Machine Learning Frameworks for Predictive PK/PD Modeling of Degraders from Fragmented Multi-Source Data

Mridula Bontha, Heta A. Gandhi, Elena Cáceres, John Eichenseer, Lik Hang Yuen, Prabhu Raman, Ryan Pemberton, Emel Ficici, Jeff Mihalic, Ryan Rountree ,Jeffery Wu, John Lee, Christopher Phelps, John Lin, Mitchell Lavaris, Alexandra Trotier, Robert Cass, Karthik Arumugam, Sasha Borodovsky, Ganesh Cherala, Daniel Chan, Wylie Palmer, Ya Wen Lu, Ge Peng, Gwenn Hansen, Paul Novick.
Nurix Therapeutics, Inc., San Francisco, CA, USA.

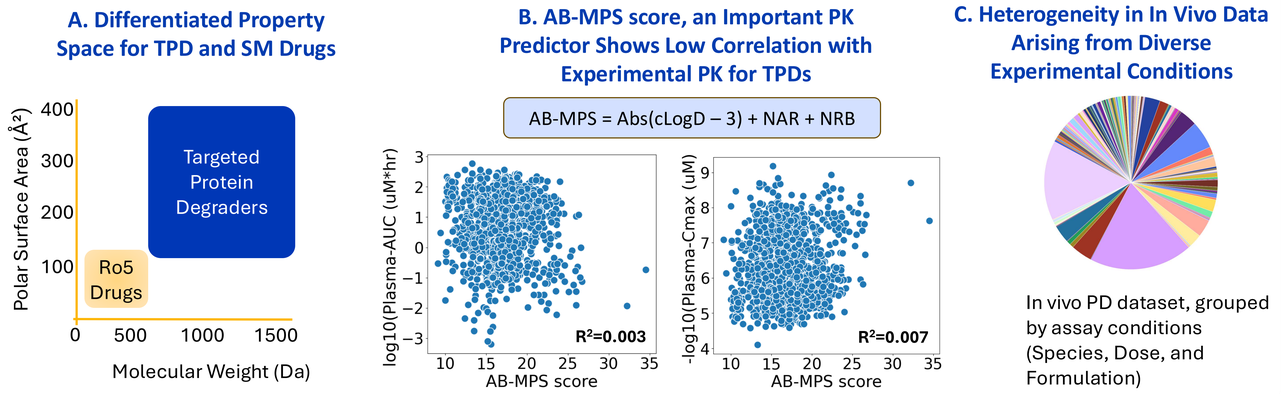
Background

Targeted protein degradation (TPD) is a rapidly advancing area of drug discovery that harnesses the cell's machinery to eliminate disease-causing proteins. Bivalent TPD molecules occupy a differential chemical space compared to traditional small-molecule drugs and display differentiated PK property profiles. As a result, rules governing optimization developed for small-molecule drugs do not translate to degrader development.

Machine learning (ML) offers an opportunity to learn directly from primary data to generate predictive models to guide TPD development. However, the diversity across **high-fidelity PK/PD** endpoints present an opportunity for developing robust, **data-efficient ML** techniques that capitalize on the complementary information encoded in each assay.

We report here a **feature engineering** and **model-stacking** approach which allows us **to aggregate data** from disparate in vivo datasets and further leverage our **DEL-AI platform** to train **ML models** with high accuracy **in PK and PD prediction.**

Figure 1. Current Challenges in the Development of TPD Therapeutics



Abbreviations

SM : Small Molecules.
TPD : Targeted Protein Degraders.
PK : Pharmacokinetics; what the body does to the drug.
PD : Pharmacodynamics; what the drug does to the body.
ML : Machine Learning.
AB-MPS : AbbVie Multiparameter Score; estimates the likelihood of successful preclinical pharmacokinetic (PK) results.
NRB : Number of Rotatable Bonds.
NAR : Number of Aromatic Rings.
TPSA : Total Polar Surface Area.
HBD : Number of Hydrogen Bond Donors.
MW : Molecular Weight (in Da).
AUC : Area Under Plasma Concentration Curve.
C_{max} : Peak concentration.
T_{max} : Time to C_{max}.
%F : Oral bioavailability ; percentage of drug that reaches. systemic circulation.
CNS : Central Nervous System.

Feature Engineering for Dataset Aggregation

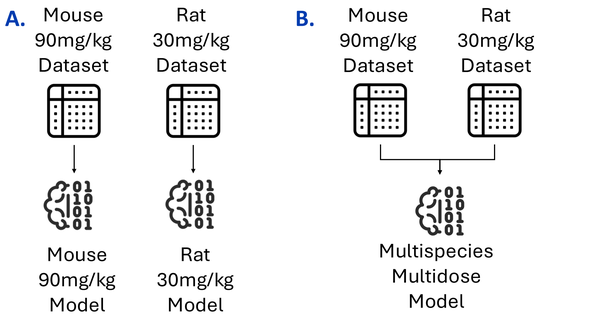


Figure 2. Dataset Aggregation through Encoding Experimental Parameters. (A) Training models on fractured datasets leads to underpowered models with limited scope. (B) Aggregating related datasets by encoding experimental parameters allows for aggregating larger training sets for more broadly applicable models. (C) Example encoding for multi-species and multi-dose models.

C.

S_M	S_R	Dose
1	0	90

S_M	S_R	Dose
0	1	30

Modeling Drug Plasma Exposure Curves

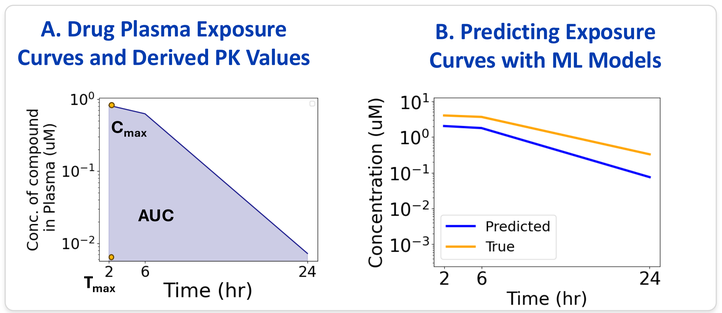


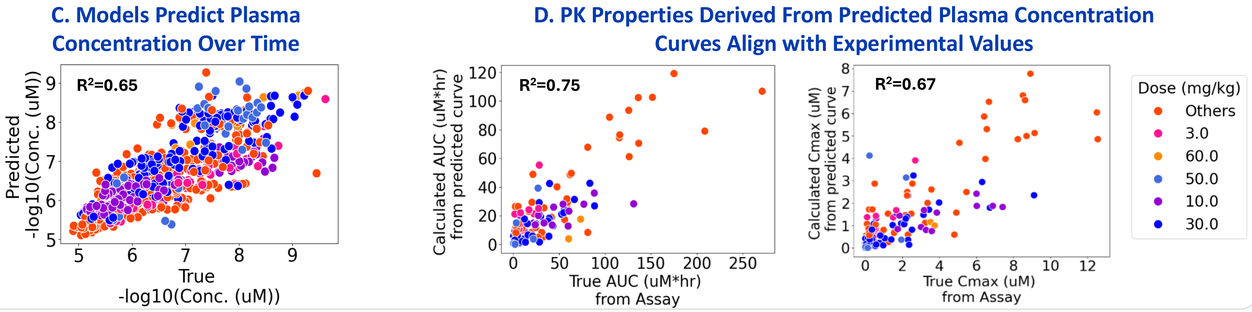
Figure 3. Modeling Plasma Exposure Curves Enables Prediction of PK endpoints.

(A) Example drug plasma exposure-curve with labelling of PK endpoints.

(B) Predicted vs Experimental exposure-curve using our ML model.

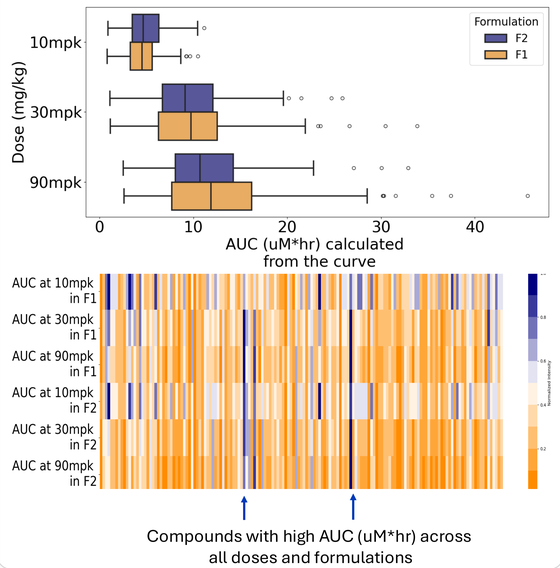
(C) Predicted vs Experimental prediction of individual (Time, Conc) points from exposure-curves.

(D) Comparison of AUC and C_{max} derived from Experimental vs Predicted exposure-curves.



AUC Predictions Prioritize Compounds and Formulations Prior To In Vivo Testing

Figure 4. (A) Boxplot Explaining Trends in Doses & Formulations (B) Heatmap Explaining Trends Across Compounds

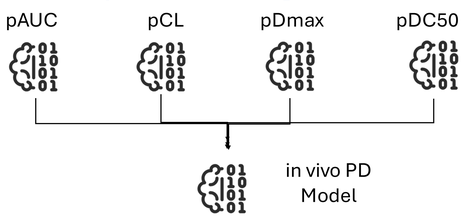


Stacked Ensemble Models Predicting PD and CNS Exposure

Figure 5. Overview of Feature Sizes and Representation Methods

Feature Type	Example Feature	Feature Size
Structural Fingerprints	C(=C)(C)N	~1000+ bits
Molecular Properties	MW, HBD	~50
Calculable Descriptors	BCUT2D, CalcNPR1	~100's
Predicted PK	**pAUC, pPPB**	**~10**

Example Model Stacking for in vivo PD



Predicted Features Enable Modeling of In Vivo PD

Figure 6. (A) Model Performance on Test Set

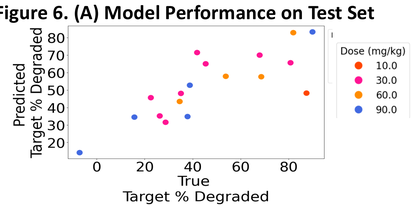
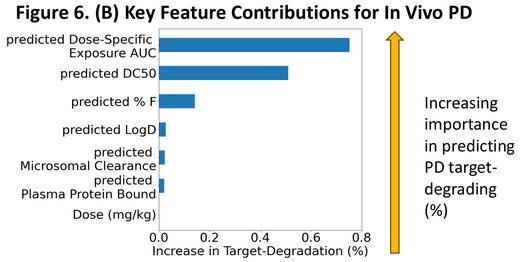


Figure 6. (B) Key Feature Contributions for In Vivo PD



Predicted Features Enable Modeling of CNS Exposure

Figure 7. (A) Model Performance on Test Set

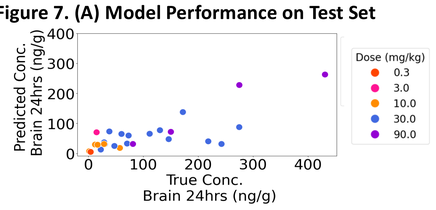
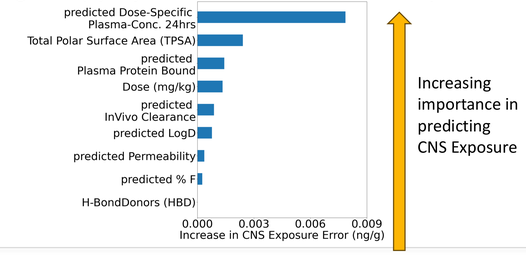


Figure 7. (B) Key Feature Contributions for CNS Exposure



Conclusions

- The inherent complexity **of in vivo PD data** mandates innovative approaches to the successful application of machine-learning methods.
- Using a combination of **feature engineering** and **model stacking**, we have developed a suite of performant machine-learning models able to predict a broad spectrum of high-value in vivo and in vitro PK and PD endpoints to a high level of accuracy.
- In data regimes characterized by **heterogeneity** due to **diversity of compounds** and **experimental conditions, predicted PK properties** serve as reliable features for developing robust machine learning models.

Presented at The Berkeley Biotech Research Day on July 22nd, 2025