

TACK: A Statistical Evaluation of Degradation Activity on a Novel TArgeting Chimeras Knowledge Dataset

Downloaded from: https://research.chalmers.se, 2026-11-09 00:29 UTC

Citation for the original published paper (version of record):

Ribes, S., Dunlop, N., Mercado, R. (2026). TACK: A Statistical Evaluation of Degradation Activity on a Novel TArgeting Chimeras Knowledge
Dataset. Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2-B: 11912-11923. http://dx.doi.org/10.1145/3770855.3819052

N.B. When citing this work, cite the original published paper.

research.chalmers.se offers the possibility of retrieving research publications produced at Chalmers University of Technology. It covers all kind of research output: articles, dissertations, conference papers, reports etc. since 2004. research.chalmers.se is administrated and maintained by Chalmers Library

(article starts on next page)

TACK: A Statistical Evaluation of Degradation Activity on a Novel TArgeting Chimeras Knowledge Dataset

Stefano Ribes
ribes@chalmers.se
Department of Computer Science and Engineering Chalmers University of Technology and University of Gothenburg
Gothenburg, Sweden

Nils Dunlop*
nilsdu@chalmers.se
Department of Computer Science and Engineering Chalmers University of Technology and University of Gothenburg
Gothenburg, Sweden

Rocío Mercado
rocio.mercado@chalmers.se
Department of Computer Science and Engineering Chalmers University of Technology and University of Gothenburg
Gothenburg, Sweden

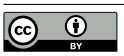
Abstract

Proteolysis-targeting chimeras (PROTACs) represent a promising therapeutic modality that induces targeted protein degradation by hijacking the ubiquitin-proteasome system. However, rational PROTAC design remains challenging due to the complex interplay between molecular structure, target proteins, E3 ligases, and the cellular context. We present TACK, a statistical evaluation of degradation activity on a novel TArgeting Chimeras Knowledge dataset of 3,514 PROTACs and 6,561 degradation endpoints aggregated from three major repositories with standardized molecular representations, protein annotations, and experimental conditions. Using scaffold-based 5×5 cross-validation, we perform a rigorous statistical comparison of three machine learning methods to predict PROTAC degradation activity across three tasks: DC_{50} and D_{max} regression, and binary activity classification. Feature ablation demonstrates that cellular context features and simple protein representations rival complex ESM protein embeddings, highlighting the importance of feature engineering over architectural sophistication. Models trained on the best performing features show that potency (pDC_{50}, $R^2 = 0.66$) is substantially more predictable than maximum degradation (D_{max}, $R^2 = 0.36$). In activity prediction, statistical tests support that classical methods (XGBoost and MLP) significantly outperform PROTAC-STAN, a domain-specific graph neural network model (ROC-AUC: 0.85 vs. 0.75, $p < 0.001$). Finally, we propose an ensemble-based uncertainty quantification approach showing that prediction variance correlates with prediction error (pDC_{50}: Spearman $\rho = 0.36$, $p < 0.001$; D_{max}: $\rho = 0.69$, $p < 0.001$), enabling confidence-aware experimental prioritization. Our findings challenge assumptions about specialized architectures for degradation prediction and provide evidence-based guidance for ML-driven PROTAC assessment.

CCS Concepts

• **Applied computing** → **Chemistry**; **Mathematics and statistics**; • **Computing methodologies** → **Machine learning**.

*The first two authors contributed equally to this research.



This work is licensed under a Creative Commons Attribution 4.0 International License.
KDD '26, Jeju Island, Republic of Korea
© 2026 Copyright held by the owner/author(s).
ACM ISBN 979-8-4007-2259-2/2026/08
https://doi.org/10.1145/3770855.3819052

Keywords

PROTAC; Protein Degradation; Machine Learning; Dataset Curation; Statistical Comparison

ACM Reference Format:
Stefano Ribes, Nils Dunlop, and Rocío Mercado. 2026. TACK: A Statistical Evaluation of Degradation Activity on a Novel TArgeting Chimeras Knowledge Dataset. In *Proceedings of the 32nd ACM SIGKDD Conference on Knowledge Discovery and Data Mining V.2 (KDD '26), August 09–13, 2026, Jeju Island, Republic of Korea.* ACM, New York, NY, USA, 12 pages. https://doi.org/10.1145/3770855.3819052

1 Introduction

Proteolysis-targeting chimeras (PROTACs) represent a paradigm shift in drug discovery, offering a catalytic mechanism to eliminate disease-relevant proteins from cells [4, 30]. By recruiting an E3 ubiquitin ligase to a protein of interest (POI), PROTACs harness the cell's native degradation machinery to achieve therapeutic effects at lower doses and against targets long considered undruggable. Nevertheless, rational design of effective degraders remains challenging due to the complex, non-additive relationship between PROTAC components (warhead, linker, E3 ligand) and degradation efficacy.

Machine learning (ML) offers a compelling path forward for accelerating PROTAC property prediction[17], yet three critical gaps hinder progress: *data scarcity*—public datasets contain only a few thousand annotated PROTACs, with >80% of database entries lacking key activity metrics (DC_{50}, D_{max}); *lack of standardized benchmarks*—heterogeneous activity thresholds, data splits, and evaluation protocols make fair model comparison difficult; *limited generalizability*—many state-of-the-art models developed for small molecules do not often generalize well to the PROTAC space. Existing databases, including PROTAC-DB [16], PROTACpedia [25], PROTAC-Patent-DB [8], and TPDdb [28], provide valuable structural and activity data but suffer from inconsistent annotations and coverage, directly impacting ML model training and generalization.

Degradation activity prediction can be framed as either binary classification or regression. Classification enables compound filtering in virtual screening, whereas regression supports compound ranking during lead optimization. Practical tools should support both paradigms. Here we introduce TACK, a curated, ML-ready benchmark for PROTAC degradation activity prediction, addressing the three aforementioned gaps:

- The *TACK dataset* harmonizes data from multiple sources with consistent activity thresholds, SMILES standardization, and rigorous quality filters.

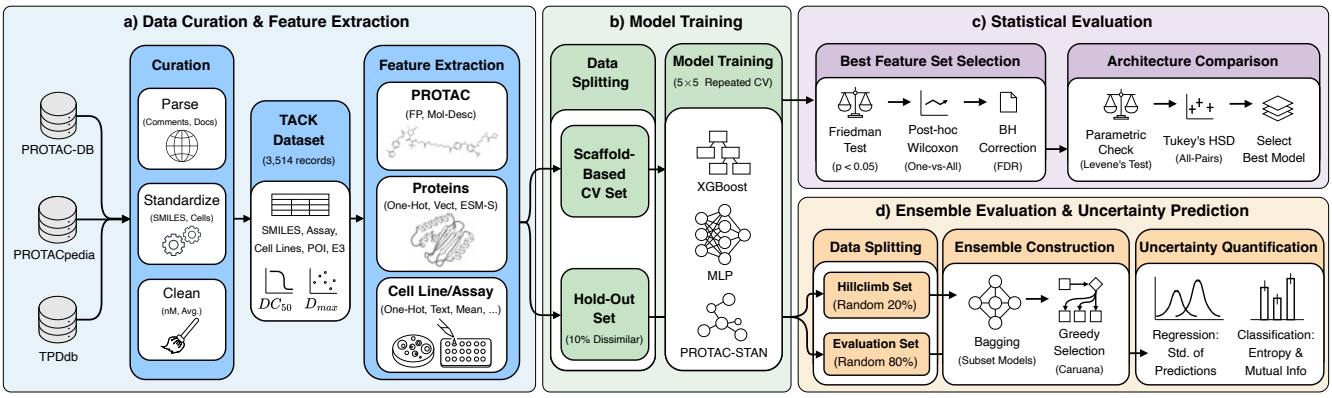


Figure 1: Pipeline summarizing our methodology. (a) Data curation includes merging and cleaning open-source databases to obtain TACK. (b) The TACK dataset is partitioned into a hold-out set, while the rest is used to train models in repeated cross-validation (CV). (c) The models are statistically compared to determine the best feature set and architecture. (d) Ensembles of CV models are used to improve performance and provide epistemic uncertainty.

- The *TACK benchmark* provides a rigorous statistical evaluation of state-of-the-art models, revealing performance gaps and failure modes.
- The *TACK ensemble* model is a fast, open-source predictor supporting both classification and regression.

All code and data are publicly available at this link. This work bridges AI & pharmaceutical science, demonstrating how careful data curation and rigorous benchmarking can advance ML-driven drug discovery for a new therapeutic modality.

2 Background

PROTACs were first established as a proof-of-concept by Sakamoto et al. [30] in 2001 using peptide-based recruiters, followed by the first all-small-molecule PROTAC in 2008 [31]. With the field rapidly expanding into clinical trials [12], there is a need for computational tools to screen the vast combinatorial space of linkers, warheads, and E3 ligands that make up PROTACs. Early computational efforts relied on physics-based modeling and molecular simulation to predict stable ternary complex formation [13, 38]. While detailed, these methods are too intensive for large-scale library screening, motivating the shift towards ML-based approaches.

Initial predictive models focused on classic ML approaches. Schneider et al. [32] used random forests and gradient boosting machines with molecular descriptors and docking scores as input features for small ligands targeting ERα. Similarly, Nori et al. [26] used LightGBM as a reward function within a generative framework. Feature importance analysis consistently identifies PROTAC molecular weight, topological polar surface area, and hydrogen-bond occupancies as key predictors of degradation potency. Recently, deep learning (DL) approaches have tried to learn degradation drivers directly from raw molecular representations. Ribes et al. [29] introduced a framework using 1D and 2D embeddings without 3D dependencies, achieving a test accuracy of 82.6% for predicting degradation activity. These methods show that high predictive performance is achievable without computationally expensive docking scores, marking a shift toward scalability.

Despite the success of these models, recent work has reintegrated structural insights to capture ternary complex intricacies. DeepPROTACs [24] introduced a graph neural network to encode molecular graphs of ligands and generated protein pockets, alongside recurrent neural networks for linker SMILES encoding. However, the model relies on protein-ligand structures, limiting its applicability when high-quality 3D data is sparse. Recently, PROTAC-STAN [11] combined a hierarchical PROTAC encoder with structure-informed POI & E3 embeddings derived from the protein language model ESM-S [39]. Using a ternary attention network, it explicitly fuses POI–PROTAC–E3 representations, achieving 88.4% accuracy on its refined dataset, outperforming DeepPROTACs & Ribes et al. [29].

Despite these architectural advances, current approaches typically frame PROTAC activity prediction as a binary classification task, simplifying experimental values into 'active' or 'inactive' labels based on thresholds (commonly DC_{50} <100 nM or D_{max} >80%) [11, 24, 29]. This thus fails to distinguish highly potent degraders from marginally active ones. Further, prior work has often relied on random data splitting, which risks overestimating performance on novel chemical structures due to scaffold similarity between train and test sets. Alongside methodological advances, the scale and quality of available data have evolved from scattered literature reports to centralized repositories. PROTAC-DB [36] was among the first resources, aggregating experimentally validated degraders from academic literature. Subsequently, the community-driven PROTACpedia [25] curated >1K degraders. Most recently, TPDdb [28] expanded the chemical space by mining patent disclosures, cataloging >22K PROTACs alongside emerging modalities like molecular glues. Together, these resources provide the data infrastructure to support the next generation of predictive models.

3 Methods

3.1 TACK Dataset Creation

3.1.1 Dataset Curation. To construct a comprehensive dataset for ML-based PROTAC degradation activity analysis (Figure 1a), we

Table 1: Overview of the curated TACK dataset vs. raw data organized by source (TPDdb, PROTAC-DB, PROTACpedia).

Data Overview				
Statistic	TPDdb	PROTAC-DB	PROTACpedia	TACK
Total Records	22,183	9,380	1,203	6,561
Unique PROTACs	21,429	6,110	1,189	3,514
Degradation Endpoints	6,518	2,170	580	6,561
POI Targets	184	441	79	164
E3 Ligases	8	21	8	9
Cell Lines	142	—	139	155
TACK Hold-out Set	774	92	53	913

aggregated data from three open-access repositories: 6,110 experimentally validated compounds from PROTAC-DB 3.0 [16], 1,189 PROTACs from PROTACpedia [25], and a collection of 21,429 PROTACs from TPDdb [28]. While the initial raw dataset exceeded 29K entries, our subsequent rigorous curation procedure, including filtering for valid DC_{50} & D_{max} values and chemical standardization, resulted in a high-quality dataset of 3,514 PROTACs.

Molecular Standardization. We implemented a robust data cleaning pipeline designed to standardize molecules and ensure dataset consistency. SMILES representations were canonicalized using RDKit [22]. POI and E3 ligase names were standardized with their corresponding UniProt identifiers; amino acid sequences for both sets of proteins were retrieved from UniProt where available [33].

Endpoint Standardization. DC_{50} & D_{max} endpoints required special attention due to heterogeneous reporting formats. Numeric values were extracted directly with appropriate unit conversion. All concentrations (DC_{50}) were converted to nanomolar units, and pDC_{50} to DC_{50}. Values with comparison operators (*e.g.*, ">100 nM") were treated as their numeric component. However, the operators ($<, >, \leq, \geq$) were stored separately to enable filtering during train-test splitting, allowing entries with inequalities to be excluded from evaluation sets. Range values (*e.g.*, "10-100 nM") were converted to their arithmetic mean while storing the original bounds. Categorical grades (A, B, C, D) for activity, common in patent disclosures, were excluded as they required patent-specific mapping to obtain quantitative values.

Assay Standardization. Cell line names were validated against the Cellosaurus database [3] to obtain standardized identifiers. In TPDdb, the online repository contains no metadata (*e.g.*, cell line, amino acid sequences, assay descriptions) in the downloadable datasets, so we developed a parsing script to extract additional data from the web pages. Assay descriptions were standardized (*e.g.*, "WB" to "Western Blot") to facilitate grouping by experimental method. For each D_{max} measurement, the treatment concentration was extracted from multiple sources including explicit metadata fields and parsed assay descriptions, where available. Patent information was sourced from the original patent database tables. For PROTAC-DB, the primary challenge was parsing assay descriptions that encoded multiple experimental conditions within single text entries. Descriptions such as "*Degradation of BRD4 short/long in HeLa cells after 24 h treatment*" required systematic extraction of target proteins, cell lines, and treatment times. We developed parsing functions to handle slash-separated multi-target entries, protein mutation annotations, and various time formats. Value columns containing multiple measurements (*e.g.*, "*0.081/0.14/0.53*") were split into individual entries. PROTACpedia required parsing of free-text comment fields where experimental values were embedded within narrative descriptions (*e.g.*, "*DC50 is 0.86 nM in LNCaP, 0.76 nM in VCaP*"). To extract cell-line-specific measurements, we implemented pattern-matching handlers for comments, which often described multi-protein degradation panels. Entries containing multiple targets within a single row were excluded for consistency.

Table 1 and Figure 2 summarize the characteristics of the TACK dataset. The aggregation and filtering processes resulted in 6,561 total experimental endpoints comprising 4,184 potency (DC_{50}) and 2,377 efficacy (D_{max}) measurements. The dataset covers a broad but consistent activity range in both the training/validation (Figure 2a) and held-out (Figure 2b) sets. We report the distribution of pDC_{50}, the negative logarithm (base 10) of DC_{50} in molar concentration, to better visualize potency. D_{max} values are mostly concentrated at high degradation levels, with a peak around 90–100%, while pDC_{50} values are more normally distributed with a slight skew towards lower potency. These trends indicate a bias towards effective degraders, though PROTAC potency varies widely. The dataset spans a diverse biological landscape totaling 164 distinct POIs and 155 cell lines, with the eight most common POIs, E3 ligases, and cell lines detailed in Figure 2c. Overall, activity labels are fairly balanced, with approximately 55% of entries classified as active in both the training/validation and held-out sets (Appendix A).

3.1.2 Scaffold Clustering and Hold-out Set. We constructed a hold-out test set isolating ∼10% of the curated TACK data based on structural dissimilarity. Using RDKit, we calculated 512-bit Morgan8 fingerprints for all PROTACs in TACK and calculated pairwise Tanimoto distances. We averaged the distances for each PROTAC to obtain a single dissimilarity score. The hold-out set was formed by selecting ∼10% of points with the highest dissimilarity score to simulate a realistic scenario in which models must predict degradation activity for novel PROTACs that differ structurally from those seen during training. This hold-out set was kept separate from the training and validation steps. For the remaining data, we clustered PROTACs using a scaffold-based grouping strategy based on Murcko scaffolds [5]. This grouping guarantees that during cross-validation (CV), no PROTACs sharing the same scaffold would appear in both training and validation sets, thereby providing a stringent evaluation of model generalization to unseen chemical structures. Compared to random splitting, which can lead to optimistic performance estimates, scaffold splitting offers a more realistic assessment of model capabilities in drug development [2].

3.2 Prediction Tasks

3.2.1 Task Definition. Our analysis focuses on three tasks related to degradation activity: (1) potency (pDC_{50}) regression, (2) efficacy (D_{max}) regression, and (3) binary classification of degradation activity. For activity classification, we define a data point as 'active' if it reports $D_{max} > 80\%$ & $DC_{50} < 100$ nM, following thresholds in Li et al. [24]. D_{max} & DC_{50} values are independently scaled with a quantile transformation [18] prior to regression. Since such normalization is data-dependent, we fit the transformation only on the training set and apply it to the validation and hold-out sets.

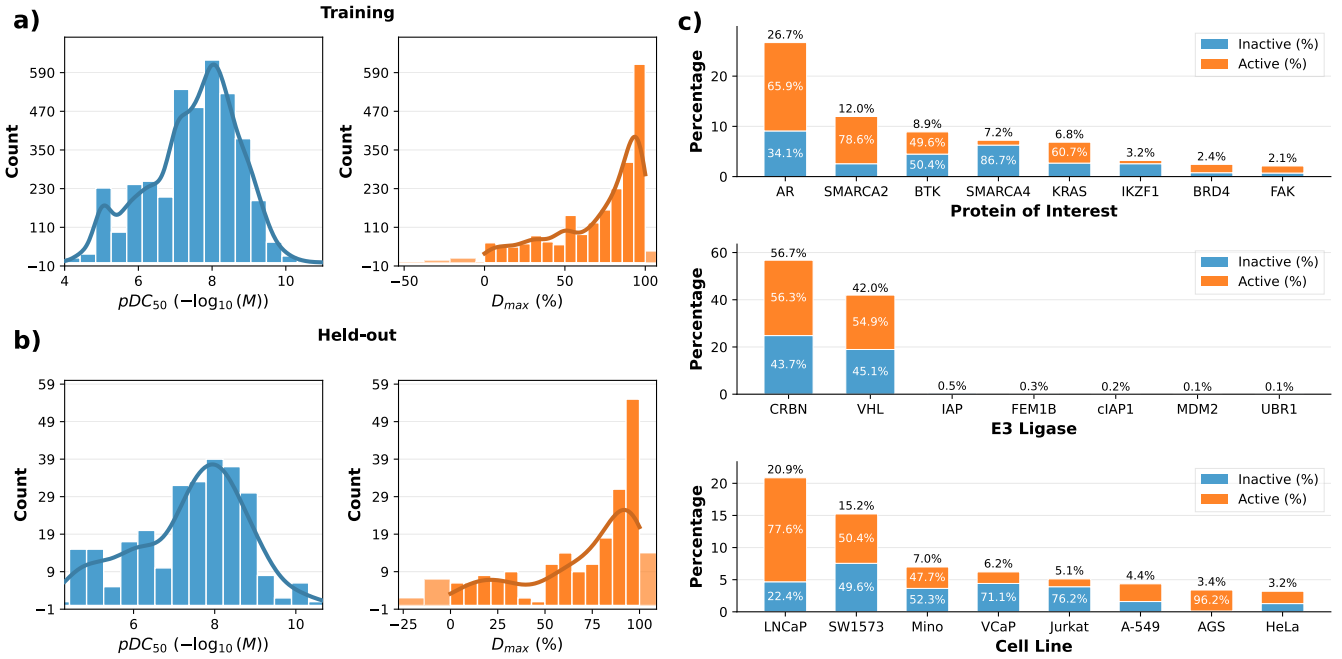


Figure 2: (a, b) Histograms of standardized degradation activity for the training and hold-out sets, showing the distributions of potency (pDC_{50}) and maximal degradation efficacy (D_{max}). Opaque bins for D_{max} were clipped between 0 and 100% for training and evaluation. (c) Stacked bar charts illustrating biological diversity, depicting the top 8 most represented POIs, E3 ligases, and cell lines with the proportions of active versus inactive measurements.

3.2.2 Feature Extraction. PROTAC ternary complexes consist of the POI, the E3 ligase, and the PROTAC molecule itself; we describe each of these three components by its own feature. For POIs and E3 ligases, we extract features based on their amino acid sequences: either as 1- and 2-grams with normalized counts (referred to as "Vec"), or as ESM-S precomputed embeddings [39]. We compress ESM-S embeddings via PCA to retain 90% of their variance, bringing the original 1280 embeddings dimension down to 44 and 7 for POI- and E3 ligase-related embeddings, respectively; we refer to these compressed embeddings simply as "ESM-S". For PROTACs, we compute either 512-bit Morgan8 fingerprints (referred to as "FP") or RDKit descriptor fingerprints (217 descriptors; referred to as "Desc") using RDKit. In some models, we use both Morgan8 and descriptor fingerprints, which we refer to as the "Desc+FP" feature set. Additionally, experimental conditions such as cell line, assay type, and treatment time influence degradation activity. For cell lines, we extract a textual description of the cell line and embed it using a pre-trained sentence transformer [29, Appendix B] (referred to as "Text"). The assay type (*e.g.*, Western blot, ELISA, HiBit), is encoded with one-hot encoding. The treatment time is instead encoded as a single numerical feature representing the duration of the assay treatment in hours. For assay conditions, we handle missing values as follows: embed the string "*Unknown cell line.*" for missing cell lines, impute missing treatment times with the mean of the training set, and use a zero vector for missing assay types. Finally, we also experimented with: a *simple* one-hot encodings of the POI gene name, the E3 ligase name, and/or the cell line (referred to as "OneHot" in all cases), and a representation that mimics the input features of PROTAC-STAN, *i.e.*, molecular fingerprints and ESM-S embeddings without applying PCA for POI and E3 ligase. The complete list of features is reported in Appendix D, Table 4.

3.3 Model

3.3.1 Selected Models. In this work, we evaluate three model architectures: XGBoost, a multi-layer perceptron (MLP), and PROTAC-STAN [11], the latter only used for the binary classification task as a literature baseline. Different combinations of feature sets (Section 3.2.2) are evaluated with both the XGBoost and MLP models. Following concatenation of the selected input features, all models predict a single scalar value, corresponding to either a regression or classification target. See Appendix C for model details.

3.3.2 CV. We perform a repeated group 5×5 CV (later referred to as *repeated CV*) on the TACK dataset, excluding the hold-out set, for a rigorous statistical evaluation as described in Ash et al. [2]. The inner CV loop splits the data into 5 folds, where each fold is set aside once for validation while the remaining 4 folds are used for training. The outer CV loop repeats this process 5 times with different random initialization seeds, leading to a total of 25 unique folds and 25 models trained per model class and task. As detailed in Section 3.1.2, groups are defined by the Murcko scaffolds of the PROTAC molecules. Figure 5, Appendix A, shows the distribution of the label values in the different folds. See Appendix E for a detailed

description of the hyperparameter optimization procedure using CV.

3.4 Statistical Evaluation

To compare model configurations and identify optimal architectures for each prediction task, we used a hierarchical statistical framework designed to control false discovery rates while maintaining statistical power across multiple comparisons. For regression tasks, we selected root mean squared error (RMSE) as our primary evaluation metric, while for the binary classification task, we use the area under the receiver operating characteristic curve (ROC-AUC). All metrics are computed on the 5×5 CV scheme, yielding 25 independent performance estimates per configuration.

3.4.1 Best Feature Set. A key challenge in applying ML to activity prediction is determining which combination of model and molecular/assay representations yields the best performance. We evaluated 10 feature combinations (Appendix D) for both XGBoost and MLP architectures across each of the three prediction tasks.

To identify top-performing configurations while accounting for variance across CV folds, we employed a rigorous two-stage statistical testing procedure. First, we applied the Friedman test [15] to assess whether significant performance differences exist among configurations. Upon detecting significant omnibus differences ($p < 0.05$), we conducted post-hoc pairwise comparisons using Wilcoxon signed-rank tests [37]: we identified the configuration with the best mean performance in the validation set as the control method and compared all other configurations against this top-performing control. We controlled the false discovery rate (FDR) using Benjamini-Hochberg (BH) correction [6] with $\alpha = 0.05$, chosen over family-wise error rate methods (*e.g.*, Bonferroni) for its superior statistical power in high-dimensional comparison settings [2]. Configurations not rejected by the BH-corrected tests were deemed statistically equivalent to the best method. This principled approach ensures that our feature recommendations generalize beyond random fold variations, an important consideration for practitioners applying these models to novel PROTAC designs.

3.4.2 Model Comparison. Having identified optimal feature configurations, we compared the XGBoost, MLP, and PROTAC-STAN models on the validation folds. For classification, when multiple feature sets were statistically equivalent, we selected the simplest configuration (*e.g.*, one-hot encodings over sequence embeddings) to prioritize computational efficiency and limit the number of comparisons, thereby avoiding inflated Type I error rates (*e.g.*, p-hacking).

The comparison procedure mirrored that used for feature selection, but in model comparison we only compare three architectures instead of 10 feature combinations, thus favoring the use of other statistical tests. We verified homogeneity of variances using Levene's test [23] and assessed normality visually, as the 5×5 CV metrics should normally distribute due to the central limit theorem (Appendix B). When assumptions were satisfied, we applied Tukey's HSD test [34], which provides tighter confidence intervals than non-parametric alternatives in small-sample comparisons.

3.5 Ensembling and Uncertainty Quantification

Reliable uncertainty estimates are critical for deploying ML models in scientific discovery workflows, where predictions guide expensive experimental validation. We construct an ensemble of the 5×5 CV models and quantify epistemic uncertainty—the uncertainty arising from limited training data and model specification—to help practitioners identify predictions requiring additional validation.

3.5.1 Ensemble Selection. We employed Caruana's greedy forward selection algorithm [10], which iteratively builds an ensemble by adding models that maximize performance on a dedicated selection set. To ensure unbiased evaluation, we partitioned the held-out set into a selection subset (19.9%) for ensemble construction and an evaluation subset (80.1%) for final assessment. For each task (pDC_{50} regression, D_{max} regression, and binary activity classification), we considered all 500 candidate models (10 feature configurations, both XGBoost and MLP models, × 25 CV folds), including lower-performing models to promote diversity.

The algorithm initializes with the top 5 models and iteratively adds the candidate that most improves selection set performance (RMSE for regression, log loss for classification). Models may be selected multiple times, with final weights normalized to sum to 1. We applied bagging (10 random samples of 50% of available models) to reduce sensitivity to selection set variability.

We evaluated two pools of candidate models for ensemble selection: (1) *model-level* selection, where candidates include all 500 individual CV fold models; and (2) *architecture-level* selection, where candidates are the 20 unique feature configurations with predictions pre-averaged across their respective CV folds. We compared against three baselines: the single best model, uniform averaging of all 500 models, and uniform averaging of the 25 CV folds for the best model (as defined by model type and feature set).

3.5.2 Uncertainty Quantification. We derive uncertainty estimates from inter-model disagreement within the ensemble. For regression, we calculated the standard deviation (σ) across member predictions. For classification, we calculate predictive entropy and decompose it to isolate mutual information as a measure of epistemic (model) uncertainty. To assess calibration, we measure Spearman correlation between uncertainty and absolute prediction error for regression tasks; well-calibrated models should exhibit positive correlation, indicating that uncertain predictions correspond to larger errors. For classification, we report expected and maximum calibration error (ECE/MCE) [27], quantifying the alignment between predicted probabilities and observed frequencies. These calibration metrics are essential for accurate compound prioritization using our models: miscalibrated confidence could lead to suboptimal resource allocation in PROTAC design campaigns.

4 Results

4.1 TACK Dataset

The TACK dataset is summarized in Table 1 & Figure 2. Among the 164 POI targets, the androgen receptor (AR, 26.7%), SMARCA2 (12.0%), and BTK (8.9%) are the most frequently studied, with the top five POIs accounting for 61.6% of all endpoints, reflecting the field's focus on oncology targets. E3 ligase diversity remains limited, with CRBN and VHL comprising 98.7% of all measurements. Cell

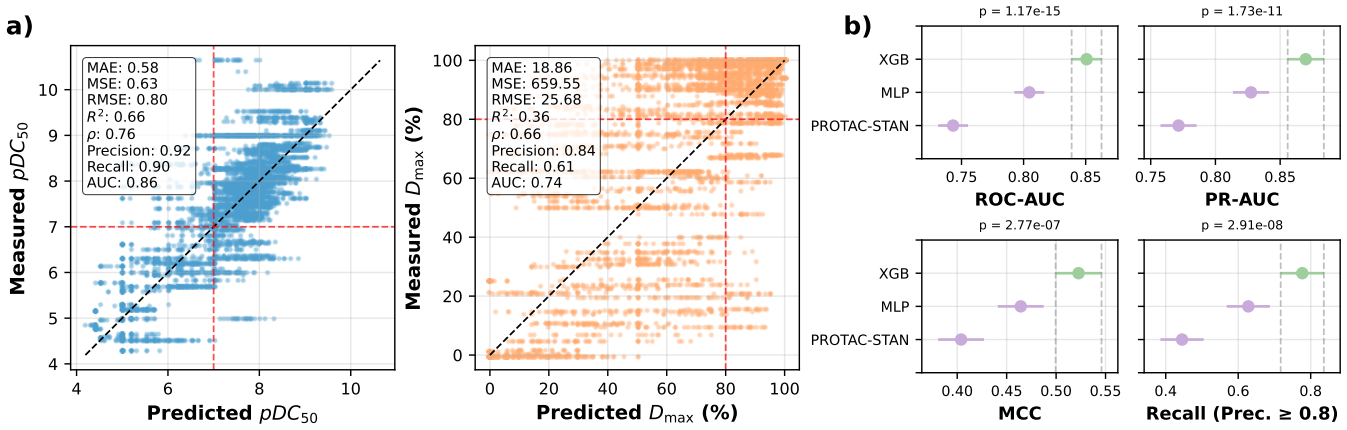


Figure 3: (a) Parity plots using MLP$^{(\dagger)}$ and XGBoost$^{(\dagger)}$ for pDC_{50} and D_{max}, respectively, on the hold-out test set (25 CV fold models). (b) Performance comparison for XGBoost$^{(\ddagger)}$, MLP$^{(\ddagger)}$, PROTAC-STAN on the binary activity classification task across four metrics. Points show means with 95% confidence intervals from 25 CV folds. Non-overlapping intervals indicate significant differences (Tukey HSD, $p < 0.05$). Omnibus ANOVA p-values shown above each panel. Precise effect size values are reported in Figure 6.

Table 2: Best feature set configurations for XGBoost and MLP models. We report the statistically equivalent feature combinations per task and per model. The metrics are calculated as the mean of the scores computed on the validation folds. Shorthand for feature sets is defined in Sec. 3.2.2.

Task	Model	Best Feature Set Configuration				Best Val. Mean
		Cell	E3	POI	Mol	
pDC_{50} (RMSE)	MLP$^{(\dagger)}$	**OneHot**	**OneHot**	**OneHot**	**Desc**	**0.770** (↓)
	MLP	OneHot	OneHot	Vec	Desc	0.787 (↓)
	XGB	Text	ESM-S-PCA	ESM-S-PCA	Desc	0.781 (↓)
	XGB	Text	OneHot	Vec	Desc+FP	0.785 (↓)
D_{max} (RMSE)	MLP	OneHot	OneHot	OneHot	Desc	20.589 (↓)
	MLP	OneHot	OneHot	Vec	Desc	20.689 (↓)
	XGB	Text	ESM-S-PCA	ESM-S-PCA	Desc+FP	18.452 (↓)
	XGB	Text	ESM-S-PCA	ESM-S-PCA	Desc	18.546 (↓)
	XGB	Text	OneHot	OneHot	Desc	18.514 (↓)
	XGB$^{(\dagger)}$	**Text**	**OneHot**	**Vec**	**Desc**	**18.340** (↓)
Activity (ROC-AUC)	MLP$^{(\ddagger)}$	OneHot	OneHot	OneHot	Desc	0.805 (↑)
	XGB$^{(\ddagger)}$	**Text**	**ESM-S-PCA**	**ESM-S-PCA**	**Desc+FP**	**0.851** (↑)
	XGB	Text	ESM-S-PCA	ESM-S-PCA	Desc	0.850 (↑)
	XGB	Text	OneHot	Vec	Desc	0.848 (↑)
	XGB	Text	OneHot	Vec	Desc+FP	0.847 (↑)
	XGB	Text	OneHot	OneHot	Desc	0.847 (↑)

(†): Models used to plot hold-out set predictions in Fig. 3a.
(‡): Models selected for architecture comparison with PROTAC-STAN in Fig. 3b.

line selection is more heterogeneous, with LNCaP (20.9%), SW1573 (15.2%), and Mino (7.0%) being the most common among 155 unique lines. Overall, 23.8% of PROTACs have multiple endpoints and 55.1% meet activity thresholds ($DC_{50} \leq 100$ nM and $D_{max} \geq 80\%$).

4.2 Statistical Evaluation Results

4.2.1 Best Feature Set. We evaluated feature configurations across XGBoost and MLP architectures using 5×5 CV on TACK, applying the statistical framework from Section 3.4.1. Table 2 summarizes configurations identified as statistically equivalent to the best-performing setup after Benjamini-Hochberg correction.

XGBoost. Friedman tests detected highly significant differences across feature configurations for pDC_{50} ($\chi^2 = 174.77$, $p = 6.2 \times 10^{-33}$), D_{max} ($\chi^2 = 131.41$, $p = 6.1 \times 10^{-24}$), and binary classification ($\chi^2 = 122.2625$, $p = 4.6 \times 10^{-22}$). The optimal XGBoost configuration for pDC_{50} combined cell line Text embeddings, ESM-S-PCA encodings for both E3 ligase and POI, and Desc for PROTACs (val. RMSE = 0.781). For D_{max} prediction, multiple XGBoost feature sets achieved statistical equivalence after Benjamini-Hochberg correction, with the best employing cell Text, E3 OneHot, POI Vec, and Desc (RMSE = 18.34). On binary classification, BH-corrected tests yielded four statistically equivalent XGBoost configurations (val. ROC-AUC = 0.847 – 0.851), with the best using ESM-S-PCA protein embeddings (cell: Text, E3 and POI: ESM-S-PCA, mol: Desc+FP; ROC-AUC = 0.851).

MLP. MLPs showed distinct patterns across tasks. For pDC_{50} , the Friedman test indicated significant configuration differences ($\chi^2 = 129.77$, $p = 1.3 \times 10^{-23}$), with two statistically equivalent feature sets emerging, both using simple OneHot encodings. Notably, the best MLP configuration (cell: OneHot, E3: OneHot, POI: OneHot, mol: Desc; RMSE = 0.770) slightly outperformed the best XGBoost setup by 1.4%. For D_{max} , two MLP configurations survived post-hoc testing ($\chi^2 = 105.25$, $p = 1.4 \times 10^{-18}$), with the best achieving RMSE = 20.59—exhibiting 12.3% higher error than the best XGBoost alternative. For binary classification, Friedman tests ($\chi^2 = 102.95$, $p = 4.0 \times 10^{-18}$) identified a unique best MLP configuration (cell: OneHot, E3: OneHot, POI: OneHot, mol: Desc; ROC-AUC = 0.805), trailing the top XGBoost model by 5.4%.

4.2.2 Generalization to Hold-out Set. To assess generalization performance, we evaluated the best performing MLP and XGBoost configurations (Table 2, marked with †) on the hold-out test set, aggregating predictions from all 25 CV fold models (Figure 3a). For prediction of pDC_{50}, the MLP model achieved on average strong performance (test MAE = 0.58, MSE = 0.63, $R^2 = 0.66$, Spearman $\rho = 0.76$), with the parity plot showing well-calibrated predictions

throughout the entire pDC_{50} concentration range and minimal systematic bias. When using a binary activity threshold of 100 nM, as reported in DeepPROTACs [24], the model exhibited high sensitivity in identifying active compounds (recall = 0.90) while maintaining high specificity (precision = 0.92, ROC-AUC = 0.86), indicating effective separation of potent degraders from weak or inactive compounds. For D_{max} predictions, XGBoost showed substantially weaker performance (MAE = 18.86, MSE = 659.55, R^2 = 0.36, Spearman ρ = 0.66), with the parity plot revealing considerable scatter and only a modest rank-order correlation. Despite the low coefficient of determination, the model retained a moderate discriminative ability to classify high vs. low degradation at a threshold of 80% [24] (ROC-AUC = 0.74, precision = 0.84, recall = 0.61), suggesting that while precise quantification of D_{max} remains challenging, coarse categorical predictions (*e.g.*, strong degraders vs. weak degraders) are feasible.

4.2.3 Model Comparison for Activity Classification. Figure 3b shows statistical comparison of the three models on the folds validation sets using the best-performing configurations (Table 2, marked with ‡). ANOVA detected significant architecture effects for all metrics: ROC-AUC ($p = 1.17 \times 10^{-15}$), PR-AUC ($p = 1.73 \times 10^{-11}$), MCC ($p = 2.77 \times 10^{-7}$), and recall at 80% precision ($p = 2.91 \times 10^{-8}$). Tukey's HSD post-hoc tests with family-wise error rate correction identified XGBoost as the best-performing model across all four metrics (ROC-AUC: 0.851, PR-AUC: 0.870, MCC: 0.523, Recall@80%P: 0.777), significantly outperforming both MLP and PROTAC-STAN on each. MLP in turn significantly outperformed PROTAC-STAN on all four metrics, with the largest margin on recall at 80% precision ($\Delta = 0.183$, $p_{\text{adj}} = 0.001$).

4.3 Ensemble Results

Table 3 shows the performance of the ensemble methods and baselines on the isolated held-out set. On pDC_{50}, the best single model achieved the lowest test RMSE (0.633), with the Caruana ensemble (RMSE= 0.672, 33 models) ranking second and outperforming both the uniform average (0.683) and best model ensemble (0.685). On D_{max}, the best single model again achieved the lowest RMSE (20.94), with the Caruana ensemble (RMSE= 21.32, 22 models) ranking second and outperforming the best model ensemble (22.36). For binary classification, the best single model achieved the lowest log loss (0.295), and ensemble strategies performed worse: best model ensemble (0.312, 25 models), Caruana ensemble (0.317, 18 models), and uniform average (0.380, 500 models).

4.3.1 Uncertainty Quantification. Ensemble uncertainty estimates demonstrate meaningful correlations with prediction errors across regression tasks (Table 3). For pDC_{50} prediction, the uniform average of all models achieved the strongest uncertainty-error correlation (Spearman $\rho = 0.355$, $p < 0.001$), while the Caruana ensemble showed a weaker but significant correlation ($\rho = 0.207$, $p < 0.01$). The Caruana ensemble's prediction standard deviation averaged 0.361 pDC_{50} units, with 46.5% of the test samples falling within $\pm 1\sigma$ intervals and 88.0% within $\pm 3\sigma$. These coverage values fall below the theoretical expectations of 68.3%, 95.5%, and 99.7% for normally distributed errors, suggesting systematic underestimation

Table 3: Ensemble performance and uncertainty quantification on held-out test set. Performance evaluated on 80% evaluation subset (20% used for Caruana selection). Uncertainty-error correlation: Spearman ρ between std and absolute error. Coverage: fraction within kσ (expected: 68.3%, 95.5%, 99.7%). Except for Best Single, all other methods are ensembles: Uniform averages all 500 models; Best Ensem. averages the 25 best models; Caruana is described in Sec. 3.5.1

Performance				Uncertainty Quantification					
Task	Method	Metric	N	$\bar{\sigma}$	ρ	1σ	2σ	3σ	ECE/MCE
pDC_{50} (RMSE)	Best Single	**0.633**	1	–	–	–	–	–	–
	Uniform	0.683	500	0.365	**0.355****	42.4	**78.3**	**91.2**	–
	Best Ensem.	0.685	25	0.284	0.303**	33.6	64.5	82.0	–
	Caruana	0.672	33	0.361	0.207*	**46.5**	70.5	88.0	–
D_{max} (RMSE)	Best Single	**20.94**	1	–	–	–	–	–	–
	Uniform	22.92	500	12.53	**0.694****	37.3	**80.5**	**91.7**	–
	Best Ensem.	22.36	25	7.59	0.614**	21.3	49.7	74.0	–
	Caruana	21.32	22	12.47	0.543**	**47.9**	78.1	91.1	–
Activity (LogLoss)	Best Single	**0.295**	1	–	–	–	–	–	–
	Uniform	0.380	500	–	–	–	–	–	11.4 / **28.2**
	Best Ensem.	0.312	25	–	–	–	–	–	**10.0** / 33.6
	Caruana	0.317	18	–	–	–	–	–	12.1 / 35.6

$^{**}p < 0.001$; $^{*}p < 0.01$. Coverage, ECE, and MCE are expressed in %.

of uncertainty by approximately 20-30%. For D_{max}, the uniform average exhibits the highest correlation ($\rho = 0.694$) alongside the best coverage at 2σ (80.5%), while the Caruana ensemble shows more balanced performance ($\rho = 0.543$, 47.9% coverage at 1σ). The best model ensemble achieves the lowest uncertainty spread ($\bar{\sigma} = 7.59$) but severely underestimates coverage (21.3% at 1σ), highlighting a tradeoff between uncertainty magnitude and correlation strength.

For binary classification, the ECE ranges from 10.0% (Best Ensem.) to 12.1% (Caruana), indicating moderate calibration with average deviations around 10%. Unlike regression tasks, MCE values are substantially lower (28.2–35.6%), suggesting that high-confidence predictions are more reliable for binary classification than for regression.

5 Discussion

Our systematic evaluation of ML approaches for PROTAC degradation reveals several insights relevant to both the AI/ML and drug discovery communities. We discuss here key findings, their implications, and limitations that inform future research directions.

5.1 Feature Representation & Simplicity

A striking finding from our feature selection analysis (Table 2) is that simple encodings (OneHot, Vec) often matched or approached the performance of computationally expensive ESM-S embeddings, particularly when paired with XGBoost. This suggests that for PROTAC activity prediction, lightweight representations may suffice when paired with tree-based models. Further, the task-specific differences in optimal feature sets warrant additional discussion. For both pDC_{50} and D_{max} predictions, the best configurations included descriptor-only, whereas Desc+FP representations performed comparably. This divergence may reflect the distinct physicochemical determinants of these endpoints, expanded on in the next subsection. The success of simple features comes with an important caveat regarding generalization. One-hot encodings, by construction, cannot represent proteins unseen during training: an unknown POI has no meaningful bit mapping. This limits applicability to novel

targets, a critical consideration given that expanding the "druggable" proteome is a key goal of PROTAC research.

5.2 Asymmetry in Endpoint Predictability

The substantial gap in predictive performance between pDC_{50} (R^2 = 0.66) and D_{max} (R^2 = 0.36) reveals a fundamental asymmetry in PROTAC predictability. Potency (pDC_{50}) is considerably easier to learn from molecular and contextual features than maximal degradation efficacy (D_{max}). This asymmetry likely reflects the underlying biology. DC_{50} is primarily governed by ternary complex formation kinetics and PROTAC-target binding affinity—properties with clearer structure-activity relationships amenable to ML modeling. In contrast, D_{max} depends on a cascade of factors: E3 ligase expression levels, target protein localization, proteasome availability, and the presence of competing binding partners that may occlude the target from PROTAC engagement [9]. These factors vary across cell lines and experimental conditions in ways that the explored molecular representations do not capture. From a practical standpoint, this finding suggests that binary classification (active/inactive) and pDC_{50} regression are more reliable for computational screening, while D_{max} predictions should be interpreted with greater caution.

5.3 Tree-Based Models Outperform DL

XGBoost outperformed both MLP and PROTAC-STAN across nearly all tasks, falling short of MLP by 1.4% on predicting DC_{50}, a result that merits examination given the recent emphasis on DL for molecular property prediction. We hypothesize several contributing factors. With <4K unique labeled PROTACs in TACK, the dataset may be insufficient for deep architectures to learn robust representations. Tree-based methods are well-established as strong performers on small-to-medium tabular datasets [20], where their design provides effective regularization. The prevalence of activity thresholds in the literature (*e.g.*, 100 nM for DC_{50}, 80% for D_{max}) creates natural decision boundaries that tree-based models can exploit through their hierarchical structure. Neural networks must learn these boundaries implicitly, requiring more data.

PROTAC-STAN, despite its attention-based architecture designed specifically for PROTAC activity prediction, underperformed relative to simpler tree-based models on TACK. This performance gap likely reflects a mismatch in model complexity and the available data size, or suboptimal hyperparameter transfer from the original setting. Notably, PROTAC-STAN was fine-tuned on the authors' curated PROTAC-DB subset, where it achieved 88% accuracy on held-out data [11]. The worse performance we observe on TACK—which incorporates structurally diverse compounds from TPDdb beyond the original PROTAC-DB distribution—suggests potential overfitting to the fine-tuning domain. We did not evaluate DeepPROTACs [24] because it requires pre-computed binding pocket structures for both the POI and E3 ligase, information which is unavailable for the majority of TACK entries. This exclusion highlights a broader challenge in the field: methods incorporating 3D inputs may offer richer representations, but face significant data availability constraints in practice. As structure prediction tools continue to mature [14], revisiting structure-aware approaches on new benchmarks represents a promising direction of future work.

5.4 Uncertainty Quantification and Data Quality

Our ensemble-based uncertainty estimates showed positive correlation with prediction error (Sec. 4.3.1), indicating reasonable calibration. However, the modest magnitude of these correlations suggests that epistemic uncertainty (reducible through additional data or improved models) may not fully explain prediction errors, especially in the case of pDC_{50} prediction. This observation points to substantial aleatoric uncertainty inherent in the data itself. Sources can include: assay variability across laboratories and protocols; cell line-specific effects not captured by current encodings; and measurement noise in reported endpoint values. If aleatoric uncertainty dominates, even perfect models cannot achieve high predictive accuracy without improved data quality. The broader dataset context supports this interpretation: of the nearly 30K experimentally characterized PROTACs across public databases, only ~9K (~30%) have associated degradation measurements. The remaining compounds lack quantitative activity data, representing both a limitation and an opportunity. Active learning methods that iteratively prioritize informative compounds for experimental characterization could efficiently expand the labeled dataset while maximizing model improvement per experiment [35].

6 Limitations and Ethical Considerations

Several limitations should be considered when interpreting our results. *Generalization to new targets.* Our models may perform markedly worse on structurally novel POIs or E3 ligases not represented in training. We explored target-based splits but found they produced highly imbalanced sets, complicating fair evaluation. Developing robust strategies for out-of-distribution generalization remains an open challenge. *Lack of 3D information.* We demonstrated that current best models rely on 1D/2D molecular representations. Incorporating 3D ternary complex structure information, now increasingly available through co-folding tools, could improve predictions by capturing binding pocket complementarity and protein-protein interfaces. This however requires reliable structure prediction for arbitrary POI-PROTAC-E3 combinations, which remains an active research area. *Cell context.* While we encode cell line identity, we do not model the underlying biological differences (*e.g.*, E3 ligase expression levels) that drive cell-specific degradation outcomes. Integrating transcriptomic or proteomic context could improve predictions but requires matched omics data currently unavailable at scale. *Benchmark scope.* TACK focuses on CRBN- and VHL-recruiting PROTACs, which dominate current datasets. Performance on emerging E3 ligases (*e.g.*, IAPs) remains untested and may differ substantially. *Dual-use concerns.* TACK may in theory be misused to predict activity against beneficial targets, with the aim of designing toxic PROTACs. However, the lack of public data on PROTAC toxicity makes this a weak concern at this stage.

7 Conclusion

We introduce TACK, a standardized benchmark dataset aggregating 3,514 unique PROTACs with 6,561 degradation endpoints, and show a rigorous statistical evaluation of ML approaches for PROTAC activity prediction. Our findings yield three actionable insights for the field. First, classical tree-based methods outperform DL models on current datasets, suggesting that architectural sophistication

should match data availability. Second, potency (DC_{50}) is substantially more predictable than efficacy (D_{max}), reflecting the greater dependence of maximum degradation on cellular factors not captured by current standard representations. Third, ensemble-based uncertainty quantification provides calibrated confidence estimates that can guide future prioritization in active learning workflows.

These results establish new baseline expectations for PROTAC property prediction and highlight the continued importance of data curation, standardized evaluation protocols, and thoughtful feature engineering. As PROTAC databases expand and structure prediction tools mature, we anticipate that the gap between simple and complex models will narrow—but principled benchmarking will remain essential to distinguish genuine methodological advances from overfitting to narrow chemical series. TACK, our evaluation framework, and all best models reported herein are made publicly available to support reproducible progress in this emerging area of ML-guided drug discovery.

GenAI Disclosure

The authors used generative AI for polishing parts of the text and the code, using the following models: Claude Sonnet 4.5, Claude Opus 4.5 and Gemini Pro. All the authors have reviewed and approve of the final version of the submitted manuscript.

Software and Data

Code to reproduce all work, including the data curation pipeline, model training, and statistical analysis, can be found in the following repository: **https://github.com/ribesstefano/TACK/**

The curated TACK dataset can be inspected and obtained from HuggingFace: **https://huggingface.co/datasets/ailab-bio/TACK**

All best models presented herein, including ensemble models, are available via Zenodo: **https://zenodo.org/records/15691822**

Acknowledgments

SR and RM acknowledge funding provided by the Chalmers Gender Initiative for Excellence (Genie). RM and ND acknowledge funding provided by the Wallenberg AI, Autonomous Systems, and Software Program (WASP), supported by the Knut and Alice Wallenberg Foundation. The authors thank Yossra Gharbi, Alexander Persson, and Felix Erngård for helpful discussions. The computations and data storage were enabled by resources provided by Chalmers e-Commons and by the National Academic Infrastructure for Supercomputing in Sweden (NAISS), partially funded by the Swedish Research Council through grant agreement no. 2022-06725. The authors declare no competing interests.

References

[1] Takuya Akiba, Shotaro Sano, Toshihiko Yanase, Takeru Ohta, and Masanori Koyama. 2019. Optuna: A next-generation hyperparameter optimization framework. In *Proceedings of the 25th ACM SIGKDD international conference on knowledge discovery & data mining*. 2623–2631.

[2] Jeremy R. Ash, Cas Wognum, Raquel Rodríguez-Pérez, Matteo Aldeghi, Alan C. Cheng, Djork-Arné Clevert, Ola Engkvist, Cheng Fang, Daniel J. Price, Jacqueline M. Hughes-Oliver, and W. Patrick Walters. 2025. Practically Significant Method Comparison Protocols for Machine Learning in Small Molecule Drug Discovery. *Journal of Chemical Information and Modeling* 65, 18 (Sept. 2025), 9398–9411. Publisher: American Chemical Society. doi:10.1021/acs.jcim.5c01609

[3] Amos Bairoch. 2018. The Cellosaurus, a cell-line knowledge resource. *Journal of Biomolecular Techniques* 29, 2 (2018), 25–38. doi:10.7171/jbt.18-2902-002

[4] Miklós Békés, David R Langley, and Craig M Crews. 2022. PROTAC targeted protein degraders: the past is prologue. *Nature Reviews Drug Discovery* 21, 3 (2022), 181–200.

[5] Guy W. Bemis and Mark A. Murcko. 1996. The Properties of Known Drugs. 1. Molecular Frameworks. *Journal of Medicinal Chemistry* 39, 15 (Jan. 1996), 2887–2893. Publisher: American Chemical Society. doi:10.1021/jm9602928

[6] Yoav Benjamini and Yosef Hochberg. 1995. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society: Series B (Methodological)* 57, 1 (1995), 289–300.

[7] James Bergstra, Rémi Bardenet, Yoshua Bengio, and Balázs Kégl. 2011. Algorithms for hyper-parameter optimization. *Advances in Neural Information Processing Systems* 24 (2011).

[8] Hong Cai, Gengyuan Yao, Yulong Shi, Tianyi Zhang, and Yuanjia Hu. 2025. PROTAC-PatentDB: A PROTAC patent compound dataset. *Scientific Data* 12, 1 (2025), 1840.

[9] Annabel Cardno, Bryony Kennedy, and Catherine Lindon. 2025. Cellular parameters shaping pathways of targeted protein degradation. *Communications Biology* 8, 1 (2025), 691.

[10] Rich Caruana, Alexandru Niculescu-Mizil, Geoff Crew, and Alex Ksikes. 2004. Ensemble selection from libraries of models. In *Proceedings of the Twenty-First International Conference on Machine Learning* (Banff, Alberta, Canada) *(ICML '04)*. Association for Computing Machinery, New York, NY, USA, 18. doi:10.1145/1015330.1015432

[11] Zhenglu Chen, Chunbin Gu, Shuoyan Tan, Xiaorui Wang, Yuquan Li, Mutian He, Ruiqiang Lu, Shijia Sun, Chang-Yu Hsieh, Xiaojun Yao, Huanxiang Liu, and Pheng-Ann Heng. 2025. Interpretable PROTAC degradation prediction with structure-informed deep ternary attention framework. *Advanced Science* 12, 47 (2025), e08138. doi:10.1002/advs.202508138

[12] Deborah Chirnomas, Keith R Hornberger, and Craig M Crews. 2023. Protein degraders enter the clinic—a new approach to cancer therapy. *Nature Reviews Clinical Oncology* 20, 4 (2023), 265–278.

[13] Michael L Drummond and Christopher I Williams. 2019. In silico modeling of PROTAC-mediated ternary complexes. *Journal of Chemical Information and Modeling* 59, 4 (2019), 1634–1644.

[14] Nils Dunlop, Francisco Erazo, Farzaneh Jalalypour, and Rocío Mercado. 2025. Predicting PROTAC-mediated ternary complexes with AlphaFold3 and Boltz-1. *Digital Discovery* 4, 12 (2025), 3782–3809.

[15] Milton Friedman. 1937. The use of ranks to avoid the assumption of normality implicit in the analysis of variance. *J. Amer. Statist. Assoc.* 32, 200 (1937), 675–701. doi:10.1080/01621459.1937.10503522

[16] Jingxuan Ge, Shimeng Li, Gaoqi Weng, Huating Wang, Meijing Fang, Huiyong Sun, Yafeng Deng, Chang-Yu Hsieh, Dan Li, and Tingjun Hou. 2025. PROTAC-DB 3.0: an updated database of PROTACs with extended pharmacokinetic parameters. *Nucleic Acids Research* 53, D1 (2025), D1510–D1515.

[17] Yossra Gharbi and Rocío Mercado. 2024. A comprehensive review of emerging approaches in machine learning for de novo PROTAC design. *Digital Discovery* 3, 11 (2024), 2158–2176.

[18] Warren Gilchrist. 2000. *Statistical modelling with quantile functions*. Chapman and Hall/CRC.

[19] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2015. Delving Deep into Rectifiers: Surpassing Human-Level Performance on ImageNet Classification. In *Proceedings of the 2015 IEEE International Conference on Computer Vision (ICCV) (ICCV '15)*. IEEE Computer Society, USA, 1026–1034. doi:10.1109/ICCV.2015.123

[20] Dejun Jiang, Zhenxing Wu, Chang-Yu Hsieh, Guangyong Chen, Ben Liao, Zhe Wang, Chao Shen, Dongsheng Cao, Jian Wu, and Tingjun Hou. 2021. Could graph neural networks learn better molecular representation for drug discovery? A comparison study of descriptor-based and graph-based models. *Journal of Cheminformatics* 13, 1 (2021), 12.

[21] Jin-Hwa Kim, Jaehyun Jun, and Byoung-Tak Zhang. 2018. Bilinear Attention Networks. In *Advances in Neural Information Processing Systems 31*. 1571–1581.

[22] Gregory Landrum. 2024. RDKit: Open-source Cheminformatics. Accessed: 2026-02-01. https://www.rdkit.org

[23] Howard Levene. 1960. Robust tests for equality of variances. *Contributions to Probability and Statistics* (1960), 278–292.

[24] Fenglei Li, Qiaoyu Hu, Xianglei Zhang, Renhong Sun, Zhuanghua Liu, Sanan Wu, Siyuan Tian, Xinyue Ma, Zhizhuo Dai, Xiaobao Yang, et al. 2022. DeepPROTACs is a deep learning-based targeted degradation predictor for PROTACs. *Nature Communications* 13, 1 (2022), 7133.

[25] Nir London and Jaime Prilusky. 2024. PROTACpedia. https://protacpedia.weizmann.ac.il. Accessed: 2026-02-01.

[26] Divya Nori, Connor W. Coley, and Rocío Mercado. 2022. De novo PROTAC design using graph-based deep generative models. In *Proceedings of the NeurIPS 2022 Workshop on AI for Science*. doi:10.48550/arXiv.2211.02660

[27] Maja Pavlovic. 2025. Understanding Model Calibration: A gentle introduction and visual exploration of calibration and the expected calibration error (ECE). arXiv:2501.19047 [stat.ME] https://arxiv.org/abs/2501.19047

[28] Xinran Qin, Yinpeng Zhang, Yajunzi Wang, Yintao Zhang, Jiachen Jing, Yuyuan Zhang, Gaoxiang Xu, Haoping Teng, Tianjun Wang, Lei Fu, et al. 2026. TPDdb:

the comprehensive database of targeted protein degrader. *Nucleic Acids Research* 54, D1 (2026), D1683–D1691.
[29] Stefano Ribes, Eva Nittinger, Christian Tyrchan, and Rocío Mercado. 2024. Modeling PROTAC degradation activity with machine learning. *Artificial Intelligence in the Life Sciences* 6 (2024), 100104. doi:10.1016/j.ailsci.2024.100104
[30] Kathleen M Sakamoto, Kyung B Kim, Akiko Kumagai, Frank Mercurio, Craig M Crews, and Raymond J Deshaies. 2001. Protacs: Chimeric molecules that target proteins to the Skp1–Cullin–F box complex for ubiquitination and degradation. *Proceedings of the National Academy of Sciences* 98, 15 (2001), 8554–8559.
[31] Ashley R Schneekloth, Mathieu Pucheault, Hyun Seop Tae, and Craig M Crews. 2008. Targeted intracellular protein degradation induced by a small molecule: En route to chemical proteomics. *Bioorganic & Medicinal Chemistry Letters* 18, 22 (2008), 5904–5908.
[32] Melanie Schneider, Jean-Luc Pons, William Bourguet, and Gilles Labesse. 2020. Towards accurate high-throughput ligand affinity prediction by exploiting structural ensembles, docking metrics and ligand similarity. *Bioinformatics* 36, 1 (01 2020), 160–168. doi:10.1093/bioinformatics/btz538
[33] The UniProt Consortium. 2025. UniProt: the Universal Protein Knowledgebase in 2025. *Nucleic Acids Research* 53, D1 (2025), D609–D617. doi:10.1093/nar/gkae1010
[34] John W Tukey. 1949. Comparing individual means in the analysis of variance. *Biometrics* (1949), 99–114.
[35] Derek van Tilborg and Francesca Grisoni. 2024. Traversing chemical space with active deep learning for low-data drug discovery. *Nature Computational Science* 4, 10 (2024), 786–796.
[36] Gaoqi Weng, Xuanyan Cai, Dongsheng Cao, Hongyan Du, Chao Shen, Yafeng Deng, Qiaojun He, Bo Yang, Dan Li, and Tingjun Hou. 2023. PROTAC-DB 2.0: an updated database of PROTACs. *Nucleic Acids Research* 51, D1 (2023), D1367–D1372.
[37] R. F. Woolson. 2005. *Wilcoxon Signed-Rank Test.* John Wiley & Sons, Ltd. arXiv:https://onlinelibrary.wiley.com/doi/pdf/10.1002/0470011815.b2a15177 doi:10.1002/0470011815.b2a15177
[38] Daniel Zaidman, Jaime Prilusky, and Nir London. 2020. PRosettaC: Rosetta based modeling of PROTAC mediated ternary complexes. *Journal of Chemical Information and Modeling* 60, 9 (2020), 4467–4480.
[39] Zuobai Zhang, Jiarui Lu, Vijil Chenthamarakshan, Aurélie Lozano, Payel Das, and Jian Tang. 2024. Structure-Informed Protein Language Model. arXiv:2402.05856 doi:10.48550/arXiv.2402.05856

A CV Folds Label Distribution

Figure 5 shows the distribution of the target labels of the 5×5 CV folds used in each of the three tasks explored in this work, together with the mean of the hold-out set labels. This plot demonstrates that the CV folds are balanced.

B Normality Diagnostic & Effect Size

Figure 4 visually confirms the assumption of normality distribution of the metrics values of the activity prediction models. This allowed us to perform parametric Tukey's HSD test to compare the performance of the three activity prediction models, XGBoost$^{(\ddagger)}$, MLP$^{(\ddagger)}$, and PROTAC-STAN, as described in Section 3.4.2. Detailed effect sizes from performing Tukey's HSD are visualized in Figure 6. The top-left-most model indicates the best performing model for a given metric.

C Model Details

We developed a multilayer perceptron (MLP) architecture with flexible depth and width. The MLP consists of a variable number of hidden layers with dimensions sampled from predefined configurations (*e.g.*, [256], [512, 256], [1024, 512, 256]), followed by a regression head (*i.e.*, a linear layer) with configurable depth (1–3 layers). Each hidden layer incorporates optional normalization (batch normalization or layer normalization), dropout regularization (0.0–0.5), and activation functions (ReLU, GELU, or SiLU). The model weights are initialized using Kaiming initialization [19], with the first layer initialized from $\mathcal{N}(0, 1/\sqrt{d_{\text{in}}})$ and subsequent layers from $\mathcal{N}(0, \sqrt{2}/\sqrt{d_{\text{in}}})$ to account for ReLU nonlinearities. Training leveraged gradient clipping (0.5–2.0), mixed-precision training (16-bit), and early stopping with a patience of 5 epochs based on the validation loss.

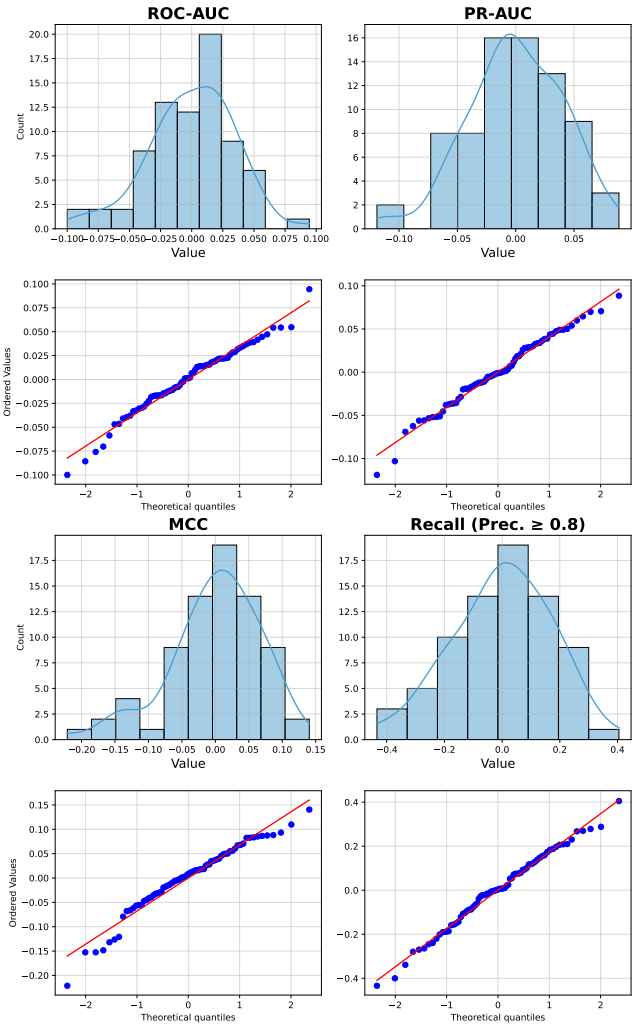


Figure 4: Normality diagnostic of the performance metric values on the fold validation sets of the three models: XGBoost$^{(\ddagger)}$, MLP$^{(\ddagger)}$, and PROTAC-STAN.

PROTAC-STAN [11] is a graph neural network architecture designed to predict the degradation activity of PROTACs using a ternary attention network (TAN) layer [21]. The model encodes PROTAC molecules using a two-layer edge-aware graph convolutional network (GCN) with global max pooling, while protein sequences (POI and E3 ligase) are represented using ESM-S [39] embeddings processed through a two-layer fully connected adapter. The TAN layer is a specialized multi-head attention mechanism that explicitly models the complex three-way interactions between the POI, PROTAC, and E3 ligase embeddings computing a joint representation via tensor outer products. The joint representation is then fed into a classifier MLP with batch normalization and ReLU activation to compute the probability of degradation activity. For

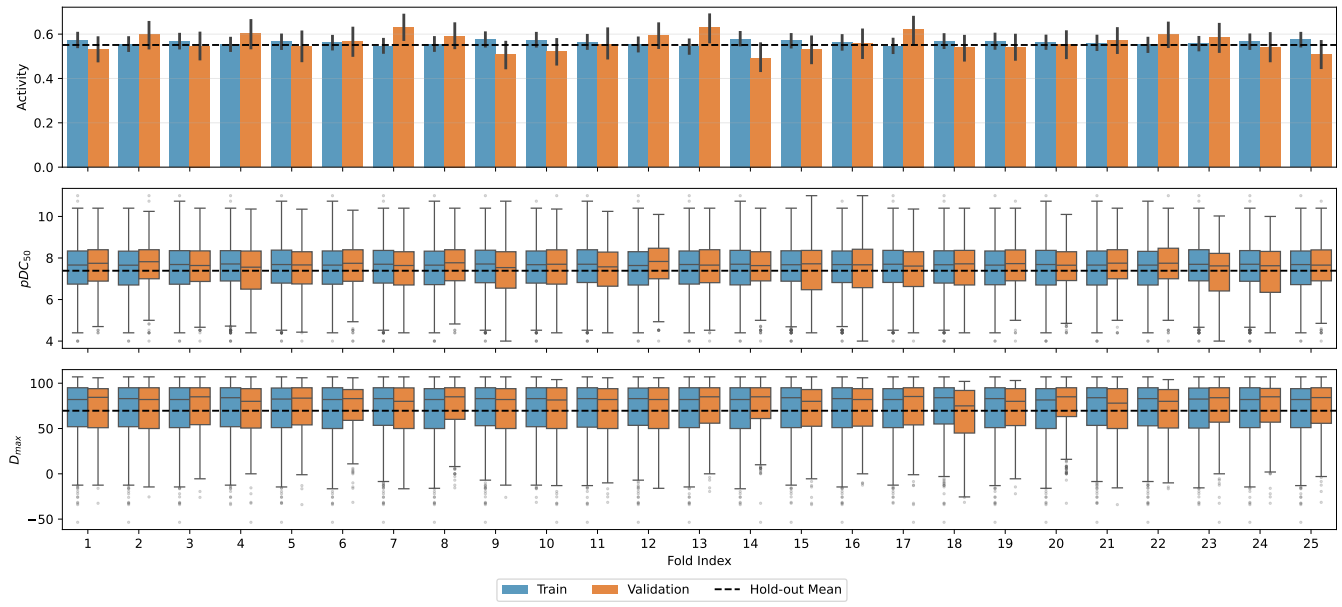


Figure 5: Distribution of the target values across the 5×5 CV folds for each prediction task.

PROTAC-STAN, we trained the model from randomly initialized weights using the hyperparameters reported in the original publication [11]. Early stopping with patience of 30 epochs was applied for consistency with the other models.

D Feature Sets

Table 4 lists the combinations of the different encodings of the features we explored in our work. We limited our analysis to the 10 most promising feature sets. For example, we avoided using 1- and 2-grams vectorization of the dozen available E3 ligase amino acid sequences, as they resulted in vectors of dimension ~4K, making the models prone to overfitting. The column 'Assay' reports whether the assay type (one-hot encoded) and the experiment time (real value) are included in the set.

Table 4: The feature set configurations evaluated across tasks. Shorthand for feature sets is defined in Sec. 3.2.2.

Config	Cell	E3	POI	Mol	Assay
1	OneHot	OneHot	OneHot	Desc+FP	✓
2	OneHot	OneHot	OneHot	FP	✓
3	OneHot	OneHot	OneHot	Desc	✓
4	OneHot	OneHot	Vec	Desc	✓
5	Text	ESM-S-PCA	ESM-S-PCA	Desc+FP	✓
6	Text	ESM-S-PCA	ESM-S-PCA	Desc	✓
7	Text	OneHot	Vec	Desc+FP	✓
8	Text	OneHot	OneHot	Desc	✓
9	Text	OneHot	Vec	Desc	✓
10	–	ESM-S	ESM-S	Desc+FP	–

E Hyperparameters

In training the XGBoost and MLP models, we performed hyperparameter optimization via Optuna [1], using 20 and 100 trials for the XGBoost and MLP models, respectively (search space summarized in Table 5). The optimization objective was to minimize the root mean squared error (RMSE) for regression tasks and maximize accuracy for classification. The hyperparameters leading to the best performance in the validation set of the first fold were selected and used to train the 25 models on all folds. Finally, at each fold, we set a different random seed for the model weights initialization to introduce variance in the models' parameters.

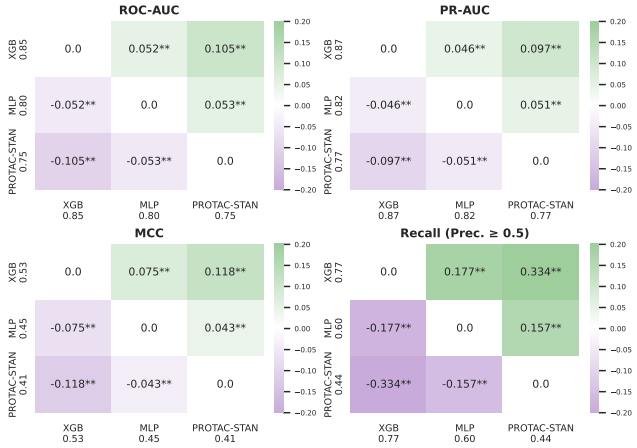


Figure 6: Effect sizes of the performance metric values on the fold validation sets of the three models: XGBoost$^{(\ddagger)}$, MLP$^{(\ddagger)}$, and PROTAC-STAN. Values marked with ** have a significance of: $0.001 \leq p < 0.1$.

Table 6: Optimized hyperparameters for the best performing models marked with (†) and (‡) in Table 2. XGBoost models used 2000 estimators with early stopping after 30 rounds. MLP models were trained for up to 200 epochs with early stopping and mixed precision (16-bit).

XGBoost Models		
Hyperparameter	D_{max} (†)	**Activity**(‡)
Learning rate	0.0159	0.0018
Max depth	7	6
Min child weight	1	1
Gamma	0.005	0.120
Subsample	0.982	0.946
Column sample by tree	0.916	0.629
L1 regularization (α)	0.007	0.447
L2 regularization (λ)	0.005	0.018
MLP Models		
Hyperparameter	pDC_{50} (†)	**Activity**(‡)
Hidden dimensions	[256]	[512, 256, 128]
Dropout	0.0	0.0
Head depth	1	1
Head dropout	0.1	0.1
Normalization	Layer	Layer
Learning rate	0.00075	0.00036
LR scheduler	Reduce-on-plateau	Reduce-on-plateau
Warmup ratio	0.0	0.05
Activation	ReLU	SiLU
Gradient clip value	1.0	1.5

E.1 Hyperparameter Space

Hyperparameters for both XGBoost and MLP models were optimized using the tree-structured Parzen estimator (TPE) algorithm [7] implemented in Optuna [1], with 20% of trials allocated for random exploration and multivariate parameter modeling enabled. Table 5 summarizes the search space explored during hyperparameter optimization for each model. For XGBoost, we tuned 8 parameters controlling tree complexity and regularization. For MLP models, we optimized 9 architectural and training parameters, including network depth/width, regularization, and learning rate schedules.

Table 5: Hyperparameter search spaces for the XGBoost and MLP models.

Model	Hyperparameter	Search Space
XGB	Learning rate	$[10^{-3}, 10^{-1}]$ (log-uniform)
	Max depth	$[3, 9]$ (integer)
	Min child weight	$[1, 25]$ (integer)
	Subsample	$[0.4, 1.0]$ (uniform)
	Column sample by tree	$[0.5, 1.0]$ (uniform)
	L1 regularization (α)	$[10^{-3}, 10]$ (log-uniform)
	L2 regularization (λ)	$[10^{-3}, 10]$ (log-uniform)
	Gamma	$[10^{-3}, 10]$ (log-uniform)
MLP	Hidden dimensions	Categorical: [256], [512], [256, 128], [512, 256], [512, 256, 128], [1024, 512], [1024, 512, 256]
	Dropout	$[0.0, 0.5]$ (step 0.1)
	Head dropout	$[0.0, 0.3]$ (step 0.1)
	Head depth	$[1, 3]$ (integer)
	Normalization type	Categorical: None, batch, layer
	Activation function	Categorical: ReLU, GELU, SiLU
	Learning rate	$[10^{-5}, 10^{-2}]$ (log-uniform)
	LR scheduler	Categorical: cosine, reduce-on-plateau
	Warmup ratio	$[0.0, 0.2]$ (step 0.05)
	Gradient clip value	$[0.5, 2.0]$ (step 0.5)

E.2 Best Models Hyperparameters

Table 6 lists the hyperparameters of the best models reported in Table 2, specifically those marked with † and ‡.