with me, for all the lunchtime chats, and for the after-work moments. Thank you!

Ranxuan Zhang, Gothenburg, 2025-06-23

Contents

List of Figures

List of Tables

1

Introduction

Targeted protein degradation (TPD) is gaining significant attention due to its potential to therapeutically target proteins that have been challenging to address with traditional small molecules. And proteolysis-targeting chimeras (PROTACs) is one of the molecules that can achieve TPD. Leveraging endogenous E3 ubiquitin ligases, PROTACs can facilitate the degradation of proteins of interest (POIs) through the ubiquitinproteasome system (UPS) [1]. To elaborate further, PROTACs are an innovative class of therapeutic agents designed to selectively degrade disease-related proteins. As shown in Figure1.1, PROTACs consist of three parts: a ligand (warhead) targeting the protein of interest (POI), a ligand that recruits an E3 ubiquitin ligase, and a chemical linker that connects the two ligands [2]. Unlike traditional inhibitors, which often block protein function, PROTACs leverage the cell's ubiquitin-proteasome system to tag and degrade target proteins. This catalytic mechanism enables PROTACs to target previously "undruggable" proteins, offering new therapeutic strategies for diseases such as cancer and neurodegenerative disorders [3], [4].

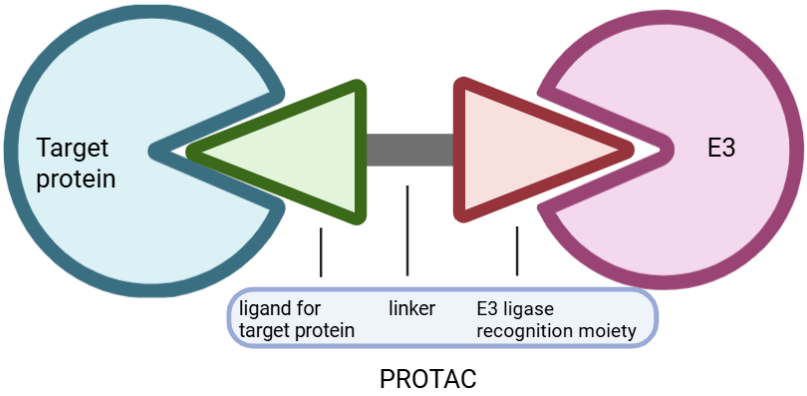


Figure 1.1: Example of PROTAC Structure Created with BioRender[5]

Although PROTAC is considered as an effective drug, manually dissecting a PROTAC molecule into these three functional components is time-consuming and inefficient for large-scale applications, a machine learning model named PROTAC-Splitter has been trained and tested on public data to automatize the decomposition pro-

cess. And the performance of this model will be evaluated using internal data in this study.

In drug discovery, the efficiency of a PROTAC is commonly evaluated using two key metrics: DC_{50}, the concentration required to achieve 50% degradation of the target protein, and $D_{\max}$, the maximum degradation achieved. Accurately predicting these values is crucial for assessing the therapeutic potential of a PROTAC and optimizing lead compounds.

The surface area around a protein that is accessible to a hypothetical solvent sphere may relate to its degradability. Solvent-accessible surface area (SASA) is a measure of this property. In this study, the relationship between DC_{50}, $D_{\max}$, and the SASA value, particularly the SASA value of lysine, will also be discussed.

This project aims to transform the current degradation predictor model from a binary classification approach into a regression model. Initially, separate regression models will be developed to predict DC_{50} and $D_{\max}$. Ultimately, the goal is to create a multi-task model capable of predicting both DC_{50} and $D_{\max}$ simultaneously. By predicting and discovering more information, this approach can potentially reduce the experimental workload involved in the early-stage design and testing of PROTAC drugs.

1.1 Risk analysis and ethical considerations

First, data quality and potential bias must be carefully considered. Ensuring a well-balanced dataset is essential to prevent biases that could impact model predictions and downstream decision-making. Data preprocessing and augmentation techniques may be necessary to improve generalizability.

This project will involve the use of internal data, which must strictly adhere to the company's data privacy policies and confidentiality agreements. All data handling and processing will comply with relevant regulatory frameworks, including GDPR, to protect sensitive information. Proper data-sharing agreements should be established to govern access and usage.

AI's role in drug discovery also raises ethical concerns. While AI can significantly accelerate research, incorrect predictions may lead to wasted resources or the selection of suboptimal drug candidates. Therefore, AI should function as an assistive tool rather than an autonomous decision-maker, with all critical evaluations and final decisions made by domain experts to ensure reliability and accountability.

2

Background

This chapter introduces the main theoretical concepts, including the underlying biochemical mechanisms, data representation.

2.1 The ubiquitin-dependent degradation pathway and PROTACs

The UPP is a crucial cellular mechanism responsible for the degradation of most intracellular proteins, maintaining protein homeostasis, regulating various cellular processes, and removing damaged or misfolded proteins. As shown in Figure2.1 The UPP begins with the ubiquitination process, where ubiquitin, a small regulatory protein, is activated by E1 enzyme (ubiquitin-activating enzyme). It is then transferred to E2 enzyme (ubiquitin-conjugating enzyme) and, with the aid of an E3 ligase enzyme, ubiquitin is covalently linked to substrate proteins through an isopeptide bond formed between its C-terminal glycine and the -amino group of a lysine residue on the target protein, serving as a post-translational modification [6]. This polyubiquitin chain tags the protein for recognition by the proteasome, a large protein complex with proteolytic activity. The proteasome unfolds and translocates the tagged protein into its catalytic core, where the protein is degraded into small peptides while ubiquitin molecules are recycled and reused [7].

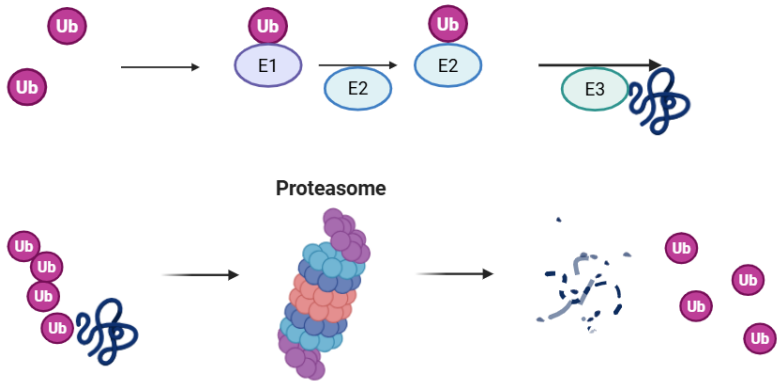


Figure 2.1: Visual representation of the ubitiquination, ubquitin proteins are referred as Ub, E1 interacts with E2, and transfers the ubiquitin molecule to E2. E2 interacts with E3-binding substrate and transfers the ubiquitin molecule to the substrate. Adapted from Fig 1 in [6]. Created with BioRender[5]

PROTACs, as mentioned, can use UPP, the three parts of PROTACs: a warhead that can bind to the target protein, a component that binds to an E3 ubiquitin ligase, and a linker that connects these two parts. When PROTAC binds to both the target protein and the E3 ligase, it brings them close together. This allows the E3 ligase to attach ubiquitin molecules to the target protein, marking it for destruction by the proteasome, the cells protein recycling center. This process reduces the amount of the target protein in the cell, offering a new way to handle proteins that cannot be targeted by traditional drugs, which could help in the treatment of different diseases [8].

2.2 Evaluation of the effectiveness of PROTAC

Because the ubiquitin-dependent degradation pathway is a complex multistep process, even for well-designed chimeric PROTACs that efficiently penetrate cells and form stable ternary complexes by tightly interacting with both binding partners, the effective degradation of the POI is not guaranteed. The degradation process also requires efficient transfer of ubiquitin to an available lysine on the surface of POI, as well as recruitment and effective degradation by the proteasome [9]. In addition, measuring PROTAC activity is challenging due to the hook effect. This phenomenon occurs when high concentrations of PROTAC are used, leading to the saturation of binary complexes such as PROTAC+E3 ligase and PROTAC+POI. This saturation hinders the formation of the ternary complex necessary for the degradation of the target protein [10]. Thus, accurately quantifying PROTAC activity is difficult in real experiments.

A critical metric for evaluating the efficacy of PROTAC molecules is the half-maximum degradation concentration (DC_{50}). This value represents the concentration of a PROTAC required to degrade 50% of the target protein under specific

experimental conditions. As a comprehensive indicator, DC_{50} reflects a PROTACs potency, binding affinity, and efficiency in engaging the ubiquitin-proteasome system. Lower DC_{50} values correspond to higher efficacy, indicating that less of the compound is required to achieve significant protein degradation. Another key metric, the maximum degradation effect ($D_{\max}$), defines the maximum level of protein degradation that a PROTAC can achieve under optimal conditions. Both DC_{50} and $D_{\max}$ are essential for a comprehensive evaluation of PROTAC efficacy, as $D_{\max}$ captures the upper limit of degradation, while DC_{50} reflects the concentration-dependent response.

2.3 Data Representation

2.3.1 Molecular Fingerprints

To apply machine learning to chemical data, molecular structures need to be translated into numerical forms. One widely used solution is the molecular fingerprint. It transforms structural features - such as atom types, bonds, rings, and functional groups - into a fixed-length binary vector. Each bit in the vector represents the presence or absence of particular chemical patterns. This abstract representation enables efficient comparison of molecules and is a standard input format in cheminformatics workflows, including virtual screening, similarity analysis, and predictive modeling. For example, the MACCS fingerprint is a type of structural key fingerprint based on a predefined list of 166 common chemical substructures, with each bit corresponding to a specific pattern such as a hydroxyl group or aromatic ring [11]. And the Morgan fingerprint generates circular substructures around each atom up to a specified radius and hashes them into a bit vector [12] . In this project, molecular fingerprints, including MACCS and Morgan, were employed to encode chemical compounds into numerical representations for use in data analysis and machine learning tasks.

2.3.2 SMILES

In addition to molecular fingerprints, Simplified Molecular Input Line Entry System (SMILES) is another chemical notation that allows a user to represent a chemical structure in a way that can be used by the computer [13]. SMILES encodes a molecule as a line of text using short ASCII strings that describe atoms and bonds in a linear form. For example, the SMILES string `CC(=O)O` represents acetic acid. This compact textual representation allows for easy storage, comparison, and conversion into other chemical formats or features, and is often used as input for cheminformatics tools and deep learning models.

2.3.3 UniProt

UniProt is a comprehensive, high-quality, and freely accessible database of protein sequence and functional information. It provides detailed annotations for proteins, including their sequence, structure, biological functions, subcellular locations, and involvement in diseases [14]. In this project, one column of the PROTAC dataset

contains the UniProt IDs of their target proteins, providing standardized protein identifiers for downstream analysis.

2.3.4 SASA

Solvent-accessible surface area (SASA) is a measure defined as the surface area around a protein that is accessible to a hypothetical solvent sphere, which interacts with the van der Waals contact surface of the protein molecule. SASA is widely used in structural biology to assess the exposure of amino acids or functional groups to the solvent. While direct experimental estimation of accurate SASA values for folded proteins at an atomic level is challenging, computational methods can provide reliable estimates from atomic coordinates [15].

SASA quantifies how much of a molecules surface is accessible to the solvent, reflecting the exposure level of amino acid residues. In this project, since ubiquitin attaches to the target protein specifically at surface-exposed lysine residues, the SASA value of lysines on the target protein is a critical feature. A higher SASA value indicates greater surface accessibility, which could correlate with the likelihood of ubiquitination and subsequent protein degradation. Therefore, incorporating lysine SASA values can improve the prediction of protein degradation.

2.3.4.1 Alphafold DB

Experimentally determining high-resolution structures for a wide range of proteins is a highly labor-intensive process. AlphaFold, an AI system developed by DeepMind can makes state-of-the-art predictions of protein structures from their amino-acid sequences [16]. With the exceptional accuracy and speed of AlphaFold, an extensive database of structure predictions at a large scale can be created now. This extensive collection is known as the AlphaFold Protein Structure Database (AlphaFold DB), a vast digital repository of predicted protein structures. Users can download the predicted structure of a given protein in PDB format for further analysis by searching the protein's UniProt id [17].

2.4 Dataset source

The data utilized in this study are from public datasets: PROTAC-DB 3.0 [18] and PROTAC-Pedia [19]. PROTAC-DB version 3.0 has been updated from version 2.0, form 5,388 entries to 9380 entries. PROTAC-Pedia comprises 1,203 entries. The number of unique SMILES in PROTAC-DB is 6,111, compared to 1,178 in PROTAC-Pedia. 1,222 SMILES entries are found in both PROTAC-DB and PROTAC-Pedia. Both datasets provide comprehensive information about PROTACs, including the compound ID, UniProt ID of the target protein, SMILES notation of the PROTAC sequence, and the name of the E3 ligase. It also contains key assay results, such as DC_{50} and $D_{\max}$, and additional assay details like the cell line and assay duration, which can be extracted for further analysis.

2.5 Related work

The emergence of machine learning in drug discovery has transformed how researchers design and optimize therapeutic candidates. 2D models are commonly used for molecular representation, with SMILES [13] encoding providing a text-based description of molecular graphs. While these approaches have been successful in fragment-based drug discovery (FBDD), they face limitations like struggle to capture the distinct chemical properties and structural complexity of large, multivalent molecules like PROTACs.

For the automatic splitting of PROTAC molecules, Zheng *et a l.* [20] proposed a novel framework for rational PROTAC design by combining an augmented transformer architecture with memory-assisted reinforcement learning. The study focused on pretraining a fragment-linking model using a transformer neural network and a dataset of quasi-PROTAC small molecules. These molecules shared a chemical space similar to PROTACs, providing a strong foundation for efficient molecular design. This innovative approach significantly advanced PROTAC optimization by enhancing the exploration of chemical space and guiding linker design. To efficiently predict PROTAC effectiveness, a machine learning model, DeepPROTACs, which takes the structure of the three PROTAC components as input, combines GNN and LSTM, achieving 77.46% accuracy on the test set [21]. For PROTACs degradation prediction, Ribes *et al.* [22], introduced instead a machine learning framework for predicting PROTAC degradation activity using curated datasets and a deep learning-based model. The study leveraged embeddings for PROTAC structures, E3 ligases, POIs, and cell types, with performance validated through multiple test scenarios. The framework achieved a top accuracy of 82.6% and ROC-AUC of 0.848, demonstrating its utility in forecasting PROTAC activity. Their approach highlighted the need for more comprehensive datasets and enhanced generalization across novel protein targets.

3

Methods

This chapter details the methodology followed throughout the project, including project foundation, data preparation, processing, model structure, and evaluation.

3.1 Previous Work and Foundation

This project builds upon the prior work conducted by Ribes *et al.* [22], who developed models for PROTAC molecule splitting and degradation prediction respectively. Our work starts from their established framework. For the PROTAC Splitter model, the exited model can split the PROTAC molecular into its three components and can achieve above 90% for the correct splitting, meaning the components splited by the model is identical with the label one. In this project ,we focuses on internal data testing and analyze the internal data and public data that used as training data. For the PROTAC degradation model, the model first ocnduct three data splitting method, including split based on how similarity the PROTAC is, the Uniprot of target protein and simply split randomly. The model can classify the PROTAC into active or non-active class by embedding PROTAC SMILES, Uniprot of target protein, the class of E3 ligase, the cell type as input. In this project, we first split out a share test set with each three splitting methods mention earlier count for 33% to use to compare the performance of model trained on different we extend the analysis by incorporating new features such as SASA values, altering the target splitting method from being based on UniProt IDs to using sequence alignment among UniProts and modifying the output from classification to regression on DC_{50} and $D_{\max}$. We also

3.2 PROTAC Splitter

The PROTAC splitter model was trained on data curated from public dataset (PROTAC-DB and PROTAC-Pedia).

To obtain accurate test results, the input data should include SMILES representations for PROTACs along with the correct labels corresponding to the three components: E3 ligand, linker, and warhead. For this purpose, the PROTAC molecules are systematically mapped to their respective fragments through an iterative process. By initially examine their SMILES representations to identify and categorize each structural component: the E3 ligand, linker, and warhead. The procedure

begins by utilizing a predefined dictionary of known fragments obtained from internal data. Each SMILES string is canonized to facilitate consistent matching, and dummy atoms are removed to ensure the structural accuracy of the fragments.

After data curation, the labels of the internal data should be accurately split and able to pass a resemblance test. This test verifies whether the three fragments from the label can be recombined to reconstruct the original SMILES string. Then the model trained on the public dataset is loaded to perform testing on internal data.

The test results are then processed using a scoring script, which compares the three components of the predicted splitting with those of the labeled data and provides a statistical summary of how many fragments are correctly split for each part.

To compare the internal dataset and public dataset that used for training, a similarity analysis was conducted on the three fragment of PROTACs using the Tanimoto similarity calculated from both MACCS and Morgan fingerprints. Additionally, we visualized the relationship between splitting correctness and similarity scores.

3.3 PROTAC Degradation Predictor

The PROTAC degradation predictor model is based on a multi-layer neural network architecture. Each input vectorconsisting of the Morgan fingerprints of the PROTAC molecule and the normalized embeddings of the protein of interest (POI), E3 ligase, and cell typeis first processed separately through independent linear layers. Each linear layers output is then passed through a softmax activation function to normalize the feature magnitudes and make them comparable across inputs. The resulting normalized vectors from the softmax layers are then summed element-wise, followed by a ReLU activation and batch normalization. Finally, the combined representation is passed through a linear layer to produce the models final output.

3.3.1 Data curation

To attain the useful information from dataset such as PROTAC compound details, cell line identifiers, E3 ligase names, protein of interest (POI), and degradation metrics such as DC_{50} and $D_{\max}$, the data curation is essential.

The total entries for PROTAC-DB is 9380, there are a lot of columns in the dataset, some entries are not complete. The columns we need is Uniprot(the Uniprot of target protein), E3 ligase(the name of E3 ligase), Smiles(the SMILES of PROTAC), DC_{50} nm $D_{\max}(\%)$,

For PROTAC-Pedia, the task involves synchronizing the names and formats of useful columns, including PROTAC SMILES, Cells, DC_{50} and $D_{\max}$, E3 ligase, target, especially for entries with multiple cell names.

Data extraction from PROTAC-DB includes pulling out assay-related details such as cell type, target protein name, DC_{50}, and $D_{\max}$ values, which are essential for performance evaluations. Textual assay descriptions were parsed using regex to

extract cell type information from statements like degradation in LNCaP cells and the extraction would be "LNCaP", remaining only the name of cells.

Then merge PROTAC-DB and PROTAC-Pedia, cannonicalize SMILES of PROTACs using RDKit to ensure standardized chemical representations. Subsequent standardization of cell line names was achieved using the Cellosaurus database, removing synonyms for clarity. Missing UniProt IDs for E3 ligases and POIs were manually added to ensure dataset completeness.

3.3.2 Data Splitting

To ensure a diverse and robust evaluation of the model, three distinct data splitting strategies were employed. The standard split involves randomly partitioning the dataset. The similarity-based split separates the training and test sets according to the Tanimoto similarity between PROTAC molecules, ensuring that structurally similar compounds do not appear in both sets. Lastly, the target-based split divides the data based on different target proteins, so that the model is evaluated on entirely unseen targets. PROTACs were divided into a 90% training-validation set and a 10% test set for all splitting methods. Uniform Manifold Approximation and Projection (UMAP) [23] is a dimensionality reduction technique that preserves both local and global structure of high-dimensional data, making it particularly useful for visualizing complex datasets in two or three dimensions. As illustrated in Figure 3.1, when the data is visualized using UMAP, it becomes evident that with random splitting, the test and training sets exhibit significant similarity. This often results in favorable test outcomes, but these results do not necessarily reflect the models true generalizability or predictive strength [24]. Consequently, in this project, two additional strategies are implemented to achieve more meaningful data separation and thus more rigorous model validation.

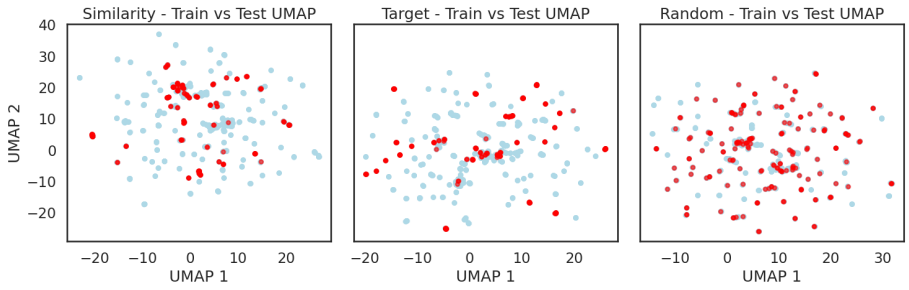


Figure 3.1: UMAP visualization of different data splitting strategies. From left to right: similarity-based splitting (based on PROTAC structural similarity), target-based splitting (based on distinct target proteins), and standard random splitting. The red points represents the test set and blue points represent the train and validation set.

3.3.2.1 Standard Splitting

In this approach, PROTAC entries are split randomly across datasets. This method serves as a control to assess model performance without systematic bias.

3.3.2.2 Similarity Splitting

PROTACs are divided into training-validation set and testing set based on Tanimoto similarity scores. This method leverages Tanimoto coefficient rankings, segmenting the dataset into a specified number of bins. Using these bins, the function organizes data entries by descending similarity, initially selecting the least similar molecules for the test set. This method helps ensure that the test set comprises molecules that are less similar to the training set, promoting robust model evaluation.

3.3.2.3 Target Splitting

This strategy involves splitting PROTACs based on the sequence similarity of their target proteins. In our dataset, there are 159 unique UniProt identifiers corresponding to different target proteins, and their sequences were retrieved from the UniProt database. Sequence alignment was then performed on these 159 sequences to compute pairwise similarities. A similarity matrix was generated from the alignment, with values ranging from 0 to 100%, representing percent identity between each pair of proteins.

To illustrate the structure of the similarity matrix, a simplified version containing only 8 UniProt entries is shown in Figure3.2. Additionally, we plotted a histogram of the highest percent identity per UniProt across all comparisons (Figure **??**).

Based on this distribution, we selected 70% as a similarity threshold: if a UniProt has a highest percent identity greater than or equal to 70% with any other UniProt, they are grouped into the same cluster, indicating sequence similarity. Otherwise, the UniProt is placed in its own singleton cluster, implying it is distinct from the others.

Using this method, we obtained 139 clusters from the 159 unique UniProt targets, of which 132 were singletons. To ensure that similar target proteins do not appear in both training and test sets, all non-singleton clusters were assigned to the training set. This approach guarantees that the test set contains only targets that are dissimilar to those in the training set, supporting the goal of evaluating model generalization across unseen target proteins.

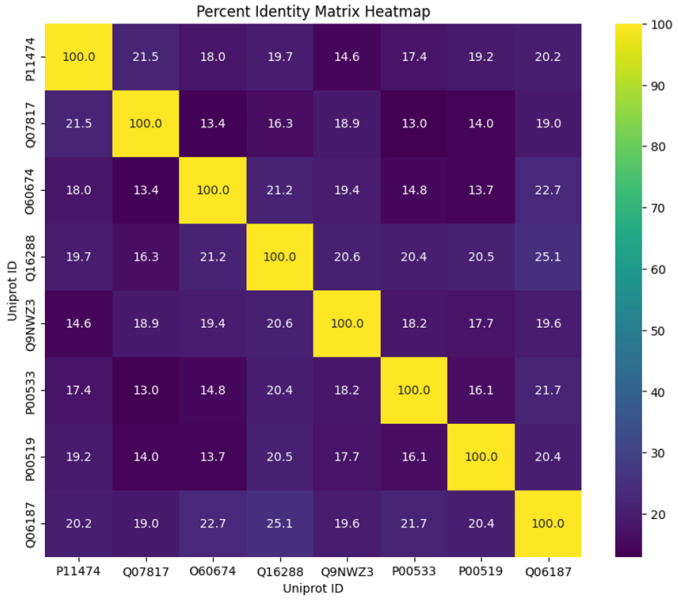


Figure 3.2: Simplified version of the similarity matrix, comparing only 8 UniProts.

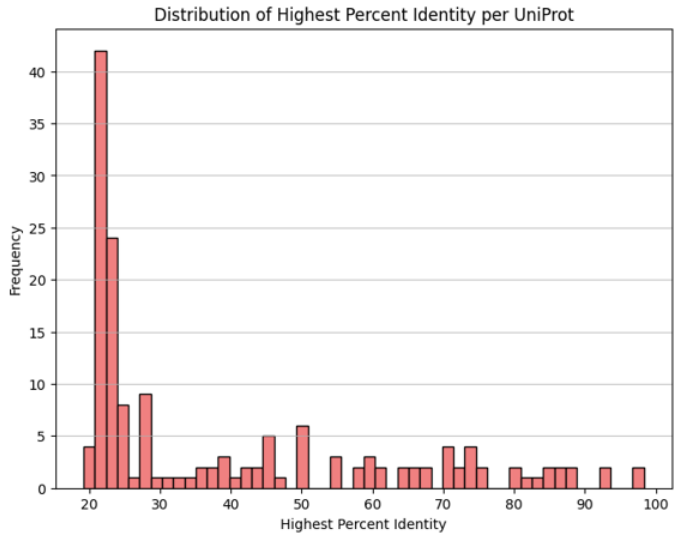


Figure 3.3: Distribution of the highest percent identity per UniProt.

3.3.2.4 Shared Test Set

To clearly compare the model's performance on different test datasets, for example, to see how a model trained using a random splitting method can perform in predicting the degradation level of PROTACs with previously unseen target proteins, we devised a specific test set splitting strategy. We began by extracting a 3.3% test set using the target splitting method. From the remaining data, another 3.3% was extracted using the similarity splitting method. Finally, we extracted 3.3% randomly from the rest. These three 3.3% test sets were then combined to form a 10% shared test set that is not used for training. The remaining data then undergoes three different dataset splitting studies. The splitting process is illustrated in Figure 3.4.

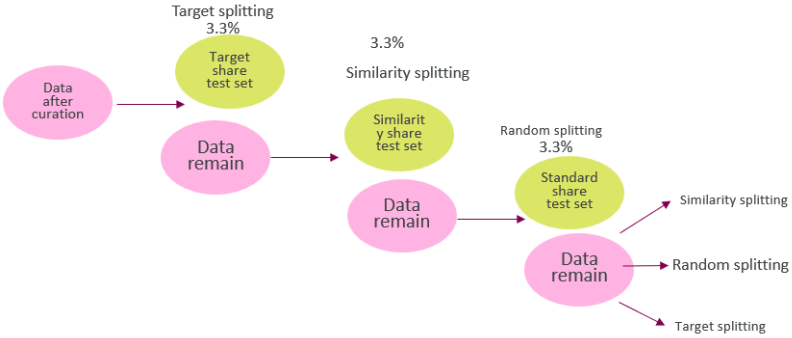


Figure 3.4: Share test set split method, green ellipse in the figure represents the portion of the shared test set.

3.3.3 Data embedding

For PROTAC molecules, their SMILES strings were converted using RDKit into 256-bit Morgan fingerprints with a radius of 2, including stereochemistry information. The choice of 256 bits corresponds to the smallest vector length that avoids any overlap between fingerprints. The two proteins, representing the E3 ligase and the protein of interest (POI), were transformed into precomputed UniProt embeddings of 1024 dimensions [25] [26]. Cell line information was obtained from the Cellosaurus database and encompasses features such as omics data, genome ancestry, doubling time, and sequence variations. These textual characteristics were ranked by uniqueness and filtered to create a concise, unified text description for each cell line [27]. Finally, a pretrained sentence Transformer model [28] was employed to encode these descriptions into 768-dimensional numerical embedding vectors.

3.3.4 SASA value and embedding

These SASA values provide insights into the accessibility of amino acid residues, in this project, indicate the accessibility of Lysine on the surface of target protein. The process of calculating SASA values begins by obtaining the protein structure predictions from AlphaFold [29] using UniProt IDs via the AlphaFold website. Once the predicted structure (PDB file) is acquired, it is processed using the Biopython library [30], which allows for detailed exploration of the protein's structure. The structure is organized according to the SMCRA (Structure/Model/Chain/Residue/Atom) architecture, facilitating the extraction of residues [31].

For each residue, if the residue name is LYS (lysine), the SASA value is calculated. This calculation considers all atoms within the residue except those in the backbone ('N', 'CA', 'C', 'O'), with the resulting SASA value representing the solvent-accessibility of the LYS side chain. A SASA value list for each UniProt entry can be generated if repeat this process. To explore the relationship between SASA value and degradation of target protein, the R 3.1 is defined as the proportion of lysine residues within a protein that have a solvent-accessible surface area (SASA) greater

than or equal to a specified threshold (n). The numerator represents the count of lysine residues that meet or exceed this SASA threshold, while the denominator indicates the total number of lysine residues in the protein.

$$R_n = \frac{\text{Number of Lys with SASA} \geq n}{\text{Number of Lys}} \tag{3.1}$$

3.3.5 Change classification model to regression model

Same as in the original classification-based prediction model, the regression model also takes as input the SMILES strings of the PROTAC molecules, UniProt identifiers of the target proteins, cell line names, and E3 ligase identifiers.

Since both DC_{50} and $D_{\max}$ are key continuous variables that quantify the degradation efficiency of a target protein, two separate regression models were built to predict each of these values. For each model, only entries with non-missing values for the corresponding label were included in the training and testing datasets. Compared to the original model that outputs a binary active/inactive label, the regression models instead predict continuous values of DC_{50} or $D_{\max}$.

Two commonly used loss functions for regression tasks are Mean Squared Error (MSE) and Mean Absolute Error (MAE). MSE penalizes larger errors more heavily by squaring the differences between predicted and true values, while MAE computes the average of the absolute differences and is more robust to outliers.

The model adopt MAE loss in place of binary cross-entropy with logits.

Subsequently, a multi-task model that jointly predicts both DC_{50} and $D_{\max}$ will be developed. The separate single-task models serve as baselines, with the expectation that the multi-task model will achieve higher accuracy by leveraging the shared information between the two related prediction tasks, leading to improved generalization.

3.3.6 Model training process

For the regression task, the target variables need preprocessing. For Dmax, the raw value is divided by 100, since it is stored as a percentage in the dataset. For DC50, The value (in nM) is first transformed to molar (M) units by multiplying by 1×10^{-9}. Then, pDC_{50} is calculated as equation 3.2.

$$\text{pDC}_{50} = -\log_{10}\left(\frac{\text{DC}_{50}\ (\text{nM})}{10^9}\right) \tag{3.2}$$

For each data splitting method, after data embedding, the model undergoes hyperparameter optimization using Optuna ($n_{\text{trials}} = 100$). The objective is to find the set of hyperparameters that minimizes the validation mean absolute error (*val_mae*) across the cross-validation folds. Using the optimal set of hyperparameters, three models are trained per data splitting strategy, each initialized with a different random seed to account for model variability.

4

Results

4.1 PROTAC Splitter

4.1.1 Comparision of public training data and internal testing data

To better understand the relationship between internal test data and public training data, we first identified the number of unique components within the PROTAC molecules. The counts for each of the three components are summarized in Table 4.1.

Table 4.1: Statistics of Unique Items in Datasets

Unique item	E3	Linker	POI
Training Set	217	1626	745
Internal Test Set	328	370	401

We also assessed the similarity between internal and public datasets by calculating the Tanimoto similarity scores using Morgan fingerprints with a radius of 2 and a fingerprint size of 512 bits. . As shown in Figure 4.1, the distributions of similarity scores for the three PROTAC components are presented separately. A similarity score of 1 indicates identical molecules. Among the components, E3 exhibits the highest similarity between internal and public datasets, indicating that E3 elements in the internal dataset are highly consistent with those in the public dataset. The linker component also displays a large number of identical molecules. In contrast, the POI component demonstrates greater variability, suggesting noticeable structural differences between the internal and public data for this region.

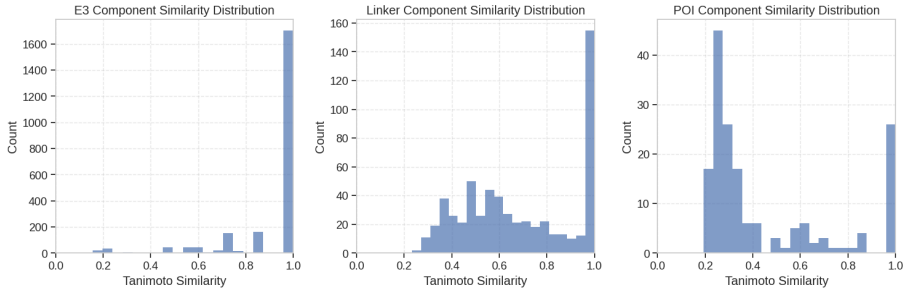


Figure 4.1: Similarity comparison between public training data and internal test data. From left to right: E3 ligase, linker, and POI.

4.1.2 Splitter result

The splitting process involves beam search, an optimization algorithm that explores a graph of possibilities by expanding the most promising nodes in a limited set. In our implementation, the model generates the top 5 results with the highest probabilities. The result of the splitter can be identified as correctly split or not, which means the split components exactly match the labeled data or not.

As mentioned, PROTACs are composed of three components. When the model correctly splits two of the three components, a fixing function can determine the remaining component by calculating the difference between the original PROTAC and the two identified components. 907 out of 2256 PROTACs were initially split correctly. After applying the fixing function, the number of correctly split PROTACs increased to 967, improving the accuracy by 2.6 % (from 40.2% to 42.8%). Using the top-5 results from beam search can increase the accuracy by an additional 5%. These results illustrated that both the fixing method and the top-5 beam search approach can improve the splitting accuracy.

To investigate the relationship between structural similarity and model performance, we analyzed how the similarity of internal test compounds to public training data correlates with splitting accuracy. Figure 4.2 presents a violin plot illustrating the distribution of Tanimoto similarity values for correctly and incorrectly predicted PROTAC components. Correct here means the prediction is totally equal with the label. As the Figure 4.2 shows, correctly split PROTACs predominantly cluster in regions of higher structural similarity to the training data. This pattern confirms our hypothesis that the model demonstrates superior performance when encountering compounds that share greater structural resemblance with its training examples. Interestingly, while both E3 ligase and linker components show high similarity distributions regardless of prediction outcome, correctly predicted instances (labeled as "true") exhibit notably less variation in their similarity profiles, suggesting more consistent structural recognition. The POI component reveals the most distinctive pattern among the three PROTAC elements, with a clear separation in similarity distributions between correctly and incorrectly predicted instances.

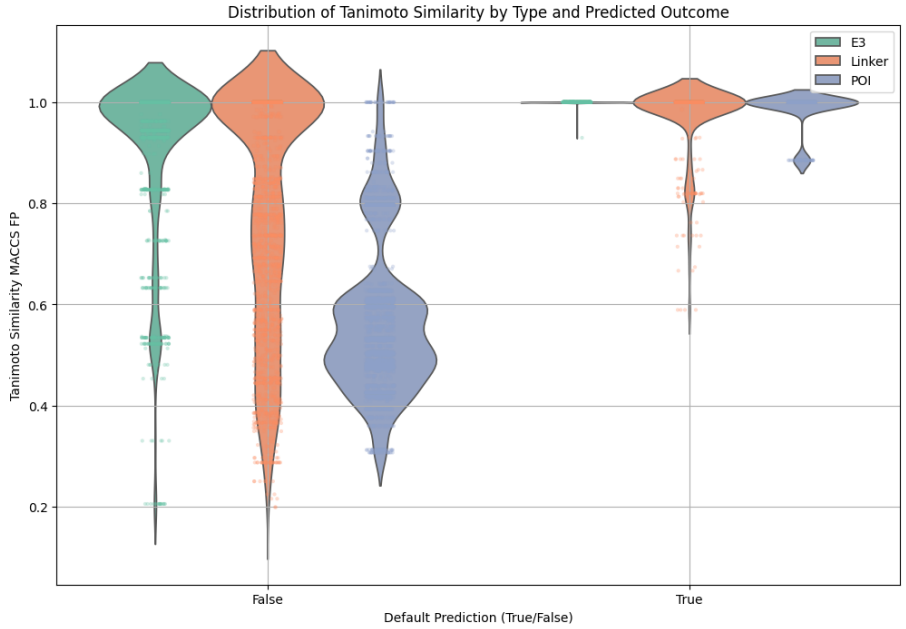


Figure 4.2: Relationship between correct splitting and Tanimoto similarity for PROTAC components. The x-axis indicates whether a component (E3 ligand, Linker, or POI) is correctly split, while the y-axis shows the Tanimoto similarity between internal and public datasets. Correctly split components tend to cluster in higher similarity regions. All three PROTAC components (E3 ligand, linker, and POI) are represented in the plot.

4.2 PROTAC Degradation Predictor

4.2.1 Relation between SASA value and Degradation

As detailed in the Methods section, we quantified the surface accessibility of lysine residues by calculating the SASA for each lysine residue within a protein. We then defined a metric R_n as the ratio of lysine residues with SASA values exceeding a specified threshold to the total number of lysine residues present in the protein. Figure 4.3 illustrates the relationship between median $D_{\max}$ for protein that have the same Uniprot and R values across various thresholds ranging from 0.2 to 0.8.

As observed, regardless of the threshold of n, most data points are concentrated around high $D_{\max}$ values (greater than 80%). This suggests that there may be no significant relationship between the amount of accessible lysine residues values and $D_{\max}$. One possible explanation is dataset bias - researchers may tend to report and upload only PROTACs with strong degradation performance, resulting in an over-representation of high $D_{\max}$ cases. Despite the graph showing no strong relationship between SASA metrics and $D_{\max}$ values, these surface accessibility measurements may still prove valuable as model inputs. While they might not be predictive on their own, SASA-derived features could interact meaningfully with other variables in a machine learning context. The R_n value can still be considered to be an input feature in the model for future analyses.

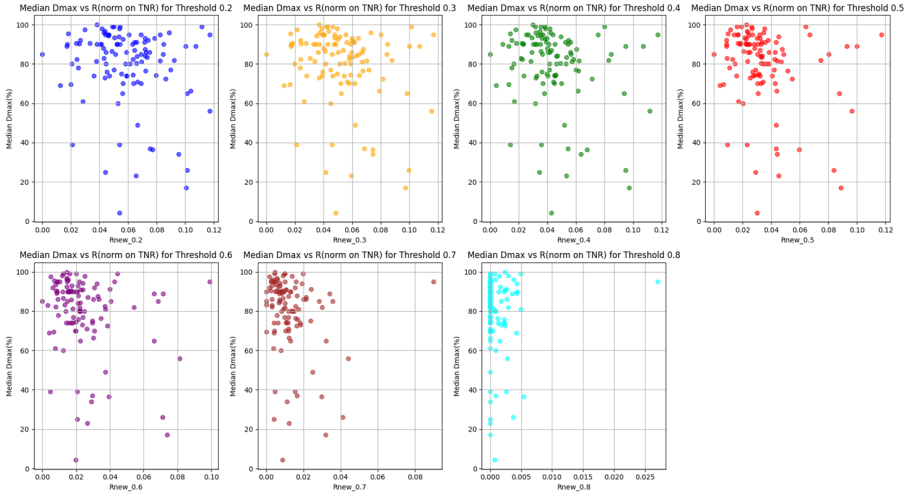


Figure 4.3: Relationship between R_n and the median $D_{\max}$ for the same UniProt, with SASA threshold n ranging from 0.2 to 0.8.

4.2.2 Data curation and splitting result

As previously mentioned, the public PROTAC dataset has been updated, increasing the number of available entries from approximately 5,000 to 9,000. After data curation, there are 3200 valid entries left, the counts for valid values for each columns can be seen in Figure 4.5. Specifically, the $D_{\max}$ and DC_{50} distribution can be seen in the venn diagram (Figure 4.4), only 1208 entries with both $D_{\max}$ and DC_{50}.

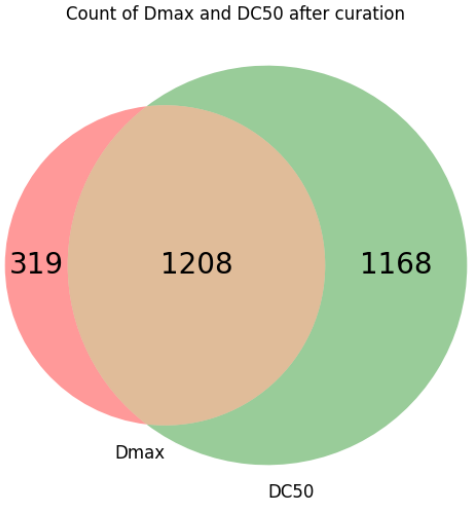


Figure 4.4: Venn plot for $D_{\max}$ and DC_{50} after data curation.

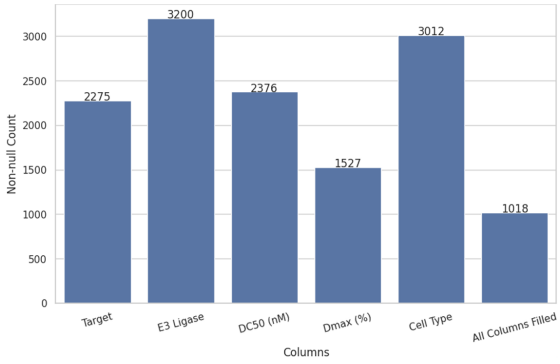


Figure 4.5: Data distribution after curation, x-axis is the name of columns that can be used as input for the model, numbers indicate the count for valid values for each columns

In the dataset, there is a column indicating the name of the target protein for each PROTAC. However, these protein sequences may include mutations compared to their canonical representations in UniProt. This feature can serve as a useful indicator to validate whether the target-based data splitting has been successfully

implemented.

To improve clarity in the visualization, we mapped each target name to a numerical label. As shown in Figure 4.6, there is no overlap in target proteins between the training and test sets, which demonstrates that the target-based splitting has been successfully carried out.

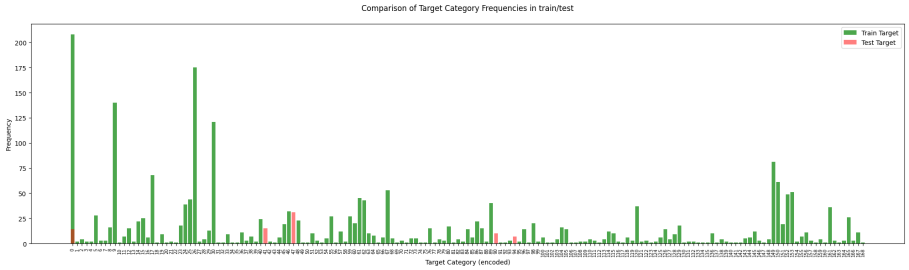
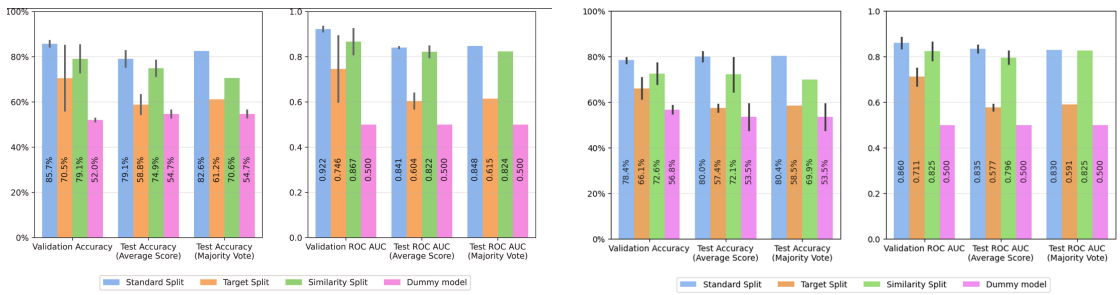


Figure 4.6: Target splitting result, with x axis stands for the name of target protein,.

4.2.3 Performance of classification model on updated data

The classification model is retrained on the curated updated dataset. Figure 4.7b and fig 4.7a shows the performance of the classification model train and test in updated dataset and old dataset respectively. For each splitting strategy (standard, target-based, or similarity-based), the plots report the mean validation accuracy and ROC-AUC scores obtained from five cross-validation models (one per fold), each trained with the best hyperparameters identified. The test performance of three models per strategy, all trained with the optimal hyperparameters but initialized with different random seeds, is also shown in the plot. For these models, we report the mean test accuracy and ROC-AUC, as well as the values obtained through majority voting across the three models. As a baseline, a dummy model that consistently predicts the majority class from the training set is included.



(a) Performance of the classification model on the old dataset across different data splitting methods.

(b) Performance of the classification model on the updated dataset across different data splitting methods.

Figure 4.7: Comparison of classification model performance between (a) old dataset and (b) updated dataset across different data splitting methods.

By comparing the two plots, despite the increase in dataset size, there is no signif-

icant improvement in test accuracy or ROC-AUC, indicating potential limitations in model generalizability.

4.2.4 Performance of classification model on share test data

The shared test set for classification comprises 151 entries, evenly distributed among the three splitting methods (random, similarity-based, and target-based). As described in the methodology section, for each data splitting strategy, the training process yields three best-performing models. We tested the three best models on shared test set and get the majority vote from from these models to determine whether a PROTAC is active or not. Then calculate the accuracy on different subset of the share test set, which were split by different data splitting strategy. The results can be seen from the table 4.2. The data presented in table 4.2 reveal that,regardless of the training data splitting method, all models generally demonstrate superior performance on the randomly split test set. Among the three models, the one trained on the randomly split dataset exhibits the best overall performance, but the performance declines when tested on PROTACs containing previously unseen target proteins.

Test splitting \ Model	Random	Similarity	Target
Random	0.822	0.822	0.711
Similarity	0.702	0.491	0.484
Target	0.694	0.450	0.537

Table 4.2: Accuracy of prediction of classification model on share test set. This table is structured as follows: the columns represent different models obtained from training on various data splitting methods (random, similarity, and target-based), while the rows indicate the sub-test sets that were split out using these corresponding strategies. This layout allows for a comprehensive comparison of model performance across different training and testing conditions.

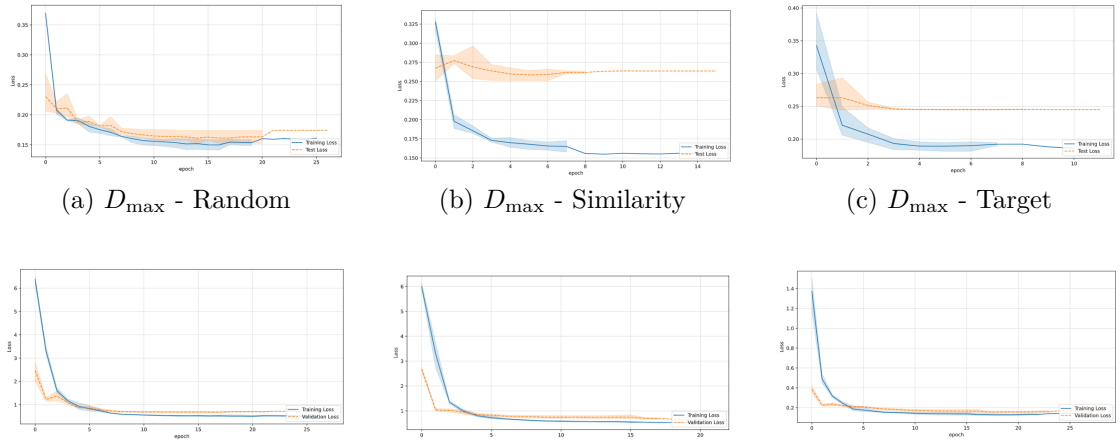
4.2.5 Result of regression model

Mean Absolute Error (MAE) is employed as the loss function for our regression tasks, also as a performance metric for evaluating the results. Separate regression models for $D_{\max}$ and pDC_{50} are trained and tested on the remaining data (after splitting out the shared test set). The results are visualized in the following plots and quantified using MAE.

For a more intuitive visualization, we present scatter plots of predicted versus true values, along with distribution histograms of both predicted and true values. Dashed lines in the histograms indicate the median values for each distribution. Figure 4.8 illustrates the results for $D_{\max}$ models trained on different data splitting methods, while Figure 4.9 shows the corresponding results for pDC_{50} models. The MAE values are reported in each subplot, providing a quantitative measure of model performance.

From these figures, we can see that, both $D_{\max}$ and pDC_{50} models demonstrate better fitting when trained on data using the random splitting strategy, as indicated by lower MAE values and improved alignment in the scatter plots. As the splitting strategy becomes more complex, progressing from random to similarity-based to target-based, the models appear to face increasing difficulties in fitting the data, as evidenced by increasing MAE values and less alignment from the plots. The training curves shown in Figure 4.10 also suggest limited effective learning occurs when employing the similarity-based and target-based splitting methods. For these latter two methods, the validation loss curve plateaus or begins to increase after only a few epochs, while the training loss continues to decrease. This pattern is indicative of overfitting, where the model performs well on the training data but fails to generalize to the validation set. In contrast, the models trained on randomly split data or $D_{\max}$ models show a more desirable learning trajectory, with both training and validation losses decreasing consistently over a longer period.

Overall, the pDC_{50} model exhibits superior fitting across all splitting methods compared to the $D_{\max}$ model when comparing the plots. As the histograms indicated, the pDC_{50} data distribution is more closely resembling a normal distribution while the $D_{\max}$ distribution appears more skewed, with a higher concentration of values in the upper range. This difference in data distribution might be the reason for the varying model performances.



(a) $D_{\max}$ - Random (b) $D_{\max}$ - Similarity (c) $D_{\max}$ - Target

(d) pDC_{50} - Random (e) pDC_{50} - Similarity (f) pDC_{50} - Target

Figure 4.10: Training curves for models trained with different data splitting strategies. The top row shows models for predicting $D_{\max}$, and the bottom row shows models for predicting pDC_{50}. Columns from left to right represent Random, Similarity, and Target data splitting strategies.

4.2.5.1 Performance on share test set

The models were further evaluated on the shared test set, which consists of 161 entries for $D_{\max}$ and 225 entries for pDC_{50}, with samples evenly distributed among the three splitting methods. For this regression task, instead of using majority voting as in the classification task, we used the average prediction from the three best

models as the final result for each model type. Table 4.3 presents the results for the $D_{\max}$ model, while Table 4.4 shows the results for the pDC_{50} model. In both tables, columns represent models trained with different data splitting strategies (random, similarity-based, and target-based), while rows represent test subsets split from the shared test set using these same strategies. The Spearman coefficient in these tables is calculated between the predicted values and true values. This coefficient ranges from -1 to +1, +1 indicates a perfect positive monotonic relationship. This coefficient indicates the strength and direction of the monotonic relationship between the predicted and actual values, providing a measure of the model's ability to correctly rank the compounds based on their $D_{\max}$ or pDC_{50} values, even if the absolute predictions may not be exact.

Scatter plots for the shared test set results are presented in Figure 4.11 for $D_{\max}$ and Figure 4.12 for pDC_{50}, providing a visual complement to the tabulated data.

From both plots and table, for $D_{\max}$, the overall results are somewhat ambiguous, particularly for models trained using similarity and target-based methods. In terms of MAE, the model trained on randomly split data generally performs better. However, all models, including the random-split model, show decreased performance when predicting the target test set. This trend is visually apparent in the scatter plots (Figure 4.11), where predictions for the target test set tend to cluster around a narrow range of values, indicating limited discriminative power when faced with novel targets. Regarding the Spearman coefficients, most values are relatively low (below 0.4), suggesting weak monotonic relationships between predicted and actual values across most models and test sets. A notable exception is the pDC_{50} model trained on randomly split data, which shows a relatively high coefficient (0.570) when tested on the random test set, but a low coefficient (0.115) when tested on the target test set. This difference in Spearman coefficients could suggest that the model's ability to rank compounds may vary depending on the test set, with potential challenges when dealing with novel targets.

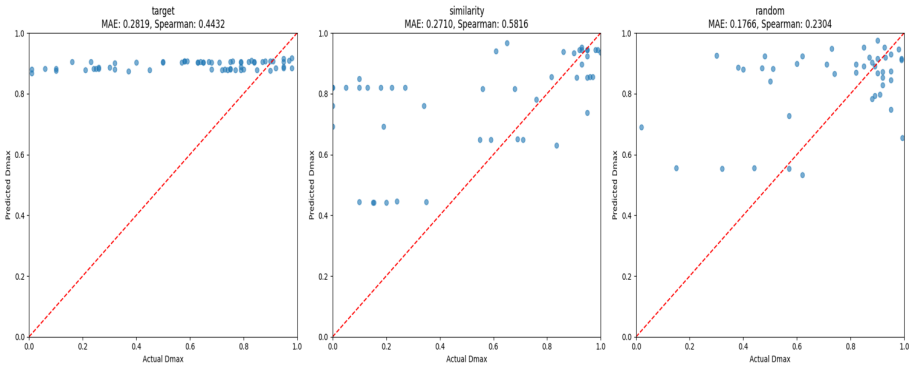
These observations underscore the challenges in generalizing PROTAC property predictions, especially to compounds with novel target proteins. The consistently better performance of models trained on randomly split data suggests that this approach may provide a more representative sampling of the chemical space. However, the decline in performance on target test sets indicates a need for strategies to improve model robustness to structural novelty in PROTACs.

Test splitting \ Model	Random		Similarity		Target	
	MAE	Spearman	MAE	Spearman	MAE	Spearman
Random	0.177	0.230	0.178	0.267	0.175	0.177
Similarity	0.271	0.582	0.367	-0.139	0.326	0.319
Target	0.282	0.442	0.243	0.311	0.235	0.279

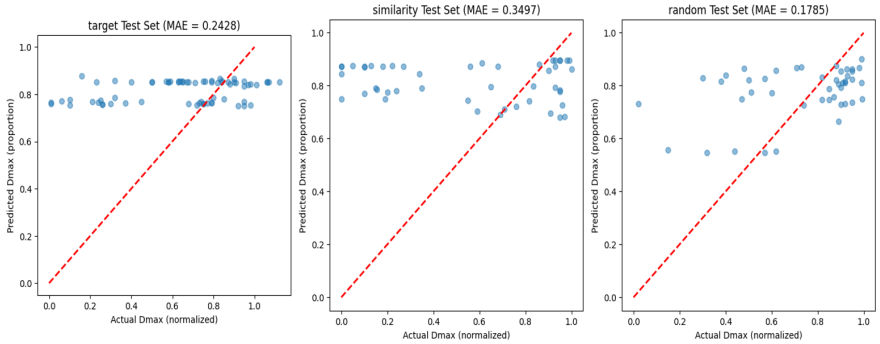
Table 4.3: $D_{\max}$ model performance comparison when applying different data splitting strategies

Model / Test splitting	Random		Similarity		Target	
	MAE	Spearman	MAE	Spearman	MAE	Spearman
Random	1.073	0.570	1.112	0.510	1.104	0.288
Similarity	0.960	0.370	1.011	0.283	1.255	0.294
Target	1.945	**0.115**	1.913	-0.537	1.497	-0.066

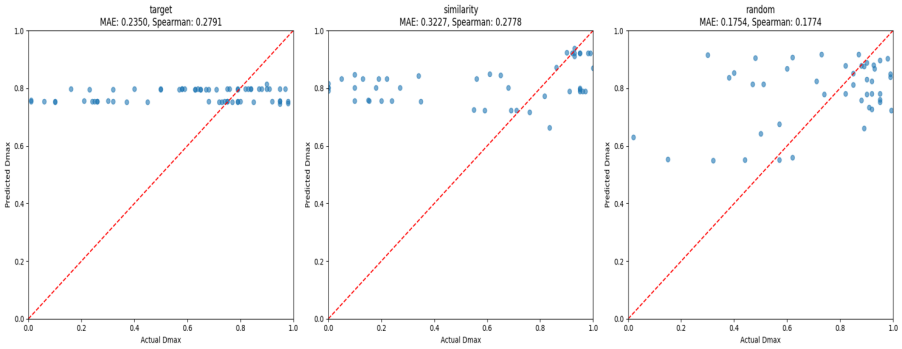
Table 4.4: pDC_{50} model performance comparison when applying different data splitting strategies



(a) Random strategy model

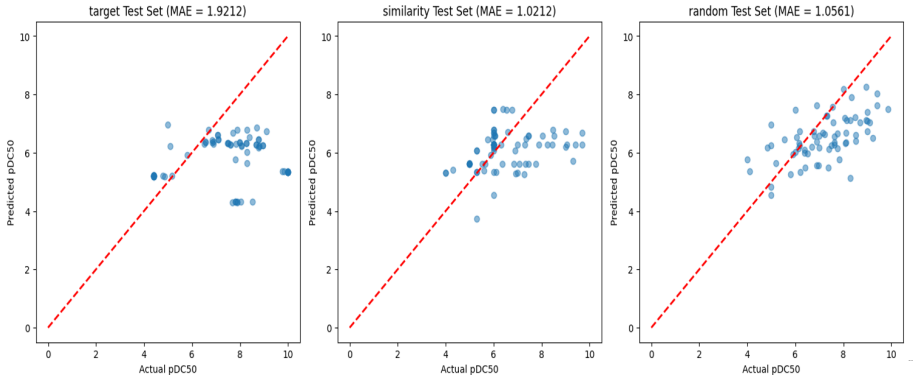


(b) Similarity strategy model

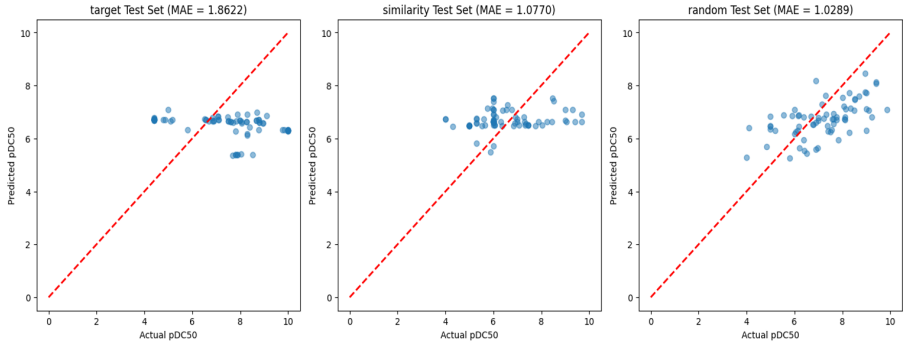


(c) Target strategy model

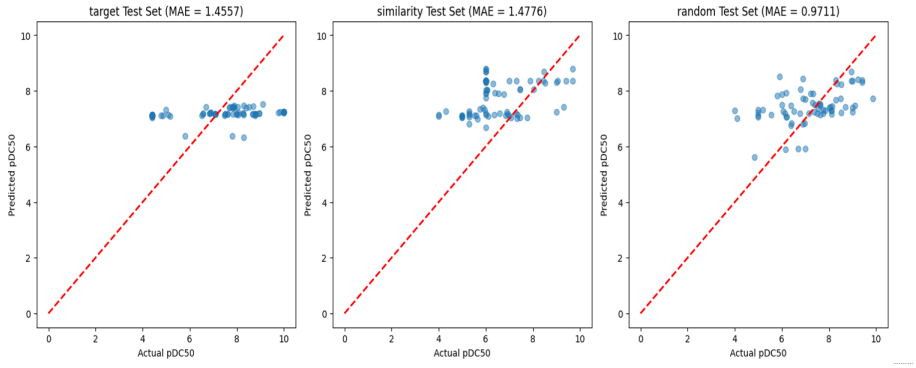
Figure 4.11: Performance comparison of $D_{\max}$ models trained with different data selection strategies on various test subsets. Each row represents a model trained using a specific strategy, while columns show performance on target, similarity, and random test subsets, respectively.



(a) Random strategy model

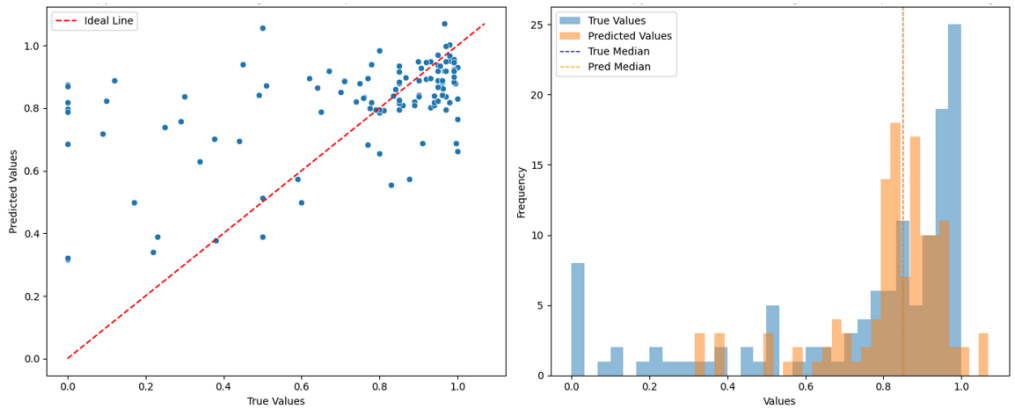


(b) Similarity strategy model

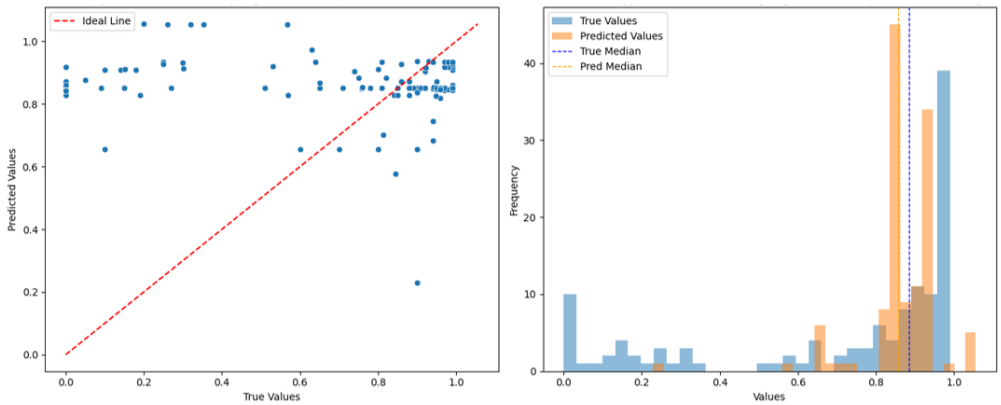


(c) Target strategy model

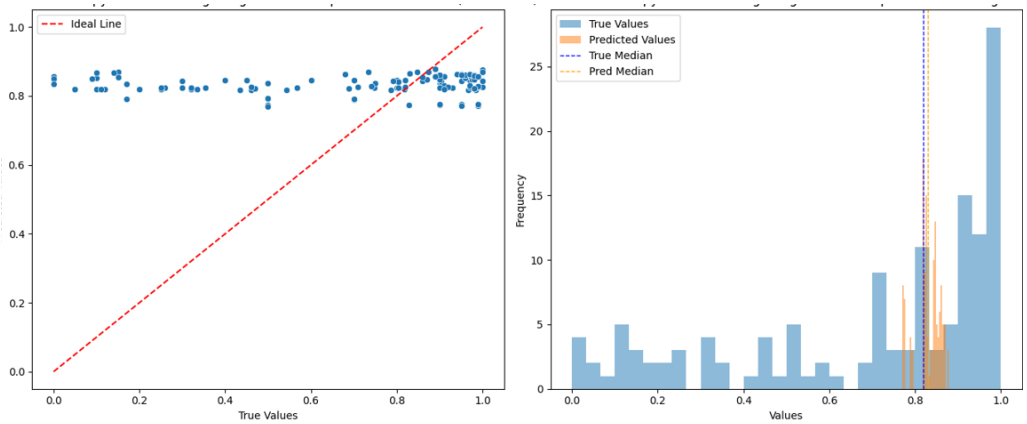
Figure 4.12: Performance comparison of $D_{\max}$ models trained with different data selection strategies on various test subsets. Each row represents a model trained using a specific strategy, while columns show performance on target, similarity, and random test subsets, respectively.



(a) Model trained with random data selection strategy, MAE=0.174

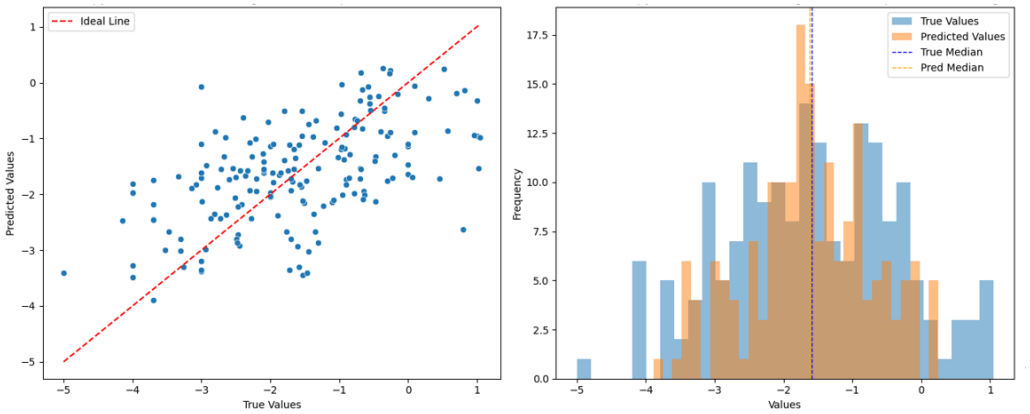


(b) Model trained with similarity data selection strategy, MAE=0.264

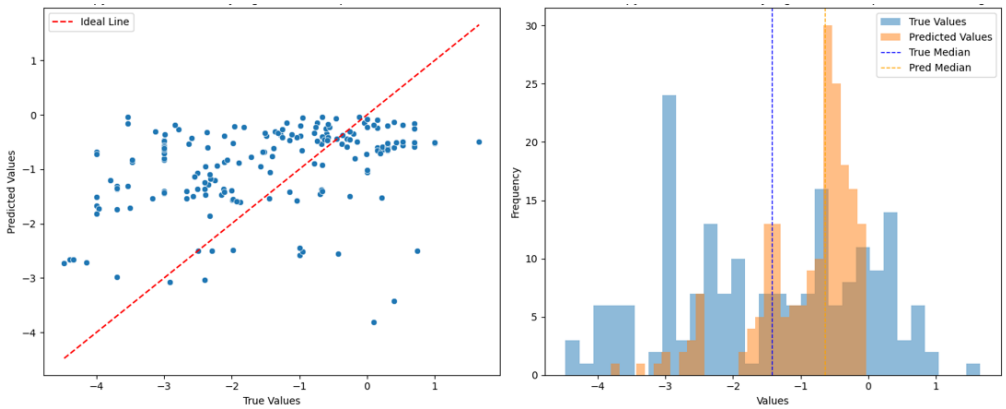


(c) Model trained with target data selection strategy, MAE=0.245

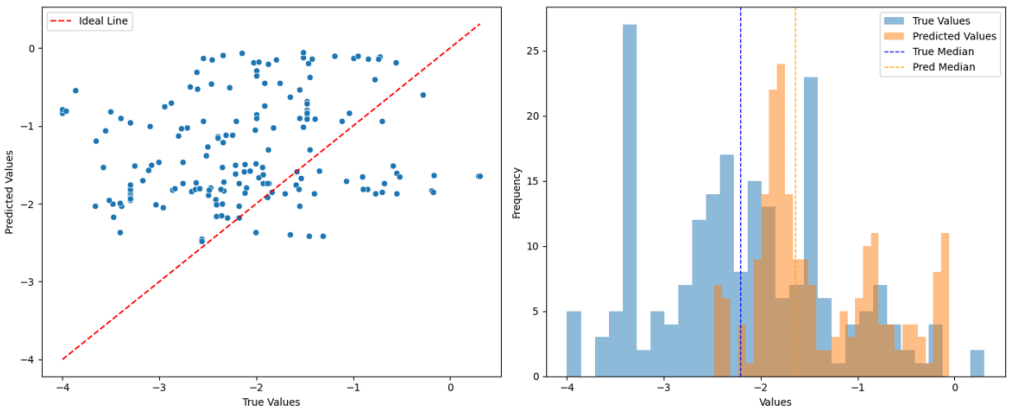
Figure 4.8: $D_{\max}$ prediction results using different data selection strategies. Each row represents a different data selection strategy: (a) random, (b) similarity, and (c) target. Left column: Scatter plots of predicted vs. true values. Right column: Distribution histograms of true (blue) and predicted (orange) values, with dashed lines indicating median values.



(a) Model trained with random data selection strategy



(b) Model trained with similarity data selection strategy



(c) Model trained with target data selection strategy

Figure 4.9: pDC_{50} prediction results using different data selection strategies. Each row represents a different data selection strategy: (a) random, (b) similarity, and (c) target. Left column: Scatter plots of predicted vs. true values. Right column: Distribution histograms of true (blue) and predicted (orange) values, with dashed lines indicating median values.

5

Conclusion

This project focuses on PROTAC molecules and can be broadly divided into two main components: PROTAC Splitter and PROTAC Degradation Predictor.

5.1 PROTAC Splitter

The PROTAC Splitter model aims to decompose a PROTAC molecule into its three constituent parts: the E3 ligase binder, linker, and POI (protein of interest) binder. We evaluated this model on an internal dataset, requiring data curation to align with the model's input format. Performance analysis was carried out by comparing the number of correct splittings before and after a fixing function, as well as using top-1 and top-5 predictions via beam search. Overall, the existing model achieves 40.2% accuracy for E3 ligand using only the top-1 prediction. Both the fixing function and top-5 beam search improved accuracy by approximately 2% and 5%, respectively. Furthermore, we analyzed the relationship between prediction accuracy and the Tanimoto similarity between internal and training data. The results suggest that PROTACs with higher similarity to the training set are more likely to be correctly split.

5.2 PROTAC Degradation Predictor

The PROTAC Degradation Predictor is designed to classify whether a given PROTAC molecule can induce degradation of its target protein or not.

Three data splitting strategies were evaluated: random splitting, in which PROTAC samples are distributed randomly between training and test sets; similarity-based splitting, where PROTACs that are less similar to the training set are assigned to the test set; and target-based splitting, where PROTACs with unseen target proteins are selected for the test set. To provide a more detailed assessment of model performance, a separate shared test set was constructed, with samples contributed equally from each splitting strategy.

Retraining on the larger dataset did not yield significant performance improvements. Models trained with random splitting performed the best, and performance dropped when tested on the target-splitting test set. This indicates difficulty for models to generalize to PROTACs with unseen target proteins.

Next, the classification approach was extended to a regression task, directly predicting key degradation metrics: DC_{50} and $D_{\max}$. The same model architecture was preserved, with only the loss function and output layer adapted for regression. Results show that random splitting again led to better predictive performance, while models using other splitting strategies did not effectively learn, especially on the target-based test set. Generally, DC_{50} was better fit than $D_{\max}$.

Then we convert the classification model into a regression model to directly predict the numerical degradation metrics: DC_{50} and $D_{\max}$,which are widely used indicators of degradation potency. Using the same model architecture but change the loss fucntion and output layer to suit the regression task. An additional observation is that the method of selecting the "best model" based solely on minimum validation MAE may not always correspond to the best overall alignment: some cross-validation models showed superior results. Alternative loss functions and further exploration of model generalizability to unseen targets are proposed as future directions.

5.2.1 SASA

The relationship between the solvent-accessible surface area (SASA) of lysine residues in target proteins and degradation efficacy (DC_{50} and $D_{\max}$) was explored. SASA values were computed from AlphaFold-predicted structures for each UniProt ID. However, no clear correlation between lysine accessibility and degradation outcomes was observed.

5.3 Future work

The distribution of $D_{\max}$ can be further normalized to reduce skewness and achieve a more uniform clustering of samples. Given the suboptimal performance of models on the similarity-based and target-based splitting strategies, future work should focus on improving model generalization to unseen data, for example, by exploring alternative loss functions, testing different model architectures, and incorporating advanced embedding techniques.

Additionally, implementing a multitask learning framework to simultaneously predict both $D_{\max}$ and DC_{50} could be a direction for further research.

Bibliography

[1] C. Wang, Y. Zhang, W. Chen, Y. Wu, and D. Xing, "New-generation advanced protacs as potential therapeutic agents in cancer therapy," *Molecular Cancer*, vol. 23, no. 1, May 2024, ISSN: 1476-4598. DOI: 10.1186/s12943-024-02024-9. [Online]. Available: http://dx.doi.org/10.1186/s12943-024-02024-9.

[2] K. M. Sakamoto, K. B. Kim, A. Kumagai, F. Mercurio, C. M. Crews, and R. J. Deshaies, "Protacs: Chimeric molecules that target proteins to the skp1cullinf box complex for ubiquitination and degradation," *Proceedings of the National Academy of Sciences*, vol. 98, no. 15, pp. 8554–8559, Jul. 2001, ISSN: 1091-6490. DOI: 10.1073/pnas.141230798. [Online]. Available: http://dx.doi.org/10.1073/pnas.141230798.

[3] J. Liu, J. Ma, Y. Liu, *et al.*, "Protacs: A novel strategy for cancer therapy," *Seminars in Cancer Biology*, vol. 67, pp. 171–179, Dec. 2020, ISSN: 1044-579X. DOI: 10.1016/j.semcancer.2020.02.006. [Online]. Available: http://dx.doi.org/10.1016/j.semcancer.2020.02.006.

[4] M. Pettersson and C. M. Crews, "Proteolysis targeting chimeras (protacs) past, present and future," *Drug Discovery Today: Technologies*, vol. 31, pp. 15–27, Apr. 2019, ISSN: 1740-6749. DOI: 10.1016/j.ddtec.2019.01.002. [Online]. Available: http://dx.doi.org/10.1016/j.ddtec.2019.01.002.

[5] BioRender. "Scientific image and illustration software." Accessed: 2025-05-07. (n.d.), [Online]. Available: https://www.biorender.com/.

[6] L. Zhao, J. Zhao, K. Zhong, A. Tong, and D. Jia, "Targeted protein degradation: Mechanisms, strategies and application," *Signal Transduction and Targeted Therapy*, vol. 7, no. 1, Apr. 2022, ISSN: 2059-3635. DOI: 10.1038/s41392-022-00966-4. [Online]. Available: http://dx.doi.org/10.1038/s41392-022-00966-4.

[7] A. Ciulli and W. Farnaby, "Protein degradation for drug discovery," *Drug Discov. Today Technol*, vol. 31, pp. 1–3, 2019.

[8] G. M. Burslem and C. M. Crews, "Proteolysis-targeting chimeras as therapeutics and tools for biological discovery," *Cell*, vol. 181, no. 1, pp. 102–114, Apr. 2020, ISSN: 0092-8674. DOI: 10.1016/j.cell.2019.11.031. [Online]. Available: http://dx.doi.org/10.1016/j.cell.2019.11.031.

[9] M. P. Schwalm, A. Krämer, A. Dölle, *et al.*, "Tracking the protac degradation pathway in living cells highlights the importance of ternary complex measurement for protac optimization," *Cell Chemical Biology*, vol. 30, no. 7, 753–765.e8, Jul. 2023, ISSN: 2451-9456. DOI: 10.1016/j.chembiol.2023.06.002.

[Online]. Available: http://dx.doi.org/10.1016/j.chembiol.2023.06.002.

[10] M. Konstantinidou, J. Li, B. Zhang, *et al.*, "Protacs a game-changing technology," *Expert Opinion on Drug Discovery*, vol. 14, no. 12, pp. 1255–1268, Sep. 2019, ISSN: 1746-045X. DOI: 10.1080/17460441.2019.1659242. [Online]. Available: http://dx.doi.org/10.1080/17460441.2019.1659242.

[11] J. L. Durant, B. A. Leland, D. R. Henry, and J. G. Nourse, "Reoptimization of mdl keys for use in drug discovery," *Journal of Chemical Information and Computer Sciences*, vol. 42, no. 6, pp. 1273–1280, Sep. 2002, ISSN: 0095-2338. DOI: 10.1021/ci010132r. [Online]. Available: http://dx.doi.org/10.1021/ci010132r.

[12] H. L. Morgan, "The generation of a unique machine description for chemical structures-a technique developed at chemical abstracts service.," *Journal of chemical documentation*, vol. 5, no. 2, pp. 107–113, 1965.

[13] D. Weininger, "Smiles, a chemical language and information system. 1. introduction to methodology and encoding rules," *Journal of Chemical Information and Computer Sciences*, vol. 28, no. 1, pp. 31–36, Feb. 1988, ISSN: 1520-5142. DOI: 10.1021/ci00057a005. [Online]. Available: http://dx.doi.org/10.1021/ci00057a005.

[14] A. Bateman, M.-J. Martin, S. Orchard, *et al.*, "Uniprot: The universal protein knowledgebase in 2025," *Nucleic Acids Research*, vol. 53, no. D1, pp. D609–D617, Nov. 2024, ISSN: 1362-4962. DOI: 10.1093/nar/gkae1010. [Online]. Available: http://dx.doi.org/10.1093/nar/gkae1010.

[15] S. Ali, M. Hassan, A. Islam, and F. Ahmad, "A review of methods available to estimate solvent-accessible surface areas of soluble proteins in the folded and unfolded states," *Current Protein amp; Peptide Science*, vol. 15, no. 5, pp. 456–476, May 2014, ISSN: 1389-2037. DOI: 10.2174/1389203715666140327114232. [Online]. Available: http://dx.doi.org/10.2174/1389203715666140327114232.

[16] J. Jumper, R. Evans, A. Pritzel, *et al.*, "Highly accurate protein structure prediction with AlphaFold," *Nature*, vol. 596, no. 7873, pp. 583–589, 2021.

[17] M. Varadi, S. Anyango, M. Deshpande, *et al.*, "Alphafold protein structure database: Massively expanding the structural coverage of protein-sequence space with high-accuracy models," *Nucleic Acids Research*, vol. 50, no. D1, pp. D439–D444, Nov. 2021, ISSN: 1362-4962. DOI: 10.1093/nar/gkab1061. [Online]. Available: http://dx.doi.org/10.1093/nar/gkab1061.

[18] G. Weng, X. Cai, D. Cao, *et al.*, "Protac-db 2.0: An updated database of protacs," *Nucleic Acids Research*, vol. 51, no. D1, pp. D1367–D1372, Oct. 2022, ISSN: 1362-4962. DOI: 10.1093/nar/gkac946. [Online]. Available: http://dx.doi.org/10.1093/nar/gkac946.

[19] Prilusky, *Protacpedia - protacs on 30012605*, en, 2016. [Online]. Available: https://protacdb.weizmann.ac.il/ptcb/detp?i=397.

[20] S. Zheng, Y. Tan, Z. Wang, *et al.*, "Accelerated rational protac design via deep learning and molecular simulations," *Nature Machine Intelligence*, vol. 4, no. 9, pp. 739–748, Sep. 2022, ISSN: 2522-5839. DOI: 10.1038/s42256-022-00527-y. [Online]. Available: http://dx.doi.org/10.1038/s42256-022-00527-y.

[21] F. Li, Q. Hu, X. Zhang, *et al.*, "Deepprotacs is a deep learning-based targeted degradation predictor for protacs," *Nature Communications*, vol. 13, no. 1, Nov. 2022, ISSN: 2041-1723. DOI: 10.1038/s41467-022-34807-3. [Online]. Available: http://dx.doi.org/10.1038/s41467-022-34807-3.

[22] S. Ribes, E. Nittinger, C. Tyrchan, and R. Mercado, "Modeling protac degradation activity with machine learning," *Artificial Intelligence in the Life Sciences*, vol. 6, p. 100 104, Dec. 2024, ISSN: 2667-3185. DOI: 10.1016/j.ailsci.2024.100104. [Online]. Available: http://dx.doi.org/10.1016/j.ailsci.2024.100104.

[23] L. McInnes, J. Healy, and J. Melville, "Umap: Uniform manifold approximation and projection for dimension reduction," 2018. DOI: 10.48550/ARXIV.1802.03426. [Online]. Available: https://arxiv.org/abs/1802.03426.

[24] P. Walters, *Some thoughts on splitting chemical data sets for improved model building*, https://practicalcheminformatics.blogspot.com/2024/11/some-thoughts-on-splitting-chemical.html, Accessed: 2025-04-17, 2024.

[25] en. [Online]. Available: https://www.uniprot.org/.

[26] C. Dallago, K. Schütze, M. Heinzinger, *et al.*, "Learned embeddings from deep learning to visualize and predict protein sets," *Current Protocols*, vol. 1, no. 5, May 2021, ISSN: 2691-1299. DOI: 10.1002/cpz1.113. [Online]. Available: http://dx.doi.org/10.1002/cpz1.113.

[27] A. Bairoch, "The cellosaurus, a cell-line knowledge resource," *Journal of Biomolecular Techniques: JBT*, vol. 29, no. 2, pp. 25–38, Jul. 2018, ISSN: 1943-4731. DOI: 10.7171/jbt.18-2902-002. [Online]. Available: http://dx.doi.org/10.7171/jbt.18-2902-002.

[28] N. Reimers and I. Gurevych, *Sentence-bert: Sentence embeddings using siamese bert-networks*, 2019. DOI: 10.48550/ARXIV.1908.10084. [Online]. Available: https://arxiv.org/abs/1908.10084.

[29] A. P. S. Database, *Alphafold protein structure database*, https://alphafold.ebi.ac.uk/, Accessed: 2025-05-15, 2023.

[30] P. J. Cock, T. Antao, J. T. Chang, *et al.*, "Biopython: Freely available python tools for computational molecular biology and bioinformatics," *Bioinformatics*, vol. 25, no. 11, pp. 1422–1423, 2009. DOI: 10.1093/bioinformatics/btp163.

[31] B. Project, *Biopython – python tools for computational biology*, https://biopython.org/, Accessed: 2025-05-15, 2023.

A

Appendix 1

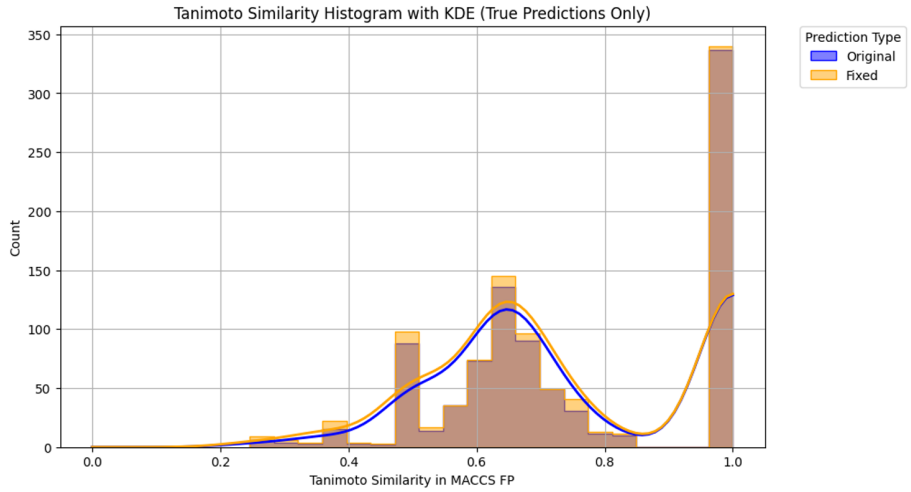


Figure A.1: Impact of function fixing on E3 ligand classification accuracy. The x-axis shows the Tanimoto similarity between internal and public datasets, while the y-axis represents the count of correctly classified E3 ligands. Blue bars indicate results before fixing, and orange bars show results after fixing the function.

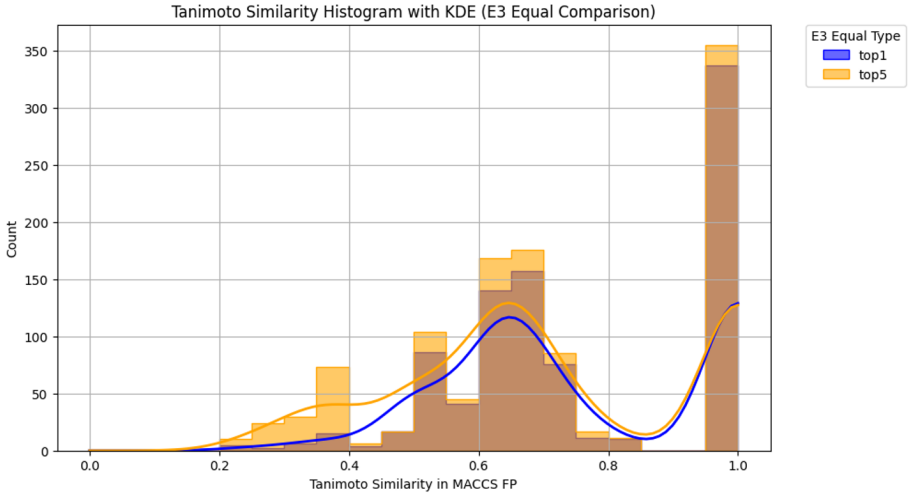


Figure A.2: Comparison of top-1 and top-5 accuracy in E3 ligand classification. The x-axis shows the Tanimoto similarity between internal and public datasets, while the y-axis represents the count of correctly classified E3 ligands. Blue bars indicate top-1 results, and orange bars show top-5 results using beam search.