Modeling PROTAC degradation activity with machine learning

Citation for the original published paper (version of record):
Ribes, S., Nittinger, E., Tyrchan, C. et al (2024). Modeling PROTAC degradation activity with machine learning. Artificial Intelligence in the Life Sciences, 6.
http://dx.doi.org/10.1016/j.ailsci.2024.100104

(article starts on next page)

Contents lists available at ScienceDirect

Artificial Intelligence in the Life Sciences

journal homepage: www.elsevier.com/locate/ailsci

Research Article

Modeling PROTAC degradation activity with machine learning

Stefano Ribes [a,*], Eva Nittinger [b], Christian Tyrchan [b], Rocío Mercado [a]

[a] Department of Computer Science and Engineering, Section for Data Science and AI, Chalmers University of Technology, Chalmersplatsen 4, Gothenburg, 412 96, Sweden
[b] Medicinal Chemistry, Research and Early Development, Respiratory and Immunology (R&I), BioPharmaceuticals R&D, AstraZeneca, Pepparedsleden 1, Mölndal, 431 83, Sweden

ARTICLE INFO

Dataset link: https://github.com/ribesstefano/PROTAC-Degradation-Predictor

Keywords:
PROTAC
Machine learning
Drug discovery
Protein degradation

ABSTRACT

PROTACs are a promising therapeutic modality that harnesses the cell's built-in degradation machinery to degrade specific proteins. Despite their potential, developing new PROTACs is challenging and requires significant domain expertise, time, and cost. Meanwhile, machine learning has transformed drug design and development. In this work, we present a strategy for curating open-source PROTAC data and an open-source deep learning tool for predicting the degradation activity of novel PROTAC molecules. The curated dataset incorporates important information such as pDC_{50}, D_{max}, E3 ligase type, POI amino acid sequence, and experimental cell type. Our model architecture leverages learned embeddings from pretrained machine learning models, in particular for encoding protein sequences and cell type information. We assessed the quality of the curated data and the generalization ability of our model architecture against new PROTACs and targets via three tailored studies, which we recommend other researchers to use in evaluating their degradation activity models. In each study, three models predict protein degradation in a majority vote setting, reaching a top test accuracy of 82.6% and 0.848 ROC AUC, and a test accuracy of 61% and 0.615 ROC AUC when generalizing to novel protein targets. Our results are not only comparable to state-of-the-art models for protein degradation prediction, but also part of an open-source implementation which is easily reproducible and less computationally complex than existing approaches.

1. Introduction

Machine learning (ML) has transformed various scientific domains, including drug design and discovery, by offering novel solutions to complex, multi-objective optimization challenges [1]. In the context of medicinal chemistry, ML techniques have revolutionized the process of identifying and optimizing potential drug candidates. Traditionally, drug discovery has relied heavily on trial-and-error experimentation, which is not only time-consuming but also expensive. ML techniques have the potential to significantly accelerate and improve this process by predicting properties of molecules *in silico*, such as binding affinity, solubility, and toxicity, with remarkable accuracy [2,3]. This in turn saves time and money in early-stage drug discovery by focusing resources on the most promising candidates. At the same time, AI models' high performance can potentially lead to better designed drugs for patients.

In order to develop ML models for chemistry, ML algorithms leverage vast datasets containing molecular structures, biological activities, and chemical properties to learn intricate patterns and relationships, also called quantitative structure-activity relationships (QSAR). These algorithms can discern subtle correlations and structure in molecular data that are difficult for human experts to identify. Consequently, ML-based approaches aid in predicting which molecules are likely to be effective drug candidates, thereby narrowing down the search space and saving resources [4].

PROTACs, or PROteolysis TArgeting Chimeras, represent an innovative class of therapeutic agents with immense potential in challenging disease areas [5–7]. Unlike traditional small molecule inhibitors, PROTACs operate by harnessing the cell's natural protein degradation machinery, the proteasome, to eliminate a protein of interest (POI), as summarized in Fig. 1(a). This catalytic mechanism of action for targeted protein degradation (TPD) offers several advantages over conventional approaches, which frequently work by having a small molecule drug bind tightly to and thus block a protein's active site. In fact, by leveraging their unique mechanism, PROTACs bypass the need for tight binding to specific protein pockets, offering a novel strategy for targeting previously "undruggable" proteins. This approach is particularly

[*] Corresponding author.
E-mail addresses: ribes@chalmers.se (S. Ribes), eva.nittinger@astrazeneca.com (E. Nittinger), christian.tyrchan@astrazeneca.com (C. Tyrchan), rocio.mercado@chalmers.se (R. Mercado).

https://doi.org/10.1016/j.ailsci.2024.100104
Received 31 May 2024; Accepted 4 July 2024
Available online 14 July 2024

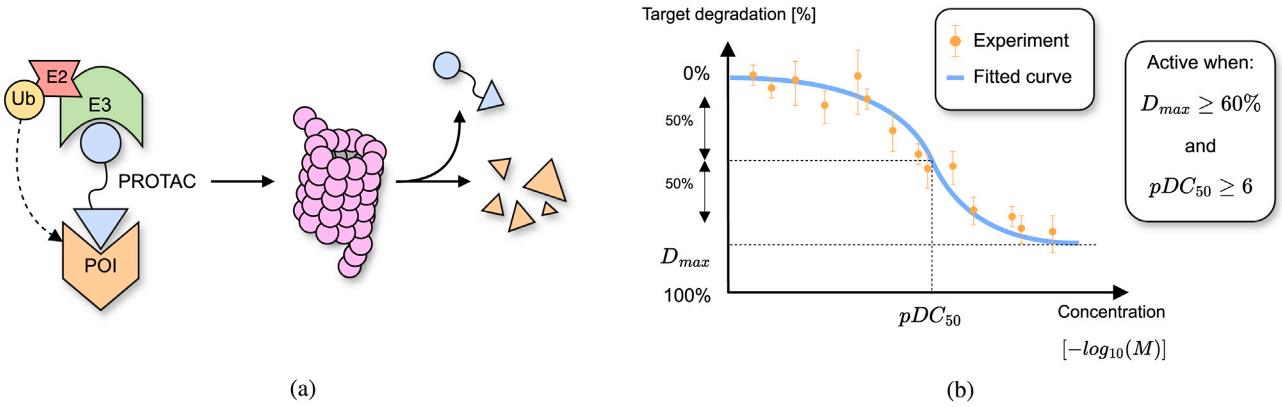


Fig. 1. (a) Schematic representation of the PROTAC mechanism of action: the proteasome (violet) degrades the ubitiquitinated POI targeted by the PROTAC. After degradation, the PROTAC becomes available again for new targets. (b) Example of a typical PROTAC dose–response curve, along with the activity thresholds used in this work.

relevant in cases where inhibiting the target's activity might not be sufficient; notable examples include certain neurodegenerative diseases like Alzheimer's, where misfolded proteins agglomerate and lead to negative downstream effects in patients [8].

By catalytically degrading POIs, PROTACs have the potential to offer more comprehensive therapeutic effects at lower doses relative to traditional inhibitors. Their capacity for TPD highlights the necessity of thorough efficacy evaluations, typically conducted through dose–response assessments (Fig. 1(b)) to determine critical parameters such as DC_{50} (the molar concentration of PROTAC at half maximum degradation of the POI; the lower the better) and D_{max} (the highest percentage of degraded POI; the higher the better) [9]. However, PROTAC development and evaluation face significant challenges due to the limited availability of open-source tools and resources specifically designed for this molecule class, a gap predominantly filled by tools aimed at small molecule inhibitors [10].

To address these challenges, our work introduces a comprehensive machine learning toolkit and curated data specifically designed for PROTAC research. We have developed predictive models that leverage the curated data to effectively forecast the degradation activity of PROTACs, achieving high predictive accuracy and ROC-AUC scores on the test set (top 82.6% and 0.848, respectively). Our system, fully open-source and easily accessible via a Python package, is designed to streamline the predictive modeling of PROTAC degradation activity, thus facilitating the rapid evaluation and optimization of new PROTAC designs. Our contribution significantly expands the available public resources for PROTAC development, setting a new baseline in the application of ML techniques to this emerging therapeutic area.

2. Materials and methods

2.1. Data curation

For this work, we collected and curated data from PROTAC-DB [11] and PROTAC-Pedia [12] that represent, to our knowledge, the two largest open datasets for PROTAC data. PROTAC-DB contains experimental data, scraped from the scientific literature, for 5388 PROTACs (as of May 2024; version 2.0). While the PROTAC-DB allows users to query, filter, and analyze PROTAC data via its online platform (*e.g.*, comparing different compounds based on their DC_{50} and D_{max}), its data is not specifically structured for ML models, but rather for online access through its web page. Wrangling the data for use in data-driven models requires significant cleaning and curation. On the other hand, PROTAC-Pedia provides 1190 crowd-sourced entries (as of May 2024), with details on PROTACs and their degradation activity.

To prepare the data for our models, we extracted and standardized the following features from the PROTAC-DB and PROTAC-Pedia datasets, where a specific combination of the features corresponds to one experiment: the PROTAC compound, cell line identifier, E3 ligase, POI, and degradation metrics (DC_{50} and D_{max}).

Each dataset entry includes the SMILES representation of the PROTAC, which was canonicalized using RDKit [13]. In PROTAC-DB, cell line information was predominantly found in textual assay descriptions, such as *"degradation in LNCaP cells after 6 h at 0.1/1000/10000 nM"*, with *"LNCaP"* being the cell type in this statement. Cell type information was extracted using regex parsing, with a few manually cleaned entries. Afterward, cell line names were standardized using Cellosaurus [14] to remove synonyms. The Uniprot IDs [15] of E3 ligases and POIs lacking that information were manually web searched and added as text to each entry.

For PROTAC-DB, some of the DC_{50} and D_{max} values were obtained by splitting entries containing information for the same PROTAC on multiple assays. A data sample is labeled as *active* when both its pDC_{50} (*i.e.*, the DC_{50} value expressed in negative log_{10} units) and D_{max} are above their respective predefined threshold values; here we used 6 and 60%, respectively. Effectively, each data point is assigned a binary label indicating degradation activity.

2.2. Data representation

Given the available data consisting of PROTACs, E3 ligases, POIs, and cell lines, our goal is to encode the diverse information into efficient numerical embeddings that an ML model can leverage. Because our pool of curated data has a limited size ($\sim 10^3$ data samples), we decided to focus on learning individual embeddings for each of the following: the PROTAC, E3 ligase, POI, and cell type for each experiment.

For PROTACs, their SMILES strings are converted, via RDKit [13], to Morgan fingerprints of 256 bits with a radius 10 and stereochemistry information included, with 256 being the smallest 2^n vector length not resulting in the overlap of any two fingerprints. The two proteins corresponding to the E3 ligase and POI are converted into precomputed Uniprot embeddings of 1024 elements [15,16]. Cell line information was extracted from the Cellosaurus database and includes omics, genome ancestry, doubling time, and sequence variations [14]. These characteristics, all in text form, are then ranked by uniqueness and filtered to form a concise single text description of a given cell line. Finally, a pretrained sentence Transformer model [17] was used to encode the text descriptions into numerical embedding vectors of 768 elements. More details on the cell line embedding process can be found in Appendix B.

Once we collected all the embeddings representations, each POI, E3 ligase, and cell embedding was normalized independently by removing the respective mean and by scaling to unit variance. The normalization parameters are learned on the given training set and kept fixed for validation and testing. Morgan fingerprints, being binary vectors, were not normalized.

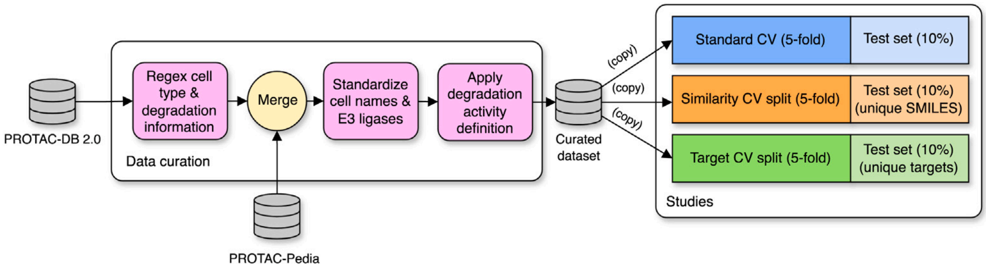

PROTAC-DB 2.0
Regex cell type & degradation information
Merge
Standardize cell names & E3 ligases
Apply degradation activity definition
Data curation
PROTAC-Pedia
Curated dataset
(copy)
(copy)
(copy)
Standard CV (5-fold)
Test set (10%)
Similarity CV split (5-fold)
Test set (10%) (unique SMILES)
Target CV split (5-fold)
Test set (10%) (unique targets)
Studies

Fig. 2. Data curation pipeline and proposed studies.

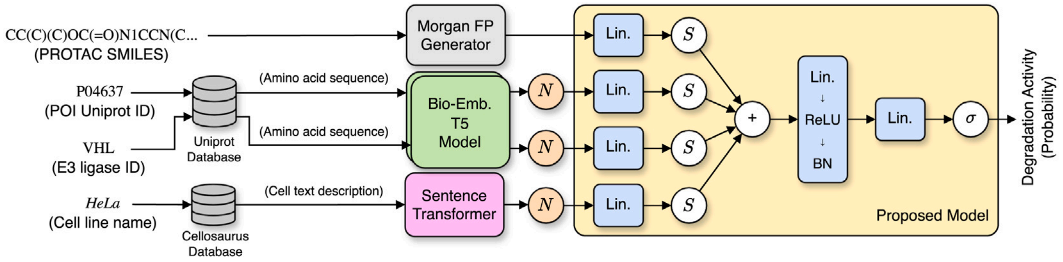


Fig. 3. Model pipeline and architecture. The normalization, softmax, and sigmoid functions are denoted as N, S, and σ, respectively. The pretrained bio-embedding model can be found in [16], while the pretrained sentence Transformer is from [17].

2.3. Model architecture

An illustration of the model architecture is shown in Fig. 3. The model includes a set of linear layers, each processing a separate input vector, *i.e.*, the Morgan fingerprints, and the normalized POI, E3 ligase, and cell embeddings, respectively. The linear layer outputs are then softmax-ed in order to make them of comparable magnitude, and finally summed together. Lastly, they are forwarded to two additional linear layers, interleaved by a ReLU activation function and a batch norm layer. The model is trained to optimize a binary cross-entropy loss (with logits). We set the batch size to 128 and reduce the learning rate by a factor of 10× whenever the validation loss increases compared to the previous training step. Finally, we apply a sigmoid function to the output of the final linear layer before returning predictions about PROTAC activity.

2.4. Evaluation strategy

To fully assess the quality of the curated data and the potential performance of DL models in predicting degradation activity, we designed a set of three studies (Fig. 2). In the first study, we seek to identify the potential upper bound of the model performance given the curated data. To do so, we randomly pick 10% of the data as a test set, and leave the remaining data for training with 5-fold cross validation (CV). This leads to an ensemble of five trained models, one per CV fold. In the next study, we explore model generalization against unseen POIs. Similar to the previous study, we carefully select 10% of the available data for testing, such that the POI does not appear in the remaining 90% of the data which is used for training (5-fold CV). Finally, we evaluate the model generalization performance to new PROTACs. To do so, we compute the average Tanimoto distance from all PROTAC Morgan fingerprints to all other PROTAC fingerprints in the full data. For generating the test set for this experiment, we isolated the data entries starting from the ones where their PROTAC is mapped to a high average Tanimoto distance, until reaching 10% of the total available data, leaving the rest for CV training.

For each study, we used stratified group CV as implemented in scikit-learn to ensure each fold has a balanced distribution of active and inactive compounds.

Table 1
Parameters optimized by Optuna: the table reports the parameter name, its type, *i.e.*, categorical (Cat) or continuous (Cont), and the range of values or options suggested in each trial. We apply SMOTE oversampling [18] to the concatenated input data, when suggested.

Parameter	Type	Options/Range
Hidden Dimension	Cat	[32, 64, 128, 256, 512]
Learning Rate	Cont (log)	[$1e^{-5}$, $1e^{-3}$]
Use SMOTE	Cat	[True, False]
SMOTE k Neighbors	Cat	[3, 4, …, 15]

2.5. Hyperparameter tuning and ablation studies

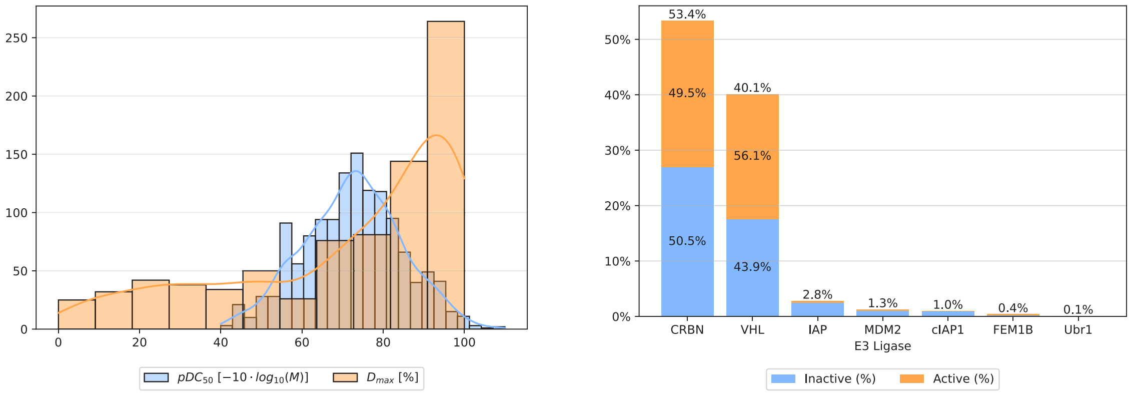
For hyperparameter tuning we leveraged the Optuna optimization framework [19]. In each study, we let Optuna spawn 150 trials to suggest a model architecture and hyperparameters to be used to train the models in the CV folds (we used 5 folds). Each trial is instructed to sample all the hyperparameters values listed in Table 1. Using Optuna, the goal is to find the best set of hyperparameters that maximize the average validation ROC-AUC score across the CV folds. The best hyperparameter configuration is then used to train three separate models per study, each with randomly initialized weights (with different seeds), in order to account for model variability. The best configuration models in each study are trained on the combined study's train and validation sets and evaluated on the respective held-out test set.

Additionally, we conducted an ablation study in which we progressively set input vectors to all zeros, and feed them to the three best models trained during the random split study.

3. Results and discussion

3.1. Degradation activity thresholds

A data sample was labeled *active* if its pDC_{50} is ≥ 6.0 (equivalent to 1 μM) and $D_{max} \geq 60\%$. The pDC_{50} threshold helps identify PROTACs with therapeutic potential, as molecules above this threshold are likely to show significant biological activity. Similarly, the D_{max} threshold helps identify PROTACs capable of achieving substantial degradation of the target protein, indicative of efficacy. pDC_{50} is particularly relevant for drug design, as it allows for the prioritization of compounds that not



(a)

(b)

Fig. 4. (a) Histogram of pDC_{50} and D_{max} in the full curated dataset. Note that the pDC_{50} values are scaled 10× to better display them along side D_{max} values, although they are not bounded by 0 and 100 as D_{max} is. (b) The percentage of curated data associated with each E3 ligase and the active/inactive percentage of data points per E3 ligase.

only bind to the POI but also lead to its effective degradation at a reasonable concentration. By choosing the above thresholds, we aimed to mitigate model bias, ensuring our dataset includes a balanced representation of both active and inactive compounds, enhancing the model's generalizability. Note that a PROTAC can be labeled active in one cell type and inactive in another, such as DT2216, a Bcl-xL degrader, which is active in MOLT-4 cancer cells ($pDC_{50}/D_{max} = 7.20/90.8\%$) and inactive in 2T60 hybrid cells ($pDC_{50}/D_{max} = 5.52/26.0\%$) [20].

3.2. Curated dataset

After data curation, we were able to extract a total of 2141 data samples, out of which 812 (37.9%) report information about D_{max}, and 1350 (63.1%) include a DC_{50} value. When applying the aforementioned definition of degradation activity, we isolated 857 data samples, 437 (50.99%) of which are labeled active and the remaining 420 (49.01%) inactive. An overview of the distribution of pDC_{50} and D_{max} values is shown in Fig. 4(a). We can see that the majority of the data samples are normally concentrated around the pDC_{50} threshold of 6.94, with a few outliers. D_{max} values, on the other hand, are more spread out, with roughly half of the samples showing a D_{max} above 60%.

Fig. 4(b) shows the distribution of E3 ligases and their frequency in the dataset, together with the percentage of active/inactive samples associated with each of them. PROTACs are equally distributed (roughly) among the two main E3 ligases, cereblon (CRBN) and von Hippel–Lindau (VHL), with a small fraction of PROTACs being evaluated with other E3 ligases. We see that CRBN and VHL are indeed the most common (53.4% and 40.1%, respectively), whereas 6% of the data samples report less common E3 ligases: IAP (2.80%), MDM2 (1.26%), cIAP1 (0.98%), XIAP (0.93%), FEM1B (0.37%), Ubr1 (0.09%), RNF114 (0.05%). Regarding the active samples distribution among E3 ligases, CRBN and VHL are quite balanced (49.5% and 56.1%, respectively), and FEM1B and Ubr1 are mostly associated with active samples. The less common MDM2, IAP, and cIAP1 are mostly associated with inactive samples.

3.3. Model performance

Fig. 5(a) reports the performance of the different models across the various studies. For each study, named after either the *standard*, *target*, or *similarity* split used, we show the mean validation accuracy and ROC-AUC scores of the five models trained during CV (one model per fold) with the best hyperparameters found. Additionally, the plots show the performance on the test set of three models trained per study with the best hyperparameters found in CV and different initial weights. For those models, we also report the mean of the test accuracy and ROC-AUC scores, alongside the test accuracy and ROC-AUC scores calculated using majority voting. A dummy model is included as a baseline, which always predicts the majority class in the training set.

The performance metrics derived from the standard CV split offer an upper bound for our model's capability, with a validation average/test average/test majority vote accuracy of 85.7%/79.1%/82.6% and a validation average/test average/test majority vote ROC AUC of 0.922/0.841/0.848. These results suggest an optimal scenario where the model has access to a diverse and representative sample of the data during training, maximizing its learning potential. The standard split serves as an upper bound estimate for model performance, as real-life scenarios generally require more constrained and specialized testing conditions.

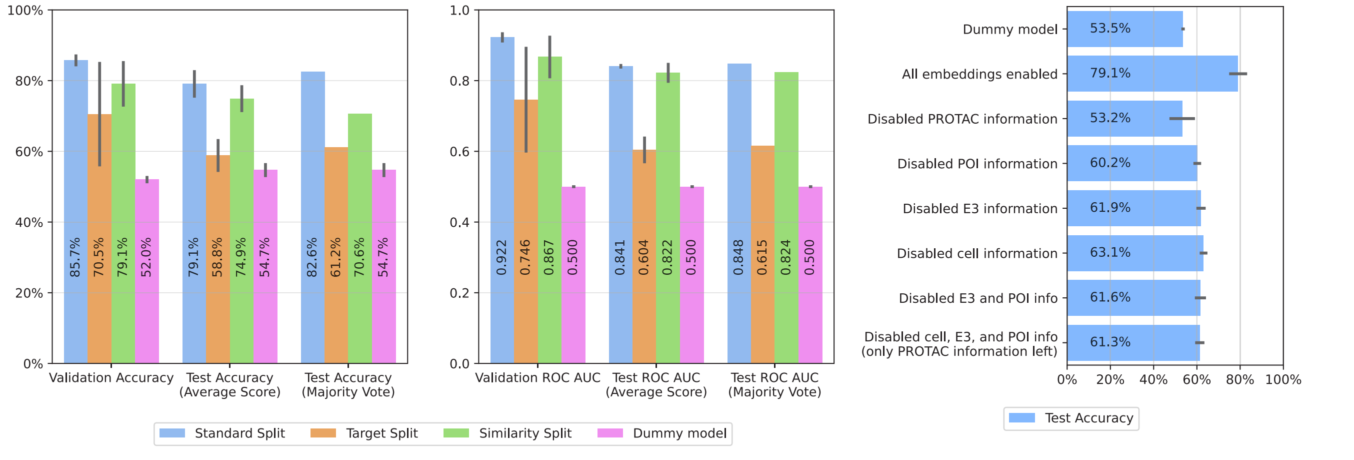
On the other hand, in the similarity CV split study, designed to evaluate the model's generalizability to unseen PROTAC compounds that do not share structural similarities with the training set, our model reached a remarkable validation average/test average/test majority vote accuracy of 79.1%/74.9%/70.6% and a validation average/test average/test majority vote ROC AUC of 0.867/0.822/0.824. The high performance in this study indicates the model's robust ability to extrapolate from known PROTACs to predict the activities of novel molecules.

Finally, the target split study presents a significant challenge for our model, as evidenced by the lower validation average/test average/test majority vote accuracy of 70.5%/58.8%/61.2% and a validation average/test average/test majority vote ROC AUC of 0.746/0.604/0.615. This study tests the model's ability to generalize across different protein targets, a critical factor for PROTAC design in novel disease mechanisms. The diminished performance suggests a need for improved protein representations or for embeddings that better capture more detailed and relevant features of the target proteins. Moreover, it underscores the necessity for more extensive and diverse datasets that include a broader array of PROTACs and targets.

Additional performance metrics are reported in Appendix A. Appendix D includes instead the performance scores of an XGBoost model evaluated on the aforementioned studies [21].

3.4. Ablation studies

The ablation study summarized in Fig. 5(b) highlights the contributions of various embeddings to model performance in PROTAC activity prediction. We focus on the average test accuracy of the three models trained with the best hyperparameters in the standard split study. With all embeddings enabled, the three models achieved an average test accuracy of 79.1%, serving as the baseline for full-feature utilization. Disabling cell, E3 ligase, and protein of interest (POI) embeddings individually led to varied decreases in performance, with test accuracies of 63.1%, 61.9%, and 60.2%, respectively. This highlights the importance of each type of embedding in enhancing predictive accuracy.



(a) (b)

Fig. 5. (a) Performance of the different models across the various studies, with model accuracy plotted on the left and ROC AUC plotted on the right. (b) Ablation results for the standard cross-validation split. Each bar indicates the embedding(s) not available to the model to process.

Notably, the model performance dropped below that of the dummy model when disabling compound information, emphasizing the importance of the PROTAC fingerprints. This is further highlighted by the test accuracy of the combination of disabled POI, E3, and cell embeddings (leaving the PROTAC information only), which reached 61.3%, close to other setups in which only a single component was disabled. In general, molecular fingerprints appear to be the most relevant input feature to the model. However, the general trend of high accuracy drops suggests that the contextual embeddings collectively contribute with significant predictive value beyond the structural information provided by molecular fingerprints alone.

Overall, this ablation study demonstrates the synergistic effect of integrating diverse embeddings, including compound structure (PROTAC fingerprint) and biological context (cell type, E3 ligase, POI), to capture the diverse determinants of biological activity in PROTACs.

4. Related work

The studies most closely aligned with our work are those of Li et al. [22] and Nori et al. [1]. Li et al. [22] introduces DeepPROTACs, a deep learning model for prognosticating PROTAC activity, whereas Nori et al. [1] proposes instead a LightGBM model for predicting protein degradation activity. LightGBM is a gradient boosting framework that uses a histogram-based approach for efficient, high-performance ML tasks [23].

The DeepPROTACs architecture encompasses multiple branches employing long short-term memory (LSTM) and graph neural network (GNN) components, all combined prior to a prediction head. Each branch processes distinct facets of the ternary complex, encompassing elements like E3 ligase and POI binding pockets, along with the individual components of the PROTAC: the warhead, linker, and E3 ligand. The model's performance culminates in an average prediction accuracy of 77.95% and a ROC-AUC score of 0.8470 on a validation set drawn from the PROTAC-DB. The LightGBM model, on the other hand, achieves a ROC-AUC of 0.877 on a PROTAC-DB test set with a much simpler model architecture and input representation.

Notwithstanding their achievements, the DeepPROTACs and LightGBM models both exhibit certain limitations. In DeepPROTACs, there is a potential risk of information loss as the PROTAC SMILES are partitioned into their constituent E3 ligands, warheads, and linkers, which are then fed into separate branches of the model. Secondly, while the authors undertake advanced molecular docking of the entire PROTAC-POI-E3 ligase complex, their subsequent focus on the 3D binding pockets of the POI and E3 ligase renders it less amenable for experimental replication and practical use. Finally, and perhaps most importantly, the potential for data leakage during hyperparameter optimization and its effects on out-of-distribution (OOD) generalization was not investigated. Data leakage between the different PROTAC components in the training and test sets of the model may artificially render a more accurate model that does not generalize well to new real-word data, necessitating more rigorous testing procedures. Because of that, generalization of the DeepPROTACs model would need to be further investigated on a separate test set.

5. Conclusions

In this work, we curated open-source PROTAC data and introduced a versatile toolkit for predicting PROTAC degradation effectiveness in three different experimental scenarios, aiming to assess the quality of our curated data and model generalizability. The performance of our models, achieving a top 82.6% test accuracy and a 0.848 ROC-AUC test score are competitive with, if not surpassing, existing methods for protein degradation prediction. Ours are also the first models to consider both DC_{50} and D_{max} in predicting degradation activity for PROTACs, a significant contribution as both properties are important to determining PROTAC efficacy. We show that our models can generalize well to unseen PROTACs, while struggling with unseen targets, highlighting the need for more comprehensive protein representations and more extensive datasets. Finally, our approach offers open-source accessibility, ease of reproducibility, and a less computationally complex alternative to previous work, making it a valuable resource for researchers working on data-driven approaches to PROTAC engineering.

Reproducibility statement and code availability

Code for this work is available at https://github.com/ribesstefano/PROTAC-Degradation-Predictor. The repository contains detailed instructions to reproduce the results presented in this work, including: the dataset curation process and the curated datasets, hyperparameter tuning, and the Optuna studies which can be used to train/retrieve all models presented herein.

CRediT authorship contribution statement

Stefano Ribes: Conceptualization, Data curation, Investigation, Methodology, Software, Validation, Visualization, Writing – original draft, Writing – review & editing. **Eva Nittinger:** Conceptualization, Methodology, Supervision, Validation, Writing – original draft, Writing

– review & editing. **Christian Tyrchan:** Conceptualization, Methodology, Supervision, Validation, Writing – original draft, Writing – review & editing. **Rocío Mercado:** Conceptualization, Funding acquisition, Methodology, Supervision, Validation, Writing – original draft, Writing – review & editing.

Declaration of competing interest

The authors declare the following financial interests/personal relationships which may be considered as potential competing interests: The author is an Editorial Board Member/Editor-in-Chief/Associate Editor/Guest Editor for *AI in the Life Sciences* and was not involved in the editorial review or the decision to publish this article.

Data availability

Data and code is available at https://github.com/ribesstefano/PROTAC-Degradation-Predictor.

Acknowledgments

The authors acknowledge computing resources from AstraZeneca and the National Academic Infrastructure for Supercomputing in Sweden (NAISS), partially funded by the Swedish Research Council through grant agreement no. 2022-06725. SR thanks the Health Engineering Area of Advance and the Gender Initiative for Excellence (Genie) for funding. RM thanks the Wallenberg AI, Autonomous Systems, and Software Program (WASP) for funding.

Appendix A. Prediction scores

This section provides a collection of all the validation and test scores computed for the evaluated models on the three proposed studies, always comparing to a dummy model that simply predicts the majority class in the training set. Figs. 6(a), 6(b), 6(c), 6(d), and 6(e) report validation and test accuracy, ROC-AUC, F1, precision, and recall scores, respectively, of the evaluated deep learning models in this work.

Appendix B. Cell line embeddings

This section details the methods used to extract cell line embedding vectors for our models. A basic approach is to assign a categorical (or one-hot encoded) label to each cell line in the dataset. While practical, this method ignores any inherent information about the cell lines and their biological similarity. To address this, we utilized the Cellosaurus database, which provides standardized information about common cell lines used in research [14]. Our approach involves isolating relevant biological information about each cell line into a text description. We then encode this text into an embedding vector by using a sentence Transformer model [17].

A sentence Transformer is designed to generate embedding representations of input sentences such that similar sentences have high cosine similarity. However, sentence Transformers have a fixed input size, accepting a maximum number of tokens. To process longer texts, we divide them into chunks of the maximum size, encode each chunk into a vector, and average the vectors into a single representation. To avoid diluting relevant information during this averaging process, we aim to summarize each cell line's information into concise, yet informative, short text descriptions.

We manually isolated columns containing relevant biological information about cells from the available database columns, such as their category (*e.g.*, “hybridoma”, “cancer cell line”, “transformed cell line”), sex (male or female), and species of origin (*e.g.*, “mus musculus”, “homo sapiens”, etc.). We discarded identification information, such as patents, synonyms, or entry dates. Additionally, Cellosaurus provides comments in various categories (*e.g.*, “monoclonal antibody target”, “sequence variation”, etc.), which we also included. The list of selected information is shown on the y-axis of Fig. 7(a).

Next, we ranked columns and comments based on the fraction of unique entries relative to their total, as illustrated in Fig. 7(a). Our intuition is that comments with a high number of unique entries help identify specific cell lines, making it easier to distinguish cell types. Following this principle and after reviewing examples, we selected the following information in this order: genome ancestry, karyotypic information, senescence, biotechnology, virology, caution, donor information, sequence variation, characteristics, transfected with, monoclonal antibody target, HLA typing, knockout cell, microsatellite instability, hierarchy (HI), breed/subspecies, derived from site, population, group, monoclonal antibody isotype, cell type, transformant, selected for resistance to, and category (CA).

Finally, for each database entry, we concatenated the strings from the selected information, removed PubMed references, and stripped extra spaces. The average text description length (*i.e.*, number of characters) of the cell lines in our curated dataset was 181.1, below the 384-token input size limit of the selected sentence Transformer model.

B.1. Cosine similarity of cell line descriptions

Table 2 presents a cosine similarity matrix for three cell line descriptions generated by following the above methodology. The cosine similarity metric quantifies the similarity between the textual descriptions of different cell lines, with values ranging from 0 to 1, where 1 indicates identical descriptions and 0 indicates no similarity.

For instance, the description of the cell line *UKF-NB-2rDACARB4* is highly similar to that of *UKF-NB-2rDOCE10*, with a cosine similarity of 0.8759. Both of these cell lines are cancer cell lines derived from the same species (*Homo sapiens*) and are part of the resistant cancer cell line (RCCL) collection. They differ primarily in their resistance to different chemotherapeutic agents: dacarbazine for *UKF-NB-2rDACARB4* and docetaxel for *UKF-NB-2rDOCE10*.

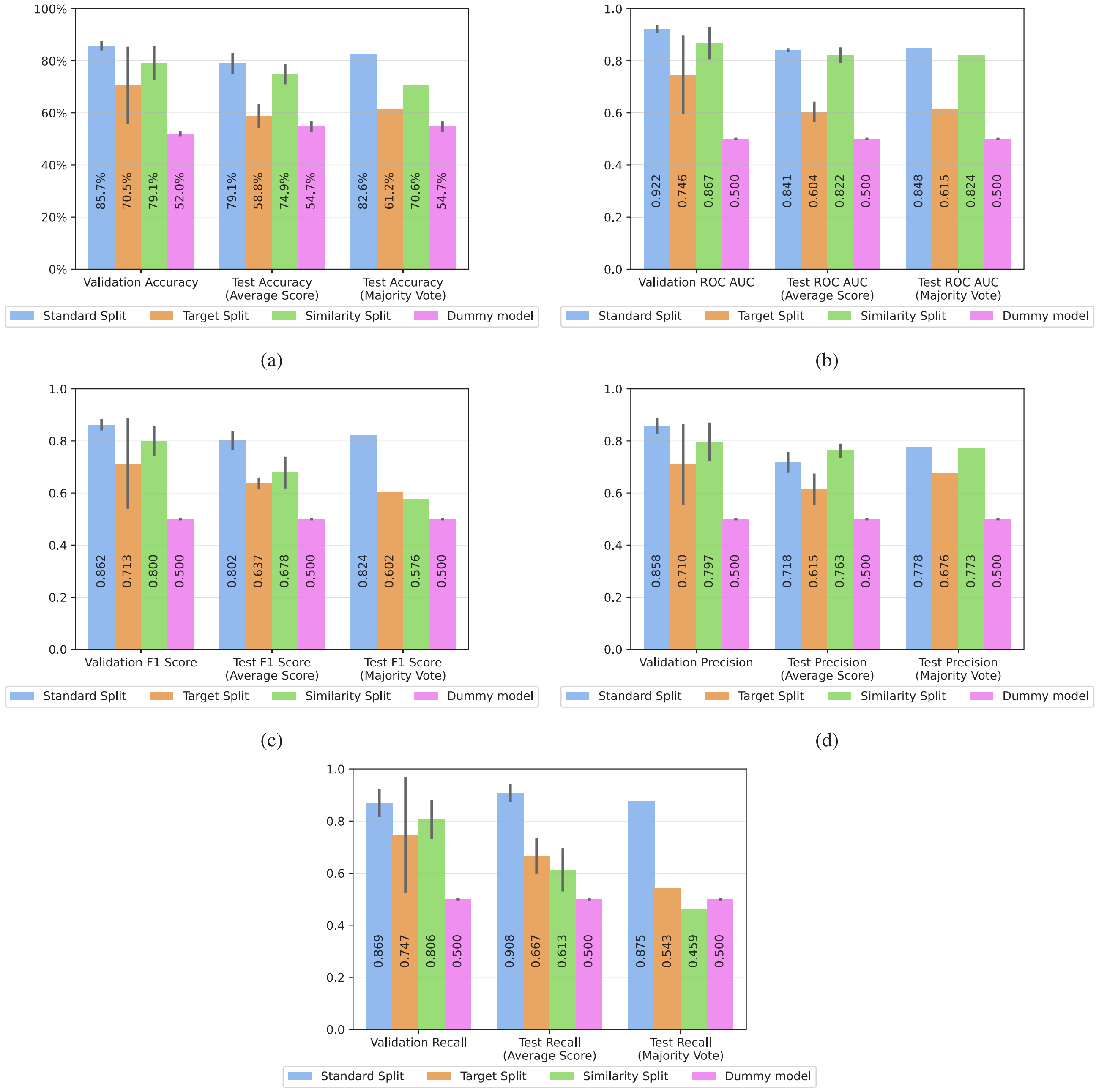
In contrast, the description of *FHS036i-sh18961C*, an induced pluripotent stem cell line, has a much lower similarity to the cancer cell lines, with cosine similarities of 0.2832 and 0.3522 to *UKF-NB-2rDACARB4* and *UKF-NB-2rDOCE10*, respectively. This lower similarity is expected given the fundamental differences in cell type, collection origin, and specific biological characteristics.

These examples illustrate how cosine similarity can effectively differentiate between cell lines based on their detailed descriptions, reflecting both broad classifications and specific attributes.

B.2. UMAP visualization of cell line embeddings

Fig. 7(b) presents a uniform manifold approximation and projection (UMAP) plot of the cell line embedding vectors. UMAP is a dimensionality reduction technique that helps visualize high-dimensional data by projecting it into a lower-dimensional space, preserving both local and global data structure [24].

The plot showcases the embedding vectors of cell lines, color-coded according to their categories. Each point represents a cell line, and its position reflects the similarity of its embedding vector to others. Similar cell lines cluster together, indicating that the embedding vectors effectively capture meaningful biological relationships. For instance, induced pluripotent stem cells (light purple) and hybridoma cell lines (light blue) form distinct, dense clusters, demonstrating the embeddings' ability to reflect their biological differences. In contrast, some categories, such as spontaneously immortalized cell lines (purple) and cancer cell lines (yellow), exhibit partial overlap, suggesting shared biological features while maintaining enough distinction to form identifiable subclusters. This visual validation underscores the embeddings' capacity to encapsulate and differentiate between various cell line categories, supporting the efficacy of our approach.



(a)

(b)

(c)

(d)

(e)

Fig. 6. Performance metrics for the presented deep learning models: (a) accuracy, (b) ROC-AUC, (c) F1 score, (d) precision, and (e) recall.

Appendix C. Datasets characteristics

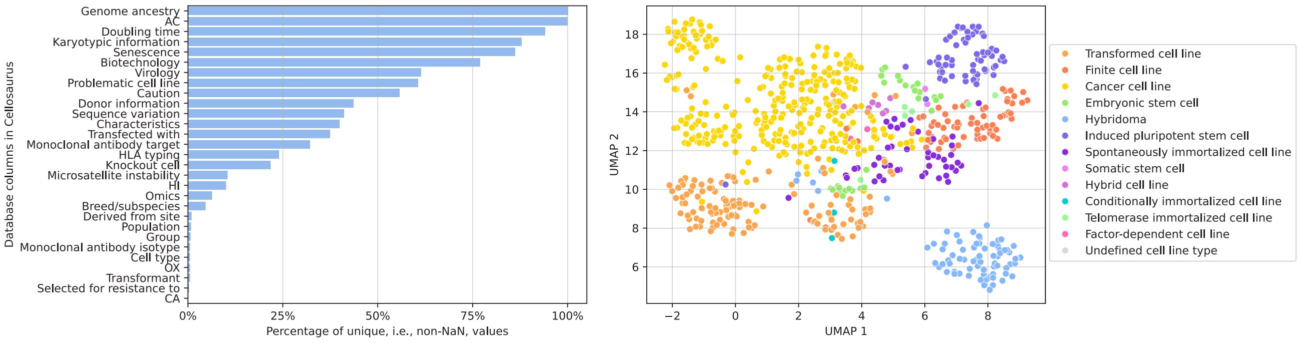
C.1. PROTAC-DB and PROTAC-Pedia

Table 3 provides an overview of the two datasets used in our study: PROTAC-DB and PROTAC-Pedia. PROTAC-DB contains a total of 5,388 entries, whereas PROTAC-Pedia comprises 1,203 entries. The number of unique SMILES in PROTAC-DB is 3,270, compared to 1,178 in PROTAC-Pedia. Unique targets in PROTAC-DB and PROTAC-Pedia are 323 and 79, respectively. A notable proportion of SMILES entries are shared between the datasets. Specifically, there are 1,222 SMILES entries that are found in both PROTAC-DB and PROTAC-Pedia. These shared SMILES are present in 22.7% of the total SMILES entries in PROTAC-DB and in 69.2% of the total SMILES entries in PROTAC-Pedia. The datasets also feature entries with SMILES that appear only once, *i.e.*, "single" SMILES, with 45.5% (2,451) of PROTAC-DB and 95.8% (1,153) of PROTAC-Pedia consisting of single SMILES. Single targets are relatively low in both datasets, at 1.4% (78) for PROTAC-DB and 1.2% (15) for PROTAC-Pedia, which is unsurprising as multiple PROTACs are generally investigated for a given target.

C.2. Cross-validation folds and test sets

Table 4 presents detailed statistics for the datasets used in the three studies proposed in our evaluation strategy.

For the standard split, each fold consists of approximately 616 training entries, 154 validation entries, and 86 test entries. The proportion of active data samples in these splits is consistent and balanced across



(a)

(b)

Fig. 7. (a) Cell line information (database columns) from Cellosaurus, ranked by their percentage of unique entries over the total number of entries in that column. (b) UMAP visualization of the generated cell line embedding vectors, color-coded by cell line categories.

Table 2
Example of cosine similarity matrix for three cell line descriptions.

Cell line text description	UKF-NB-2rDACARB4	UKF-NB-2rDOCE10	FHS036i-sh18961C
UKF − NB − 2rDACARB4: CVCL_RT02, Cancer cell line, NCBI_TaxID=9606; ! Homo sapiens (Human), Part of: Resistant Cancer Cell Line (RCCL) collection, Selected for resistance to: ChEBI; CHEBI:4305; Dacarbazine (DTIC; (5-(3,3-dimethyl-1-triazeno)imidazole-4-carboxamide)), Derived from site: Metastatic; Bone marrow; UBERON=UBERON_0002371, NCIt; C3270; Neuroblastoma, ORDO; Orphanet_635; Neuroblastoma, CVCL_9902 ! UKF-NB-2	1.0000	0.8759	0.2832
UKF − NB − 2rDOCE10: CVCL_RR83, Cancer cell line, NCBI_TaxID=9606; ! Homo sapiens (Human), Part of: Resistant Cancer Cell Line (RCCL) collection, Selected for resistance to: ChEBI; CHEBI:4672; Docetaxel anhydrous (Taxotere), Derived from site: Metastatic; Bone marrow; UBERON=UBERON_0002371, NCIt; C3270; Neuroblastoma, ORDO; Orphanet_635; Neuroblastoma, CVCL_9902 ! UKF-NB-2, Cancer cell line	0.8759	1.0000	0.3522
FHS036i − sh18961C: CVCL_YY67, Induced pluripotent stem cell, NCBI_TaxID=9606; ! Homo sapiens (Human), Part of: Framingham Heart Study (FHS) collection, Part of: Next Generation Genetic Association studies (Next Gen) program cell lines, Population: Caucasian, Sequence variation: Mutation; HGNC; 3231; CELSR2; Simple; c.*919G; dbSNP=rs12740374; Zygosity=Homozygous; Note=Major haplotype (PubMed=28388431), Omics: Transcriptome analysis by RNAseq, Derived from site: In situ; Peripheral blood; UBERON=UBERON_0000178, CVCL_YY66 ! FHS035i-sh18961A	0.2832	0.3522	1.0000

folds, with the training and validation sets containing around 51.4% active samples, and the test set 46.5%. Notably, a significant percentage of entries have leaking Uniprot identifiers (around 83%) and a smaller proportion have leaking SMILES (around 10%). The average Tanimoto distance between PROTACs in the test set is 0.381, indicating moderate structural similarity.

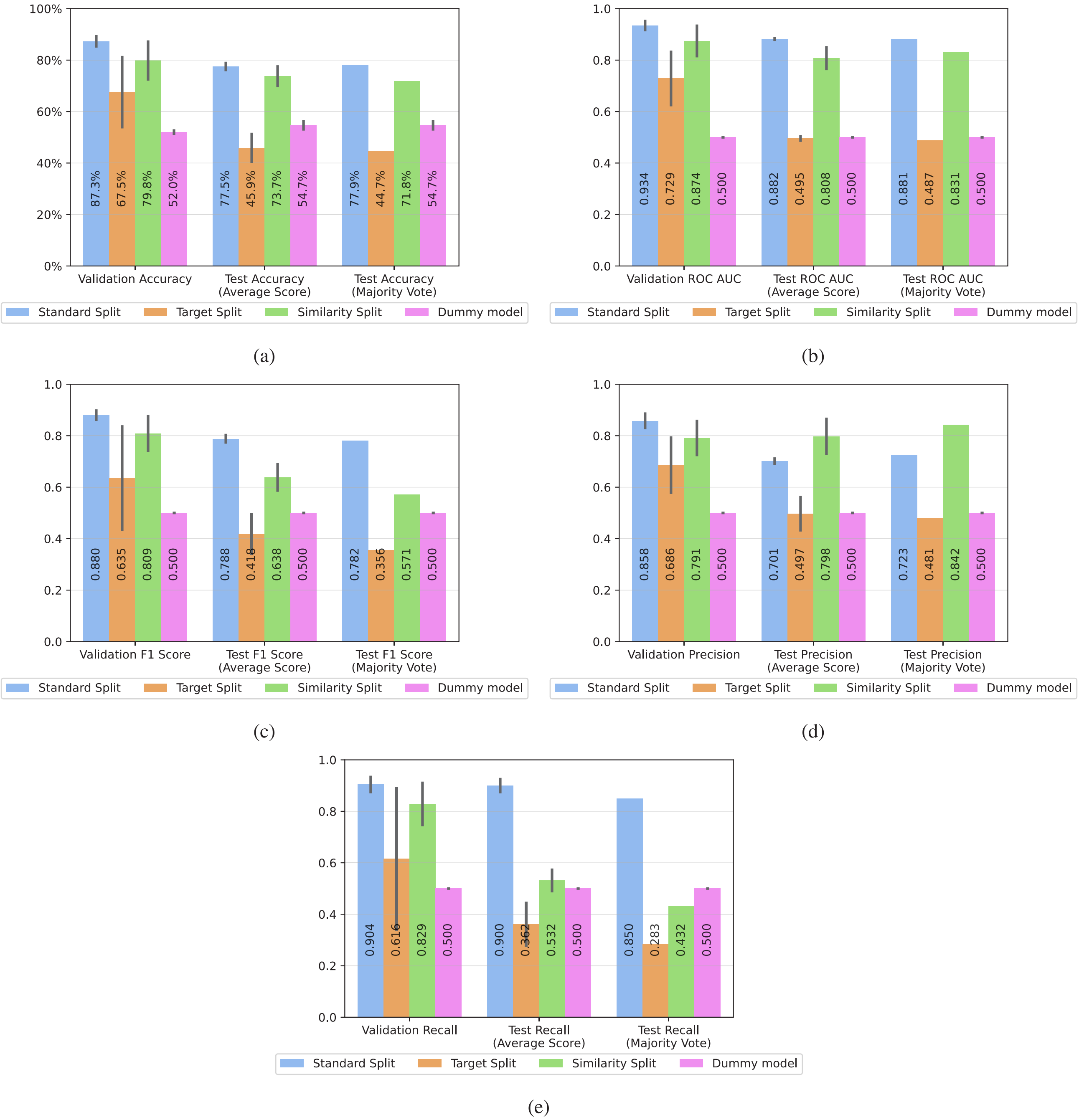
The target split aims to evaluate model generalization to unseen POIs. The training set sizes vary between 546 and 693, with the validation set sizes ranging from 79 to 226, and the test set consistently containing 85 entries. Because of stratified folds, the active data proportions in the training, validation, and test sets vary more widely than in the standard split. In fact, there are no leaking Uniprot identifiers in this split, and the proportion of leaking SMILES is below 1.5%. The average Tanimoto distance between PROTACs in the test set is slightly higher at 0.395.

For the similarity split, designed to test generalization to new PROTACs, the training set sizes range from 589 to 660, validation sets from 112 to 183, and the test set again consistently contains 85 entries. The active sample proportion in the training sets average around 51.5%, with the validation set showing slightly more variation. The leaking Uniprot identifiers are around 57%, and there are no leaking SMILES, by construction. The average Tanimoto distance between PROTACs in the test set is the highest among the splits at 0.420, reflecting the structural novelty of the test PROTACs in this specific study.

Appendix D. XGBoost performance

Given the experimental setup and evaluation strategy described in Section 2.4, we first trained different XGBoost models in a CV setting via Optuna. We then trained, with the best hyperparameters found, three models and evaluated them on the held-out test sets. As with the deep learning models, we evaluated the XGBoost models both individually by computing their average performance, and together via majority voting. Fig. 8 compares the performance metrics for the trained XGBoost models on the different studies.

The comparison of test performances between the trained XGBoost models and the proposed deep learning models highlights some key differences across the various studies. In the standard random split,



(a)

(b)

(c)

(d)

(e)

Fig. 8. Performance metrics for the XGBoost models: (a) accuracy, (b) ROC-AUC, (c) F1 score, (d) precision, and (e) recall.

Table 3
Characteristics of PROTAC-DB and PROTAC-Pedia datasets. The term *single* here indicates entries for which the SMILES or target appears only once in the corresponding dataset.

Dataset	Total entries	Unique SMILES	Unique targets	Shared SMILES	Shared SMILES %	Single SMILES	Single SMILES %	Single targets	Single targets %
PROTAC-DB	5388	3270	323	1222	22.7%	2451	45.5%	78	1.4%
PROTAC-Pedia	1203	1178	79	832	69.2%	1153	95.8%	15	1.2%

deep learning models achieve slightly higher test accuracies (up to 82.56%) compared to XGBoost (up to 79.07%). For the target split, deep learning models outperform XGBoost with test accuracies ranging from 55.29% to 63.53%, while XGBoost's performance is significantly lower and more variable, ranging from 41.18% to 51.76%. In the similarity split, deep learning models again show better performance with accuracies reaching up to 78.82% compared to XGBoost's 76.47%.

Regarding ROC-AUC scores, in the standard split, both models perform robustly, but XGBoost has slightly higher scores (up to 0.884) compared to deep learning's 0.848. In the target split, deep learning models have a clear advantage with ROC-AUC scores up to 0.633, while XGBoost's scores hover around 0.5, indicating poor performance. In the similarity split, deep learning models again demonstrate better performance with ROC-AUC scores up to 0.850, compared to XGBoost's 0.836. Overall, deep learning models generally show superior or comparable test performance, especially in the target and similarity splits.

Table 4
Statistics of datasets used in different studies. The term *leaking* indicates the percentage of entries in the training set with either a SMILES or target that also appears in the test set data samples. The *avg Tanimoto distance* refers to the average Tanimoto distance between PROTACs in the test set.

Fold	Study split	Train size	Val size	Test size	Train active %	Val active %	Test active %	Leaking Uniprot %	Leaking SMILES %	Avg Tanimoto distance
0	Standard	616	155	86	51.5%	51.6%	46.5%	82.5%	11.2%	0.381
1	Standard	617	154	86	51.4%	51.9%	46.5%	84.0%	10.2%	0.381
2	Standard	617	154	86	51.5%	51.3%	46.5%	83.8%	9.4%	0.381
3	Standard	617	154	86	51.5%	51.3%	46.5%	82.3%	10.4%	0.381
4	Standard	617	154	86	51.5%	51.3%	46.5%	83.8%	10.0%	0.381
0	Target	560	212	85	54.5%	40.6%	54.1%	0.0%	1.1%	0.395
1	Target	627	145	85	51.7%	46.2%	54.1%	0.0%	0.8%	0.395
2	Target	662	110	85	50.6%	50.9%	54.1%	0.0%	1.2%	0.395
3	Target	546	226	85	48.4%	56.2%	54.1%	0.0%	1.5%	0.395
4	Target	693	79	85	48.5%	69.6%	54.1%	0.0%	1.3%	0.395
0	Similarity	660	112	85	51.5%	53.6%	43.5%	57.7%	0.0%	0.420
1	Similarity	589	183	85	49.7%	58.5%	43.5%	56.4%	0.0%	0.420
2	Similarity	616	156	85	54.2%	42.3%	43.5%	57.3%	0.0%	0.420
3	Similarity	598	174	85	52.8%	48.3%	43.5%	56.5%	0.0%	0.420
4	Similarity	625	147	85	50.7%	56.5%	43.5%	57.0%	0.0%	0.420

References

[1] Nori D, Coley CW, Mercado R. De novo PROTAC design using graph-based deep generative models. 2022, arXiv preprint arXiv:2211.02660.

[2] Mercado R, Rastemo T, Lindelöf E, Klambauer G, Engkvist O, Chen H, et al. Graph networks for molecular design. Mach Learn: Sci Technol 2021;2(2):025023.

[3] Blaschke T, Arús-Pous J, Chen H, Margreitter C, Tyrchan C, Engkvist O, et al. REINVENT 2.0: an AI tool for de novo drug design. J Chem Inf Model 2020;60(12):5918–22.

[4] Wu Z, Zhu M, Kang Y, Leung ELH, Lei T, Shen C, et al. Do we need different machine learning algorithms for QSAR modeling? A comprehensive assessment of 16 machine learning algorithms on 14 QSAR data sets. Brief Bioinform 2021;22(4):bbaa321.

[5] Liu J, Ma J, Liu Y, Xia J, Li Y, Wang ZP, et al. PROTACs: a novel strategy for cancer therapy. In: Seminars in cancer biology, vol. 67. Elsevier; 2020, p. 171–9.

[6] Tomoshige S, Ishikawa M. PROTACs and other chemical protein degradation technologies for the treatment of neurodegenerative disorders. Angew Chem, Int Ed 2021;60(7):3346–54.

[7] Hu Z, Crews CM. Recent developments in PROTAC-mediated protein degradation: From bench to clinic. ChemBioChem 2022;23(2):e202100270.

[8] Békés M, Langley DR, Crews CM. PROTAC targeted protein degraders: the past is prologue. Nat Rev Drug Discov 2022;21(3):181–200.

[9] Gesztelyi R, Zsuga J, Kemeny-Beke A, Varga B, Juhasz B, Tosaki A. The Hill equation and the origin of quantitative pharmacology. Arch Hist Exact Sci 2012;66(4):427–38. http://dx.doi.org/10.1007/s00407-012-0098-5.

[10] Mostofian B, Martin HJ, Razavi A, Patel S, Allen B, Sherman W, et al. Targeted protein degradation: Advances, challenges, and prospects for computational methods. J Chem Inf Model 2023;63(17):5408–32. http://dx.doi.org/10.1021/acs.jcim.3c00603.

[11] Weng G, Shen C, Cao D, Gao J, Dong X, He Q, et al. PROTAC-DB: an online database of PROTACs. Nucleic Acids Res 2021;49(D1):D1381–7.

[12] London N, Prilusky J. PROTACpedia. 2024, https://protacpedia.weizmann.ac.il/. [Accessed 21 May 2024].

[13] Landrum G. rdkit.Chem.rdmolops module — The RDKit 2023.03.1 documentation. 2010, URL https://www.rdkit.org/docs/source/rdkit.Chem.rdmolops.html.

[14] Bairoch A. The cellosaurus, a cell-line knowledge resource. J Biomol Tech 2018;29(2):25.

[15] EMBL-EBI. UniProt. 2023, URL https://www.uniprot.org/.

[16] Dallago C, Schütze K, Heinzinger M, Olenyi T, Littmann M, Lu AX, et al. Learned embeddings from deep learning to visualize and predict protein sets. Curr Protocols 2021;1(5):e113. http://dx.doi.org/10.1002/cpz1.113, URL https://currentprotocols.onlinelibrary.wiley.com/doi/abs/10.1002/cpz1.113.

[17] Reimers N, Gurevych I. Sentence-BERT: Sentence embeddings using Siamese BERT-networks. In: Proceedings of the 2019 conference on empirical methods in natural language processing. Association for Computational Linguistics; 2019, URL https://arxiv.org/abs/1908.10084.

[18] Chawla NV, Bowyer KW, Hall LO, Kegelmeyer WP. SMOTE: synthetic minority over-sampling technique. J Artif Intell Res 2002;16:321–57.

[19] Akiba T, Sano S, Yanase T, Ohta T, Koyama M. Optuna: A next-generation hyperparameter optimization framework. In: Proceedings of the 25th ACM SIGKDD international conference on knowledge discovery & data mining. 2019, p. 2623–31.

[20] Khan S, Zhang X, Lv D, Zhang Q, He Y, Zhang P, et al. A selective BCL-XL PROTAC degrader achieves safe and potent antitumor activity. Nat Med 2019;25(12):1938–47.

[21] Chen T, He T, Benesty M, Khotilovich V, Tang Y, Cho H, et al. XGBoost: extreme gradient boosting. 2015, p. 1–4, R package version 0.4-2, 1 (4).

[22] Li F, Hu Q, Zhang X, Sun R, Liu Z, Wu S, et al. DeepPROTACs is a deep learning-based targeted degradation predictor for PROTACs. Nature Commun 2022-11-21;13(1):7133. http://dx.doi.org/10.1038/s41467-022-34807-3, URL https://www.nature.com/articles/s41467-022-34807-3 Number: 1 Publisher: Nature Publishing Group.

[23] Ke G, Meng Q, Finley T, Wang T, Chen W, Ma W, et al. LightGBM: A highly efficient gradient boosting decision tree. Adv Neural Inf Process Syst 2017;30.

[24] McInnes L, Healy J, Melville J. Umap: Uniform manifold approximation and projection for dimension reduction. 2018, arXiv preprint arXiv:1802.03426.