Enzyme kinetic parameters prediction

Interns: Bowei, Pin-Chi

Mentor: Cheng Wang

July 29, 2025

Outline

- Background & Aims
- Methods
- Results
- Future Work
- Reflection

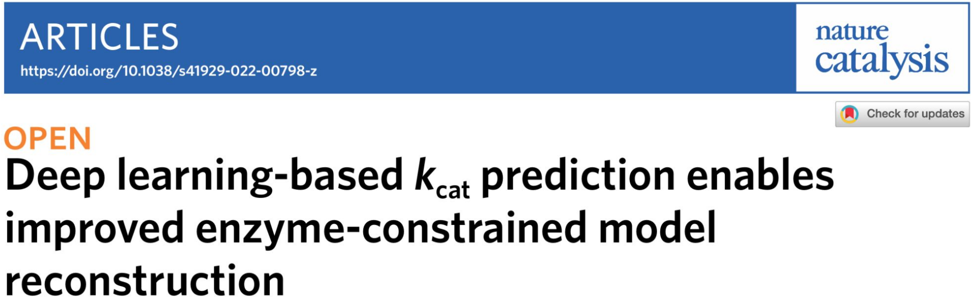
Background & Aims

Bo-Wei

Background

- Why It Matters
  - Lab tests are slow and costly, hard to measure
  - Prediction can guide wet-lab work

DLKcat →



ARTICLES

https://doi.org/10.1038/s41929-022-00798-z

nature catalysis

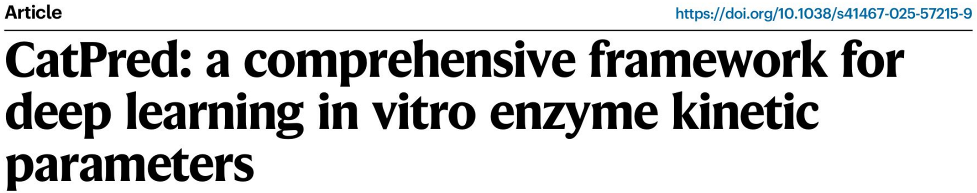
Check for updates

OPEN

Deep learning-based k_{cat} prediction enables improved enzyme-constrained model reconstruction

CatPred ↓

nature communications



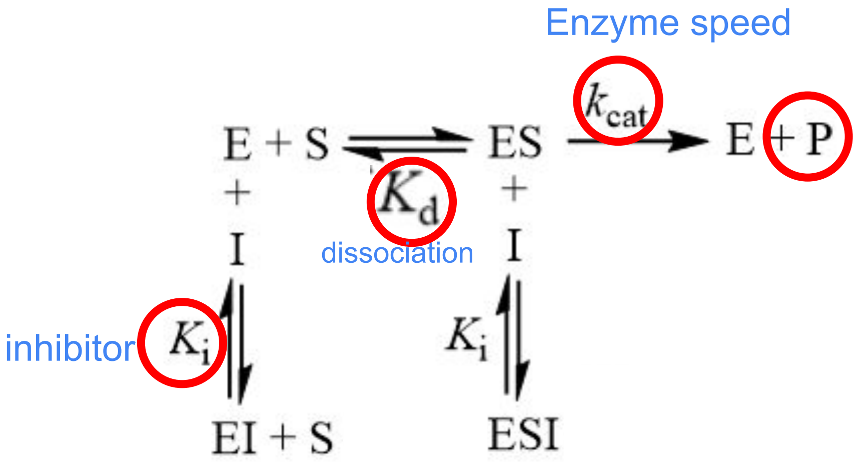
Article

https://doi.org/10.1038/s41467-025-57215-9

CatPred: a comprehensive framework for deep learning in vitro enzyme kinetic parameters

Background

- Predict enzyme reactivity and product
- It can help us to understand the reaction without experiment



$$K_d = \frac{[E][S]}{[ES]} = \frac{k_{-1}}{k_1}$$

dissociation

$$K_m = \frac{k_{-1+}k_{cat}}{k_1} = [S]@\frac{1}{2}v_{max}$$

Parameter	Unit	Meaning	Biological Interpretation
kcat	s^{-1}	Turnover number: number of substrate molecules converted per second per enzyme active site	Enzyme speed
Km	M, mM, µM	Michaelis constant: substrate concentration at half-maximal velocity (Vmax/2)	Substrate affinity
Ki	M, µM	Inhibition constant: binding affinity between enzyme and inhibitor	Inhibitor strength

Aims

1. Which regression models is the best for enzyme reactivity prediction?

2. What if test set has new proteins/molecules? (cold-protein/cold-molecule)

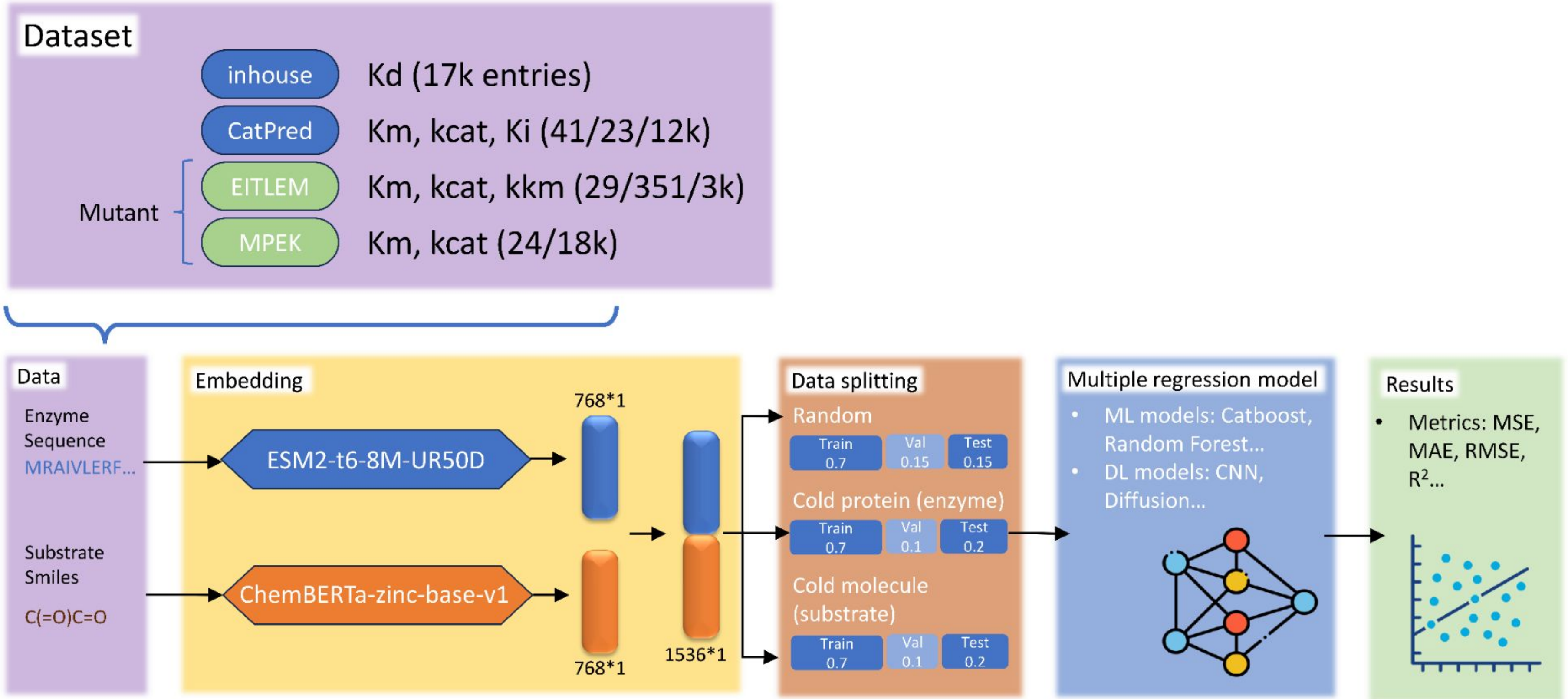
3. Can we predict kinetic parameters of mutant enzyme?

- What's new
  - new splitting methods
  - mutant enzyme
  - embedding method

Methods

Bo-Wei

Methods: Overview



Dataset

- inhouse: Kd (17k entries)
- CatPred: Km, kcat, Ki (41/23/12k)
- EITLEM: Km, kcat, kkm (29/351/3k)
- MPEK: Km, kcat (24/18k)

Mutant: EITLEM, MPEK

Data

Enzyme Sequence MRAIVLERF...

Substrate Smiles C(=O)C=O

Embedding

ESM2-t6-8M-UR50D → 768*1

ChemBERTa-zinc-base-v1 → 768*1

→ 1536*1

Data splitting

Random

Train	Val	Test
0.7	0.15	0.15

Cold protein (enzyme)

Train	Val	Test
0.7	0.1	0.2

Cold molecule (substrate)

Train	Val	Test
0.7	0.1	0.2

Multiple regression model

- ML models: Catboost, Random Forest...
- DL models: CNN, Diffusion...

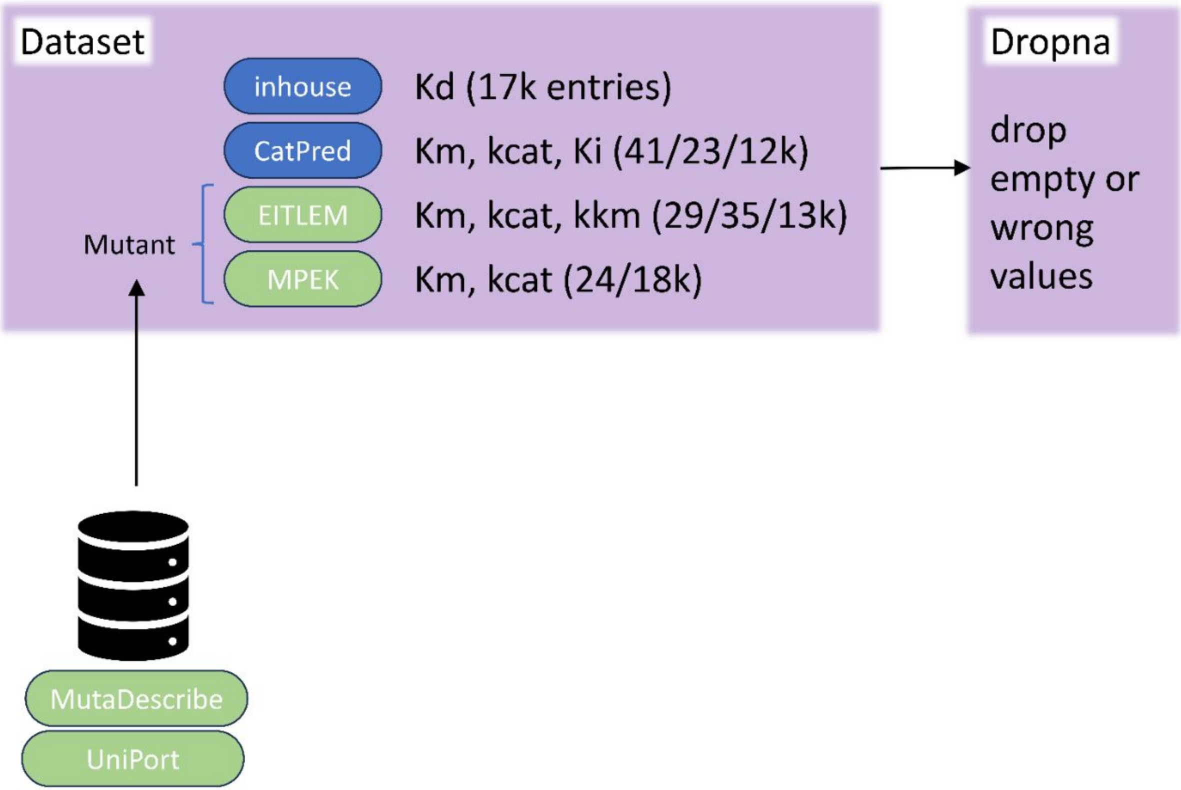
Results

- Metrics: MSE, MAE, RMSE, R^2...

https://cdn-icons-png.flaticon.com/512/9304/9304615.png

https://cdn-icons-png.flaticon.com/512/8253/8253535.png

Methods: Data



Dataset

inhouse Kd (17k entries)

CatPred Km, kcat, Ki (41/23/12k)

Mutant

EITLEM Km, kcat, kkm (29/35/13k)

MPEK Km, kcat (24/18k)

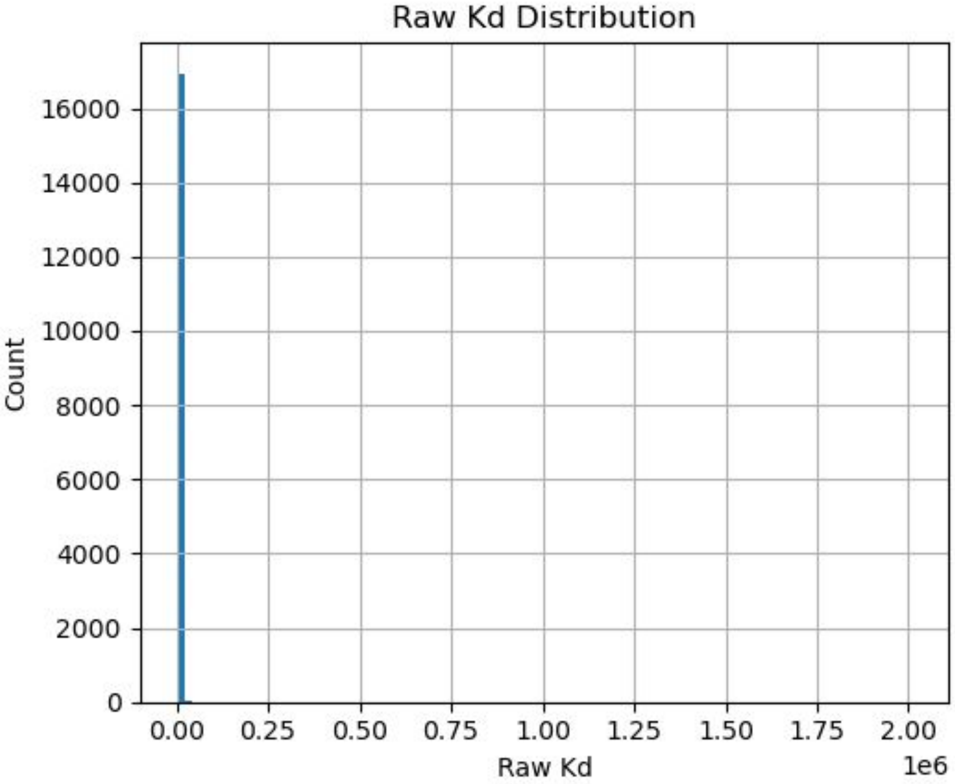
Dropna

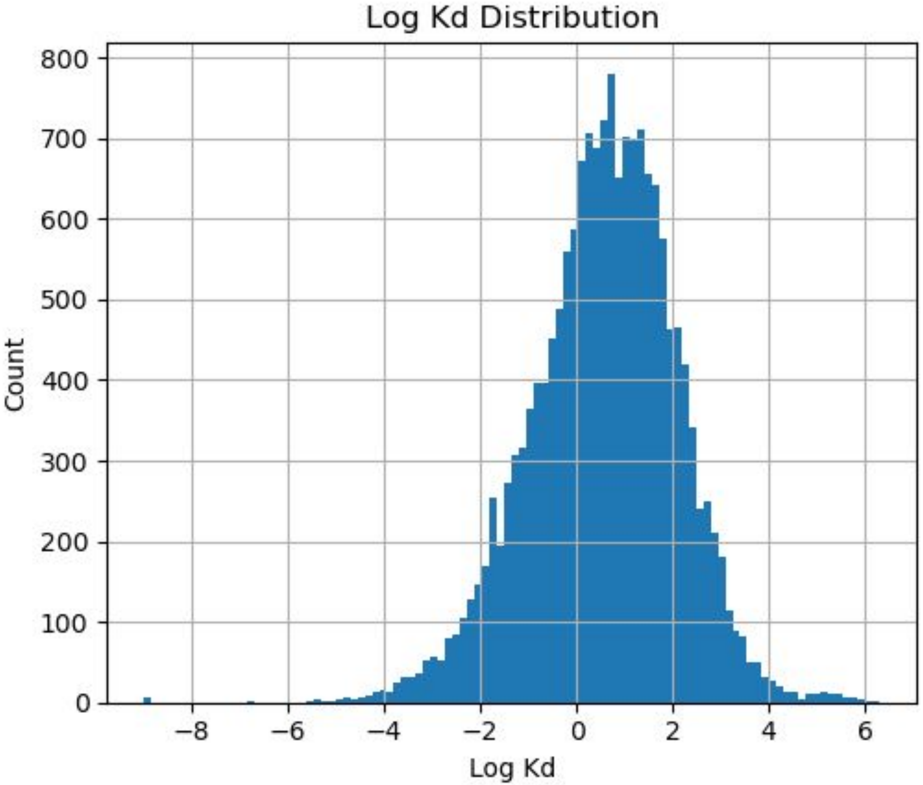
drop empty or wrong values

MutaDescribe

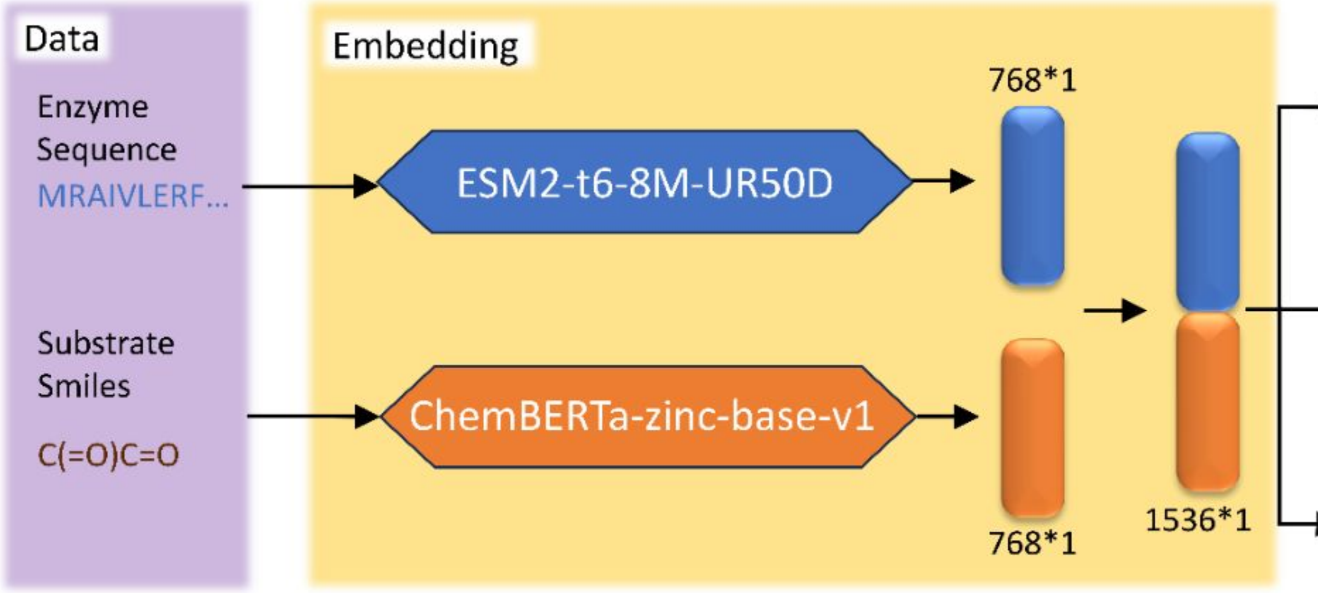
UniPort

Methods: Data





Methods: embedding



Methods: regression model training & evaluation

3 splitting methods

10 models

7+1 metrics

Methods: regression model training & evaluation

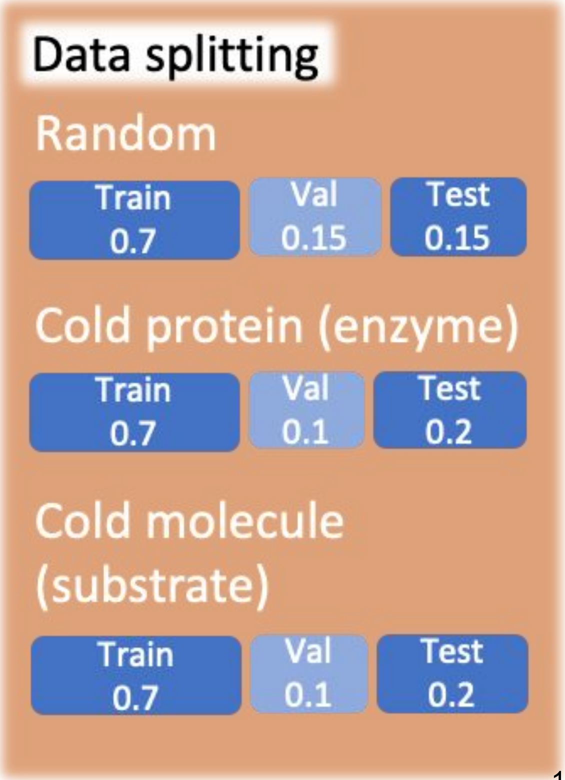
3 splitting methods

- random
- cold protein: to ensure validation and test protein are unseen in training
- cold molecule: to ensure validation and test molecule are unseen in training

purpose: to evaluate model generalizability under different biological scenarios

10 models

7+1 metrics



Data splitting

Random

Train 0.7 | Val 0.15 | Test 0.15

Cold protein (enzyme)

Train 0.7 | Val 0.1 | Test 0.2

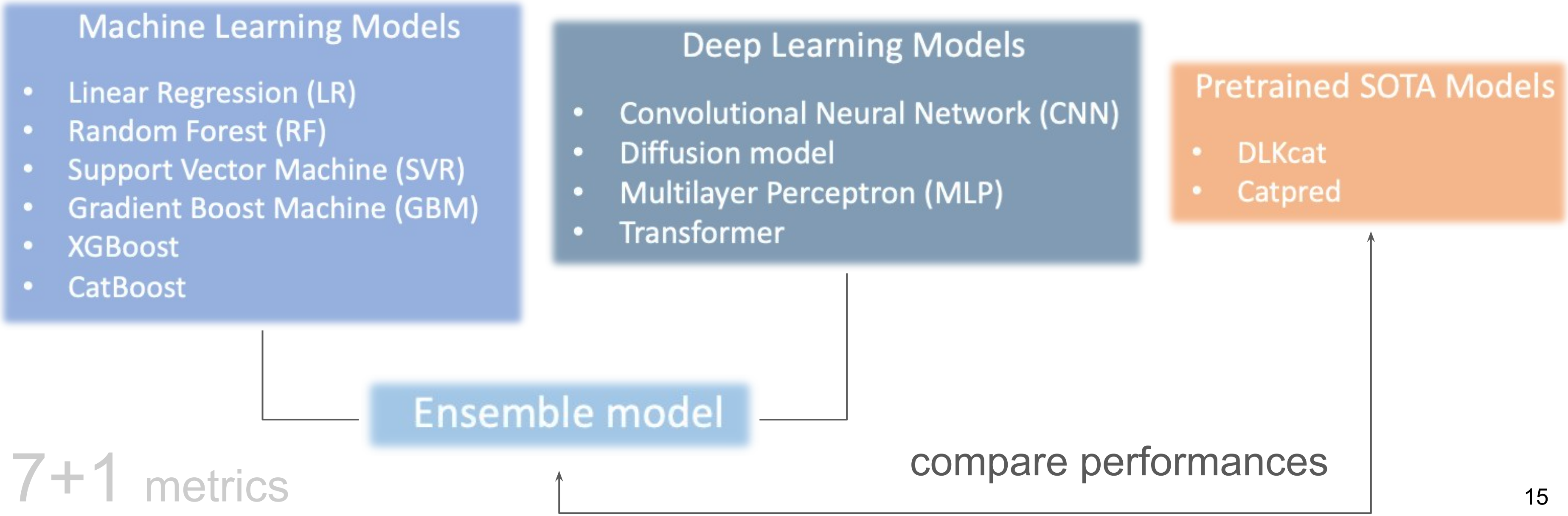
Cold molecule (substrate)

Train 0.7 | Val 0.1 | Test 0.2

Methods: regression model training & evaluation

3 splitting methods: random, cold protein, cold molecule

10 models



Machine Learning Models

- Linear Regression (LR)
- Random Forest (RF)
- Support Vector Machine (SVR)
- Gradient Boost Machine (GBM)
- XGBoost
- CatBoost

Deep Learning Models

- Convolutional Neural Network (CNN)
- Diffusion model
- Multilayer Perceptron (MLP)
- Transformer

Pretrained SOTA Models

- DLKcat
- Catpred

Ensemble model

compare performances

7+1 metrics

Methods: regression model training & evaluation

7+1 metrics

Metric	Formula	Meaning
MAE	$\frac{1}{n}\sum_{i=1}^{n}\lvert y_i - \hat{y}_i\rvert$	reflects prediction accuracy; lower is better
MSE	$\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$	reflects prediction accuracy; lower is better
RMSE	$\sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}$	reflects prediction accuracy; lower is better
Median Absolute Error	$median\left(\lvert y_1 - \hat{y}_1\rvert, \ldots, \lvert y_n - \hat{y}_n\rvert\right)$	reflects prediction accuracy; lower is better
Pearson	$\frac{\sum(y_i-\bar{y})(\hat{y}_i-\bar{\hat{y}})}{\sqrt{\sum(y_i-\bar{y})^2\sum(\hat{y}_i-\bar{\hat{y}})^2}}$	reflects prediction trend; closer to ±1 is better
R^2	$1 - \frac{\sum(y_i-\hat{y}_i)^2}{\sum(y_i-\bar{y})^2}$	reflects explanatory power; higher is better
Explained Variance	$1 - \frac{\mathrm{Var}(y-\hat{y})}{\mathrm{Var}(y)}$	reflects explanatory power; higher is better

Methods: regression model training & evaluation

7+1 metrics

Metric	Formula	Meaning
MAE	$\frac{1}{n}\sum_{i=1}^{n}\lvert y_i-\hat{y}_i\rvert$	reflects prediction accuracy; lower is better
MSE	$\frac{1}{n}\sum_{i=1}^{n}(y_i-\hat{y}_i)^2$	reflects prediction accuracy; lower is better
RMSE	$\sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i-\hat{y}_i)^2}$	reflects prediction accuracy; lower is better
Median Absolute Error	$median\left(\lvert y_1-\hat{y}_1\rvert,\ldots,\lvert y_n-\hat{y}_n\rvert\right)$	reflects prediction accuracy; lower is better
Pearson	$\frac{\sum(y_i-\bar{y})(\hat{y}_i-\bar{\hat{y}})}{\sqrt{\sum(y_i-\bar{y})^2\sum(\hat{y}_i-\bar{\hat{y}})^2}}$	reflects prediction trend; closer to ±1 is better
R²	$1-\frac{\sum(y_i-\hat{y}_i)^2}{\sum(y_i-\bar{y})^2}$	reflects explanatory power; higher is better
Explained Variance	$1-\frac{\mathrm{Var}(y-\hat{y})}{\mathrm{Var}(y)}$	reflects explanatory power; higher is better

Methods: regression model training & evaluation

7+1 metrics

Metric	Formula	Meaning
MAE	$\frac{1}{n}\sum_{i=1}^{n}\lvert y_i - \hat{y}_i\rvert$	reflects prediction accuracy; lower is better
MSE	$\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$	reflects prediction accuracy; lower is better
RMSE	$\sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}$	reflects prediction accuracy; lower is better
Median Absolute Error	$median(\lvert y_1 - \hat{y}_1\rvert, \ldots, \lvert y_n - \hat{y}_n\rvert)$	reflects prediction accuracy; lower is better
Pearson	$\frac{\sum(y_i-\bar{y})(\hat{y}_i-\bar{\hat{y}})}{\sqrt{\sum(y_i-\bar{y})^2\sum(\hat{y}_i-\bar{\hat{y}})^2}}$	reflects prediction trend; closer to ±1 is better
R²	$1-\frac{\sum(y_i-\hat{y}_i)^2}{\sum(y_i-\bar{y})^2}$	reflects explanatory power; higher is better
Explained Variance	$1-\frac{\mathrm{Var}(y-\hat{y})}{\mathrm{Var}(y)}$	reflects explanatory power; higher is better

Methods: regression model training & evaluation

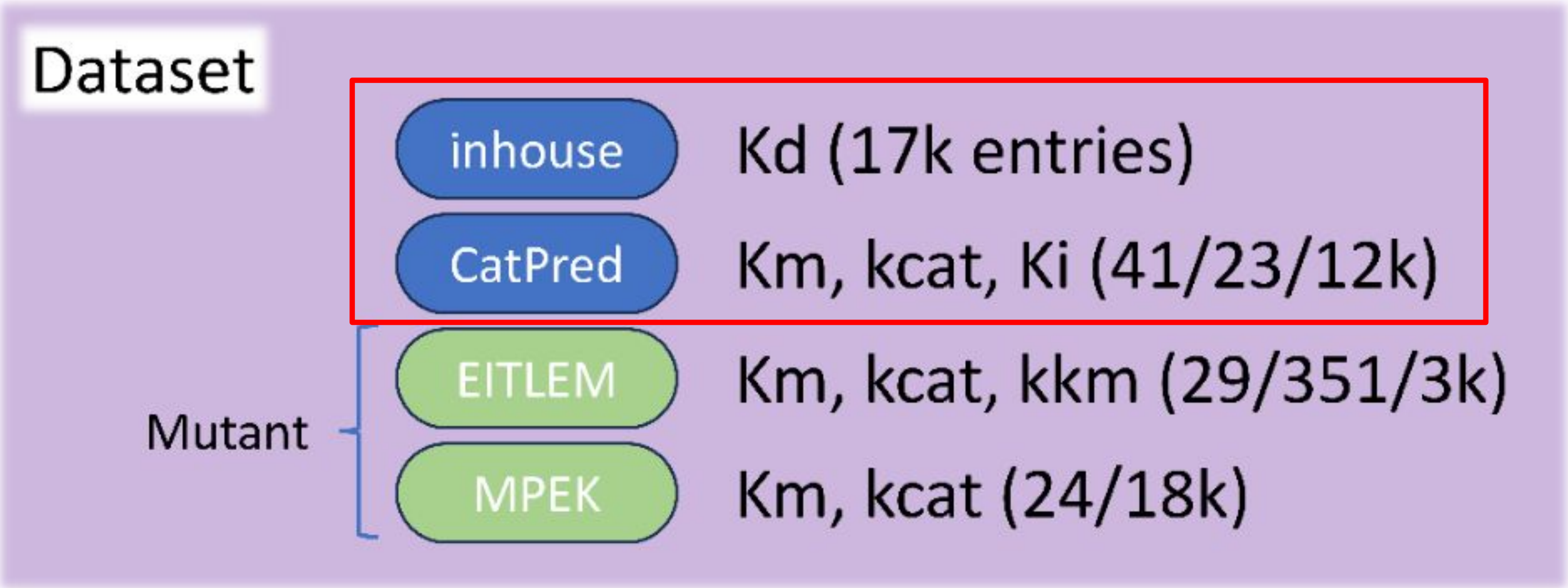
7+1 metrics

Rank Score: sort model performance for each metric to determine their rank, then sum the ranks across different splitting methods to calculate the rank score

Metric	Formula	Meaning
MAE	$\frac{1}{n}\sum_{i=1}^{n}\lvert y_i - \hat{y}_i \rvert$	reflects prediction accuracy; lower is better
MSE	$\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$	reflects prediction accuracy; lower is better
RMSE	$\sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}$	reflects prediction accuracy; lower is better
Median Absolute Error	$median(\lvert y_1 - \hat{y}_1 \rvert, \ldots, \lvert y_n - \hat{y}_n \rvert)$	reflects prediction accuracy; lower is better
Pearson	$\frac{\sum(y_i - \bar{y})(\hat{y}_i - \bar{\hat{y}})}{\sqrt{\sum(y_i - \bar{y})^2 \sum(\hat{y}_i - \bar{\hat{y}})^2}}$	reflects prediction trend; closer to ±1 is better
R²	$1 - \frac{\sum(y_i - \hat{y}_i)^2}{\sum(y_i - \bar{y})^2}$	reflects explanatory power; higher is better
Explained Variance	$1 - \frac{\mathrm{Var}(y - \hat{y})}{\mathrm{Var}(y)}$	reflects explanatory power; higher is better

Results

Results

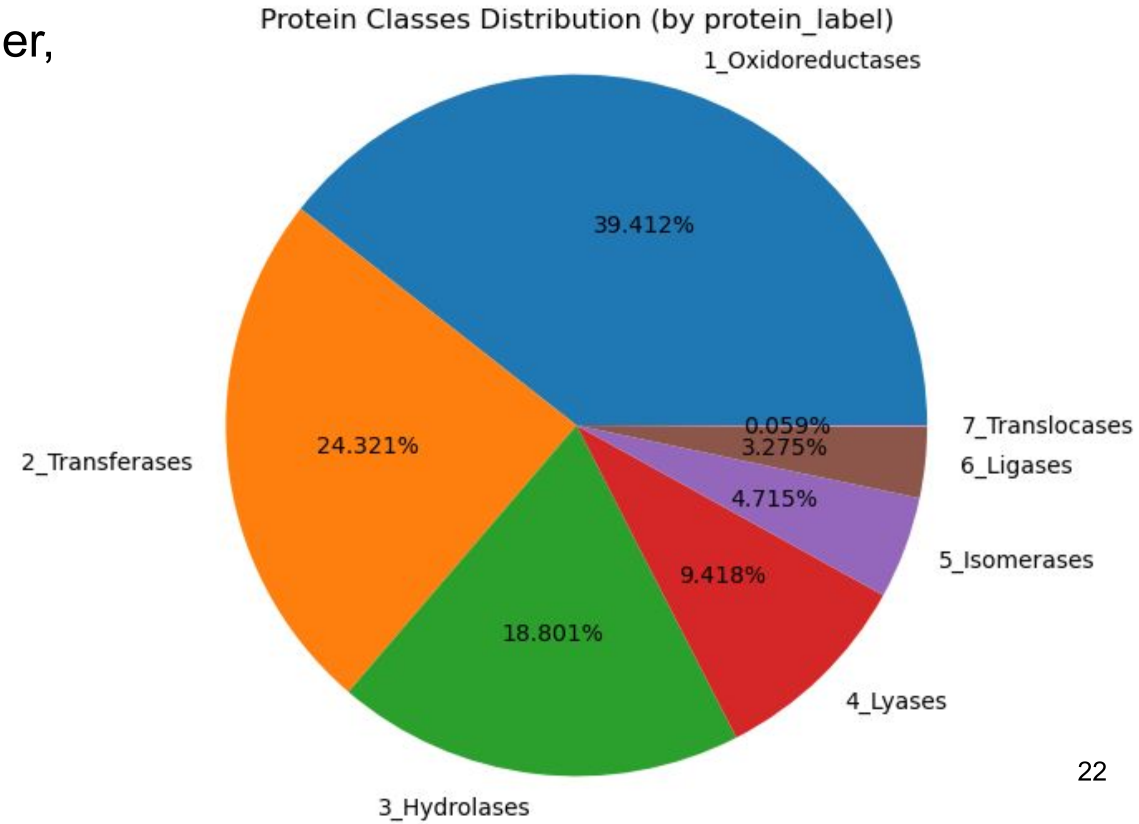


Dataset

	inhouse	Kd (17k entries)
	CatPred	Km, kcat, Ki (41/23/12k)
Mutant	EITLEM	Km, kcat, kkm (29/351/3k)
	MPEK	Km, kcat (24/18k)

Result: Enzyme class distribution overview

- Enzyme Commission Number, or EC Number
- Each EC Number has four parts. (e.g., EC 1.1.1.1).



Results: Model performances

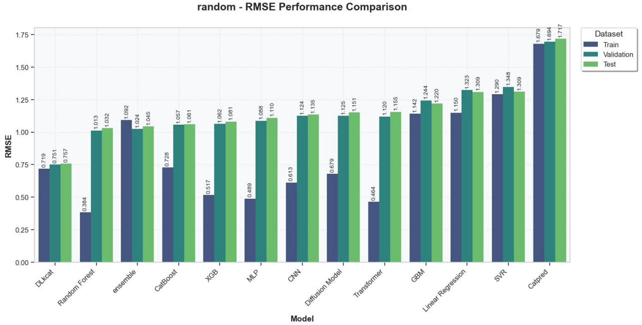
Rank score of ML and DL models

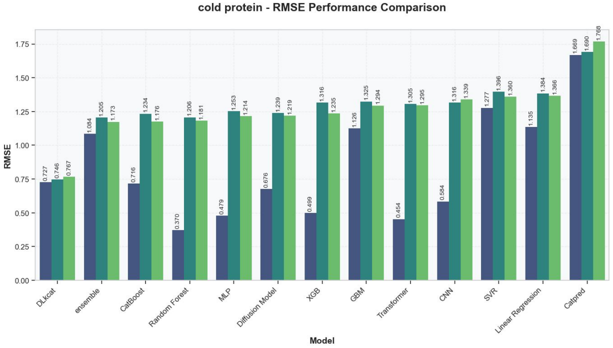
model	Catpred_kcat	Catpred_Km	Catpred_Ki	inhouse_Kd
MLP	45	45	50	66
Random Forest	57	77	79	29
CatBoost	69	75	102	91
Diffusion Model	93	108	57	117
Transformer	120	66	101	116
XGB	125	120	139	90
CNN	142	141	148	143
GBM	159	185	162	146
SVR	184	155	157	195
Linear Regression	201	209	210	204

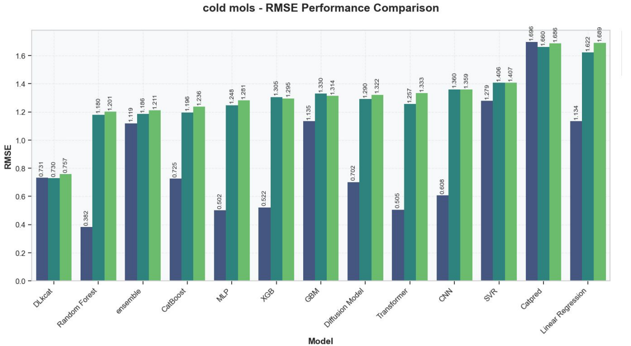
Results: In-house dataset

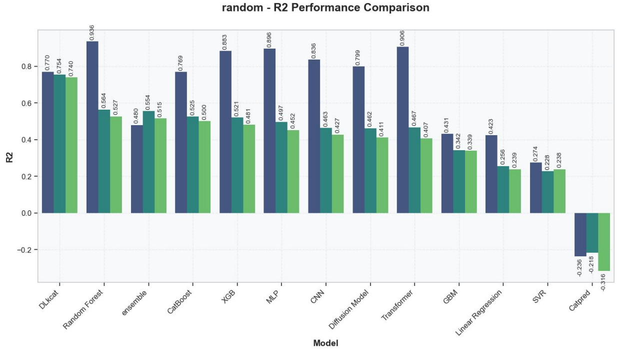
model	DLKcat	ensemble	RF	CatBoost	MLP	XGB	Transformer	Diffusion Model	CNN	GBM	SVR	LR	Catpred
Rank Score	21	56	58	91	108	132	157	160	186	190	241	255	256

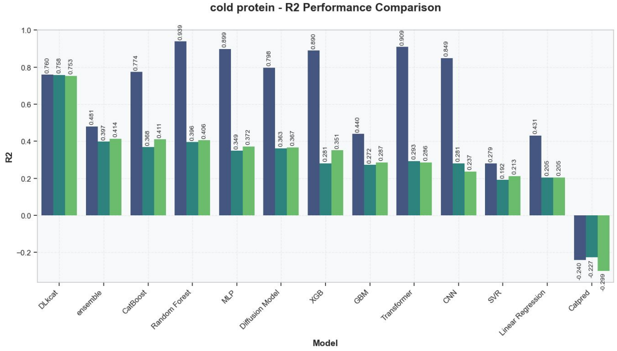
Results: In-house dataset

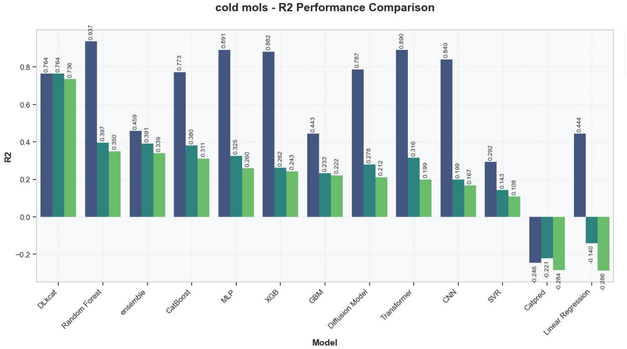

random - RMSE Performance Comparison


cold protein - RMSE Performance Comparison

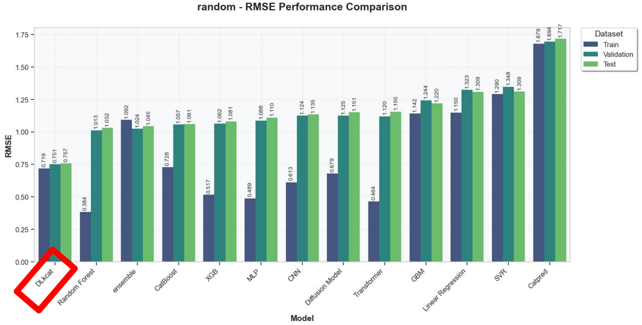

cold mols - RMSE Performance Comparison

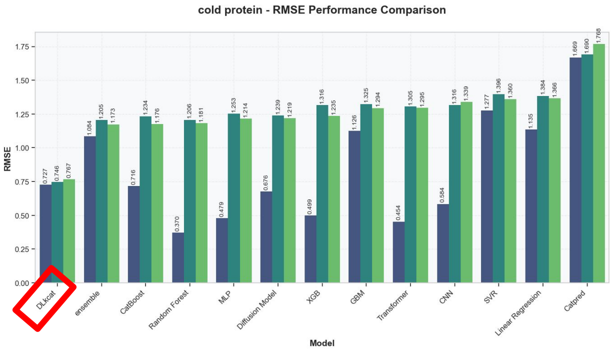

random - R2 Performance Comparison

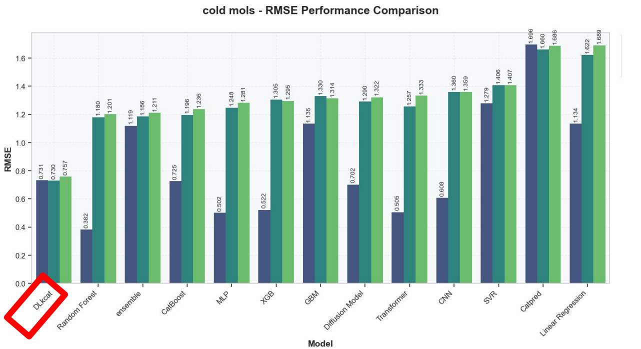

cold protein - R2 Performance Comparison


cold mols - R2 Performance Comparison

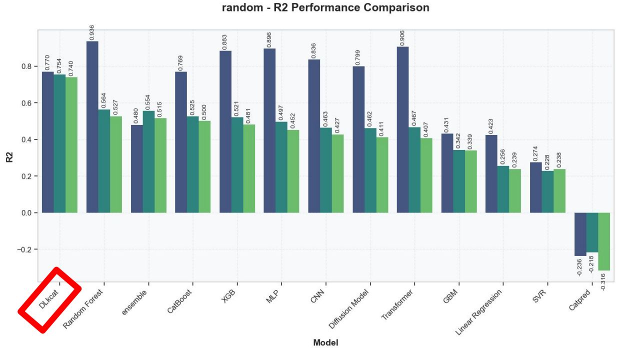
Results: In-house dataset

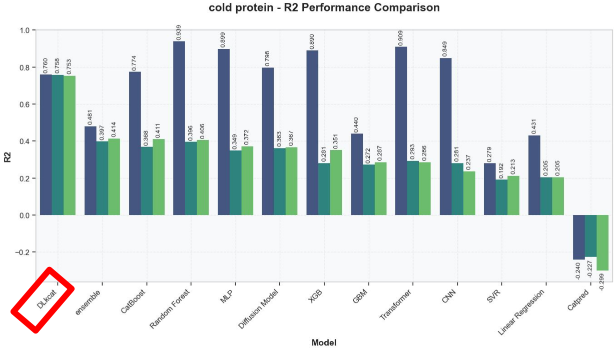

random - RMSE Performance Comparison

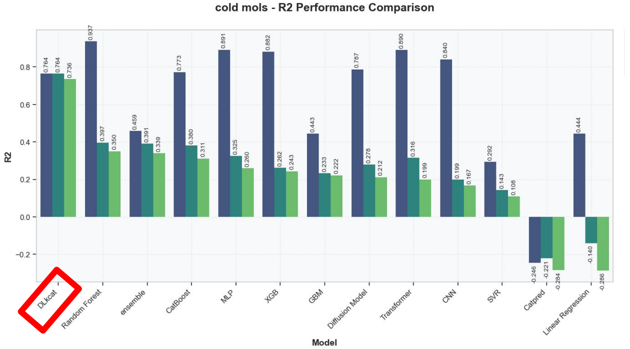

cold protein - RMSE Performance Comparison


cold mols - RMSE Performance Comparison

DLKcat performed the best across different splitting methods in the in-house dataset

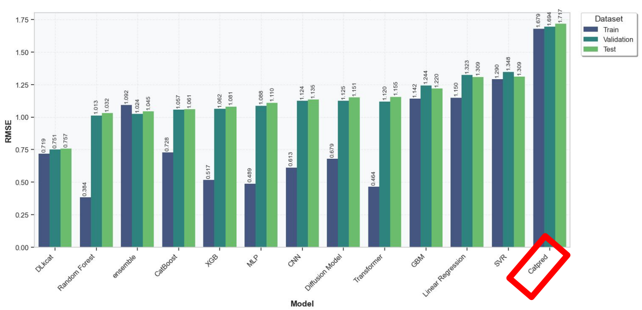

random - R2 Performance Comparison


cold protein - R2 Performance Comparison

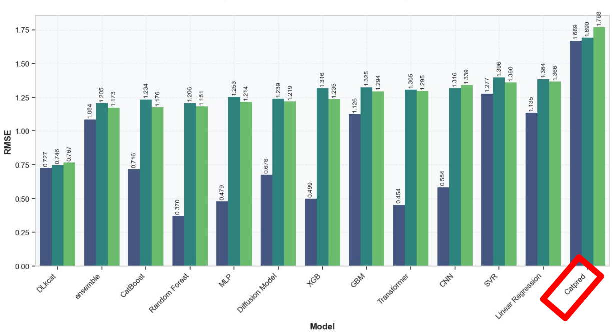

cold mols - R2 Performance Comparison

Results: In-house dataset

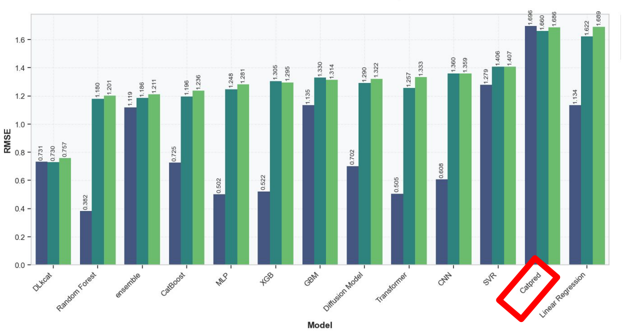
random - RMSE Performance Comparison



cold protein - RMSE Performance Comparison

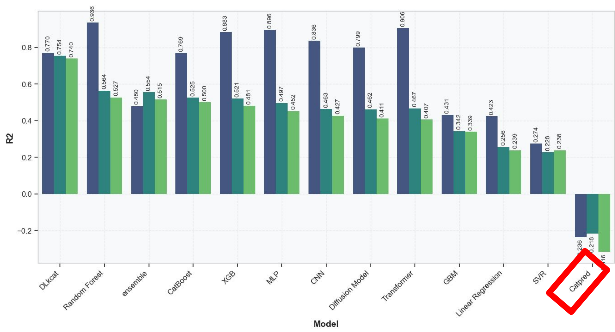


cold mols - RMSE Performance Comparison

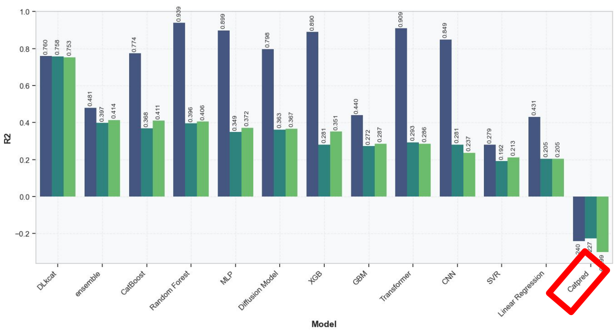


Catpred exhibited poor predictive performance on the in-house dataset

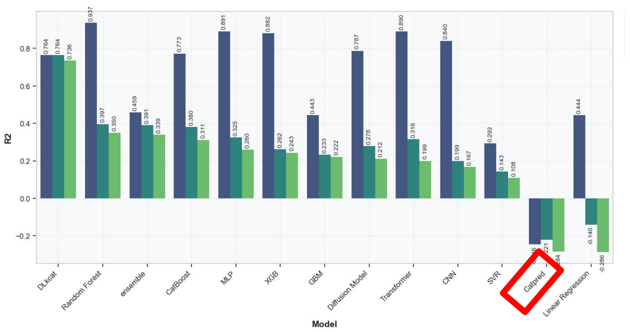
random - R2 Performance Comparison



cold protein - R2 Performance Comparison



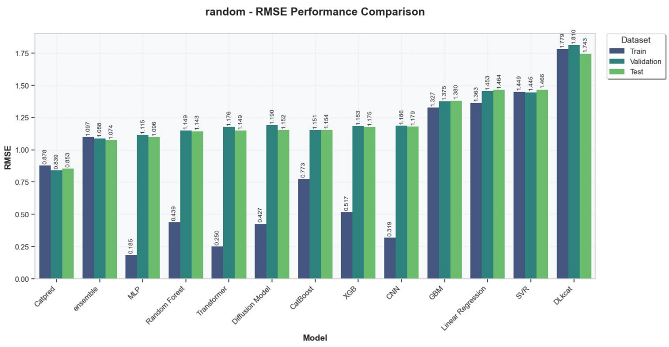
cold mols - R2 Performance Comparison



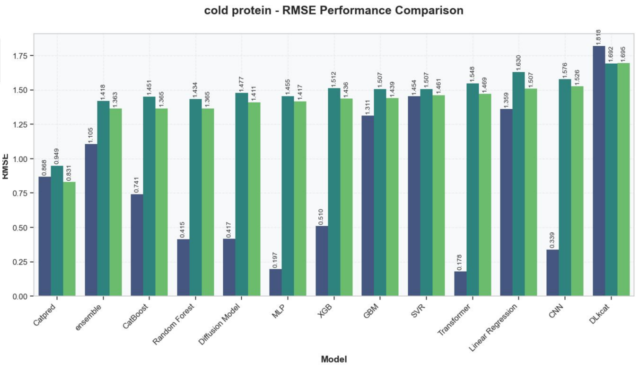
Results: Catpred dataset

model	Catpred	ensemble	MLP	RF	CatBoost	Transformer	Diffusion Model	XGB	CNN	GBM	SVR	LR	DLkcat
kcat	21	48	84	98	111	120	135	167	184	201	226	243	273
Km	28	45	84	119	117	66	150	162	183	227	197	251	-
Ki	21	56	84	121	102	141	95	181	190	204	199	252	-

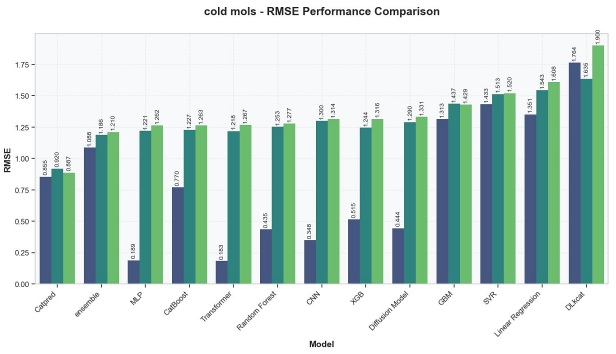
Results: Catpred dataset (take kcat as an example)



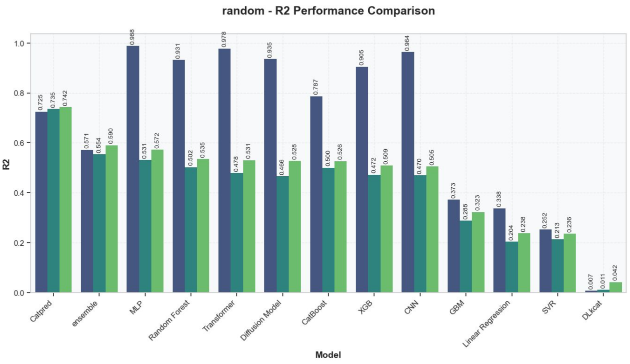
random - RMSE Performance Comparison



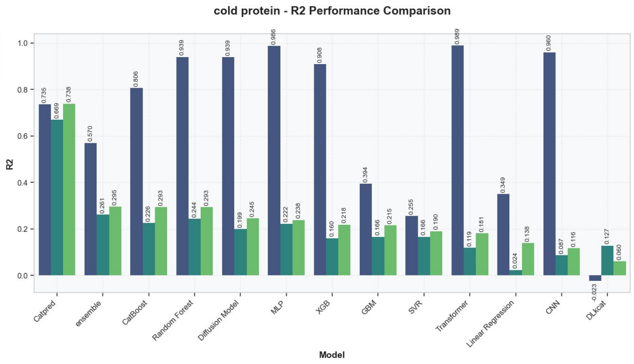
cold protein - RMSE Performance Comparison



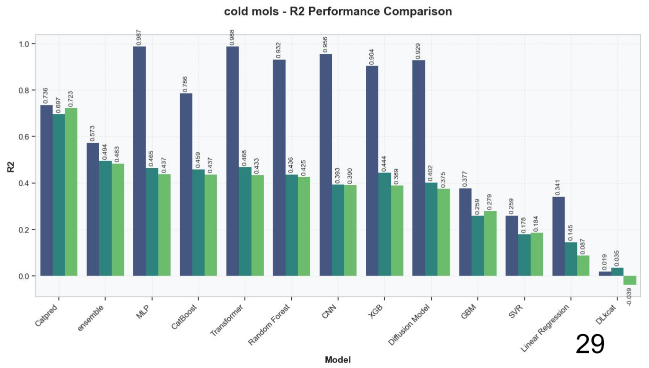
cold mols - RMSE Performance Comparison



random - R2 Performance Comparison

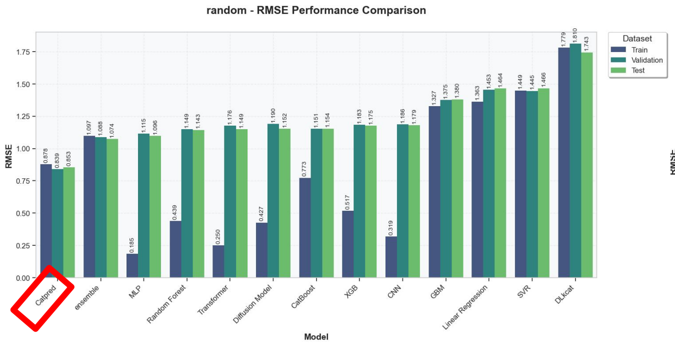


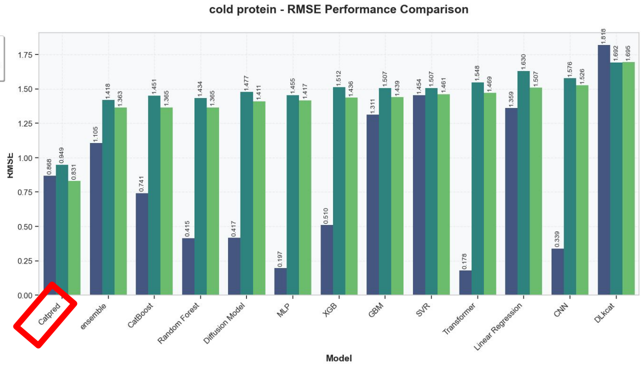
cold protein - R2 Performance Comparison

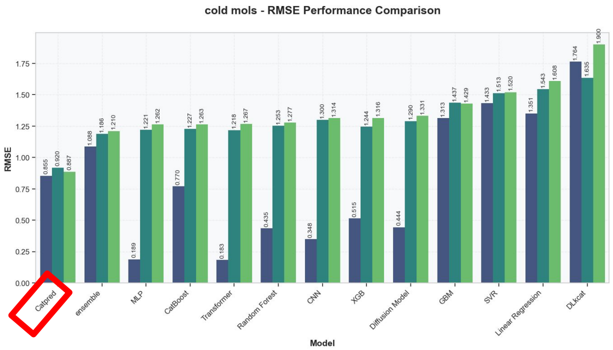


cold mols - R2 Performance Comparison

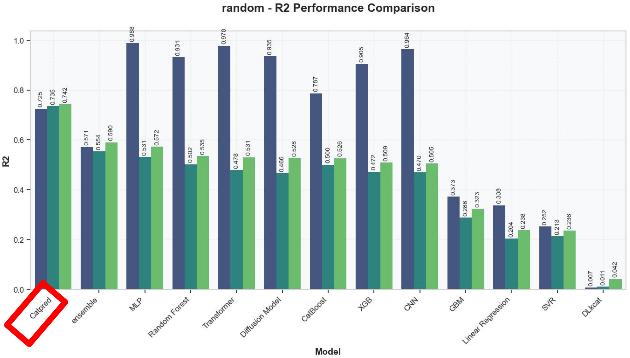
Results: Catpred dataset (take kcat as an example)

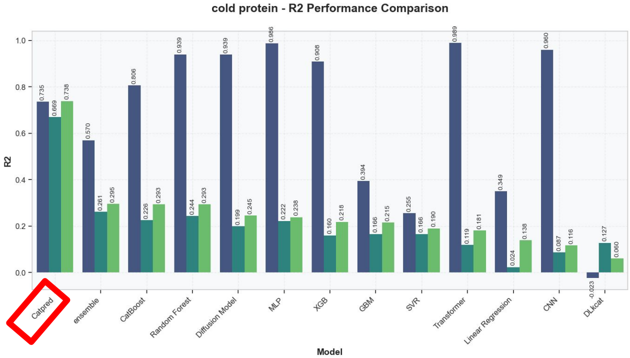

random - RMSE Performance Comparison

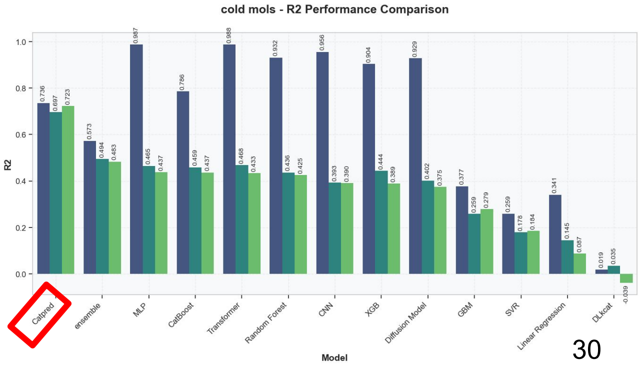

cold protein - RMSE Performance Comparison


cold mols - RMSE Performance Comparison

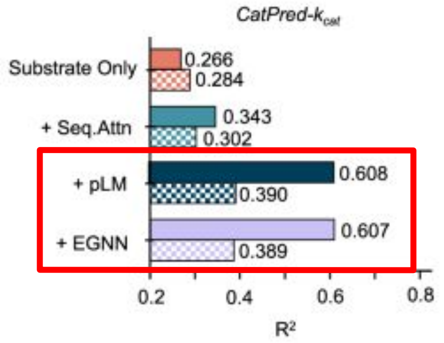
Catpred performed the best across different splitting methods in resplit Catpred dataset


random - R2 Performance Comparison


cold protein - R2 Performance Comparison

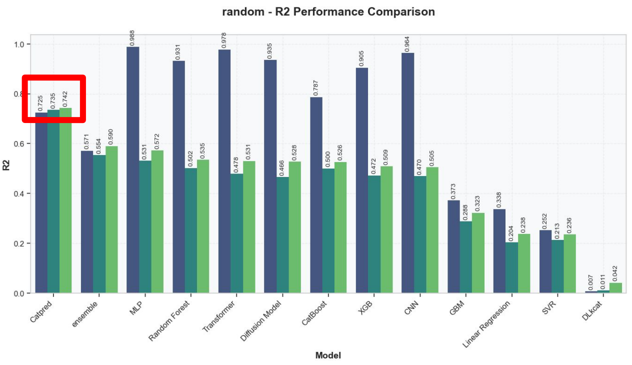

cold mols - R2 Performance Comparison

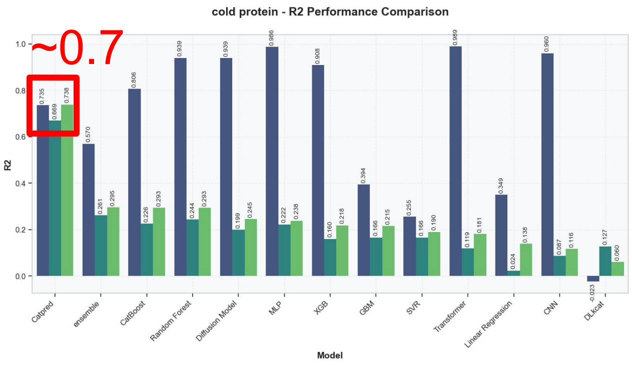
Results: Catpred dataset (take kcat as an example)



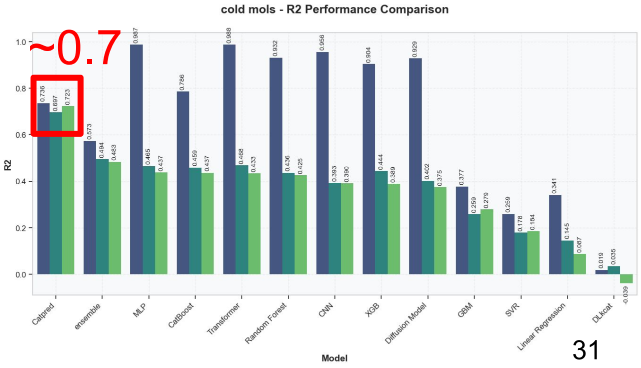
Catpred achieved better performance on the resplit dataset compared to the original study
 → cause: data leakage
 → solution: retrain Catpred model using in-house and resplit Catpred dataset

Boorla, V.S., Maranas, C.D. CatPred: a comprehensive framework for deep learning in vitro enzyme kinetic parameters. *Nat Commun* 16, 2072 (2025).


random - R2 Performance Comparison


cold protein - R2 Performance Comparison

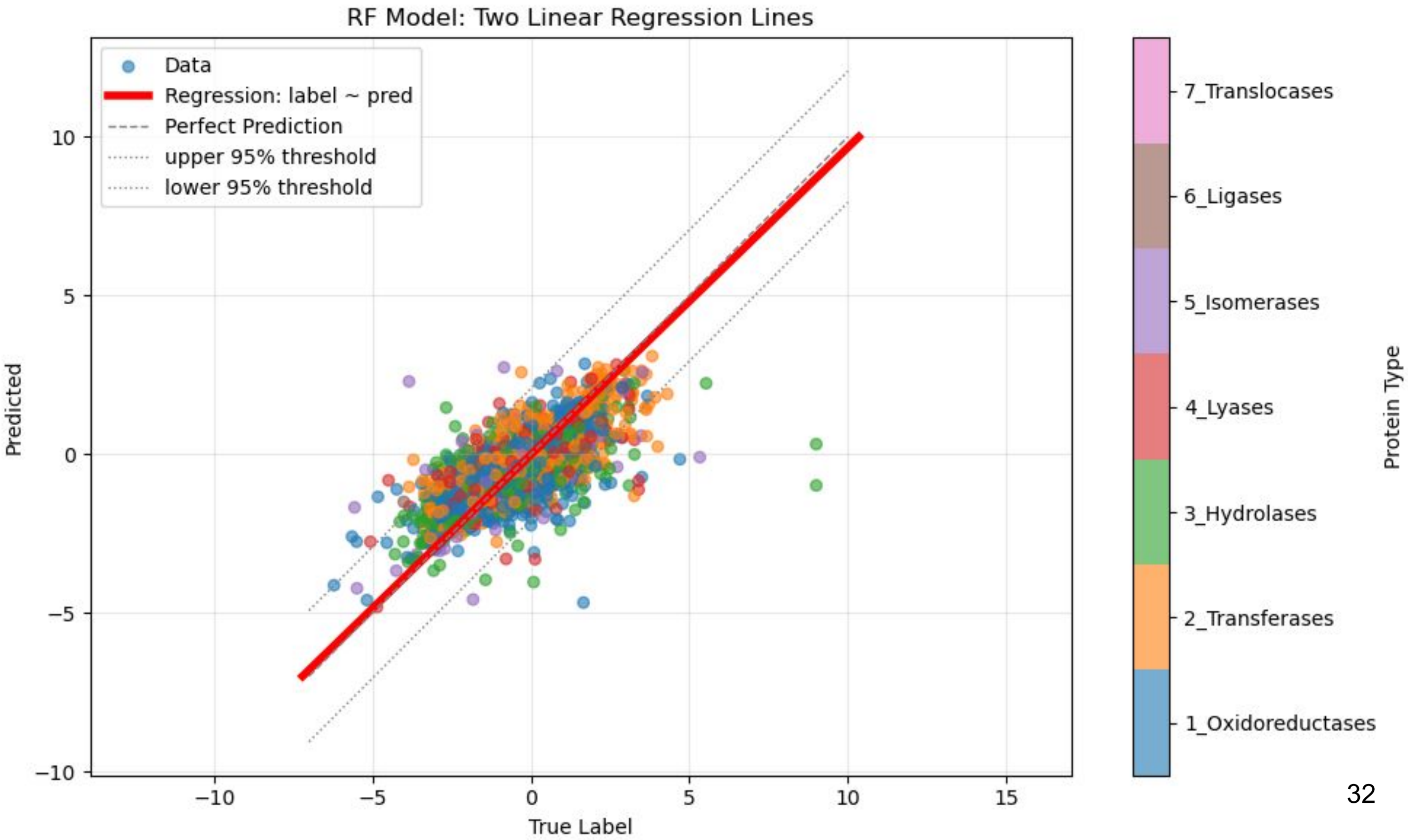
~0.7


cold mols - R2 Performance Comparison

~0.7

Results: Initial results show good prediction performance

- random split
- inhouse dataset
- **95%(5% outlier)**
- model=RF



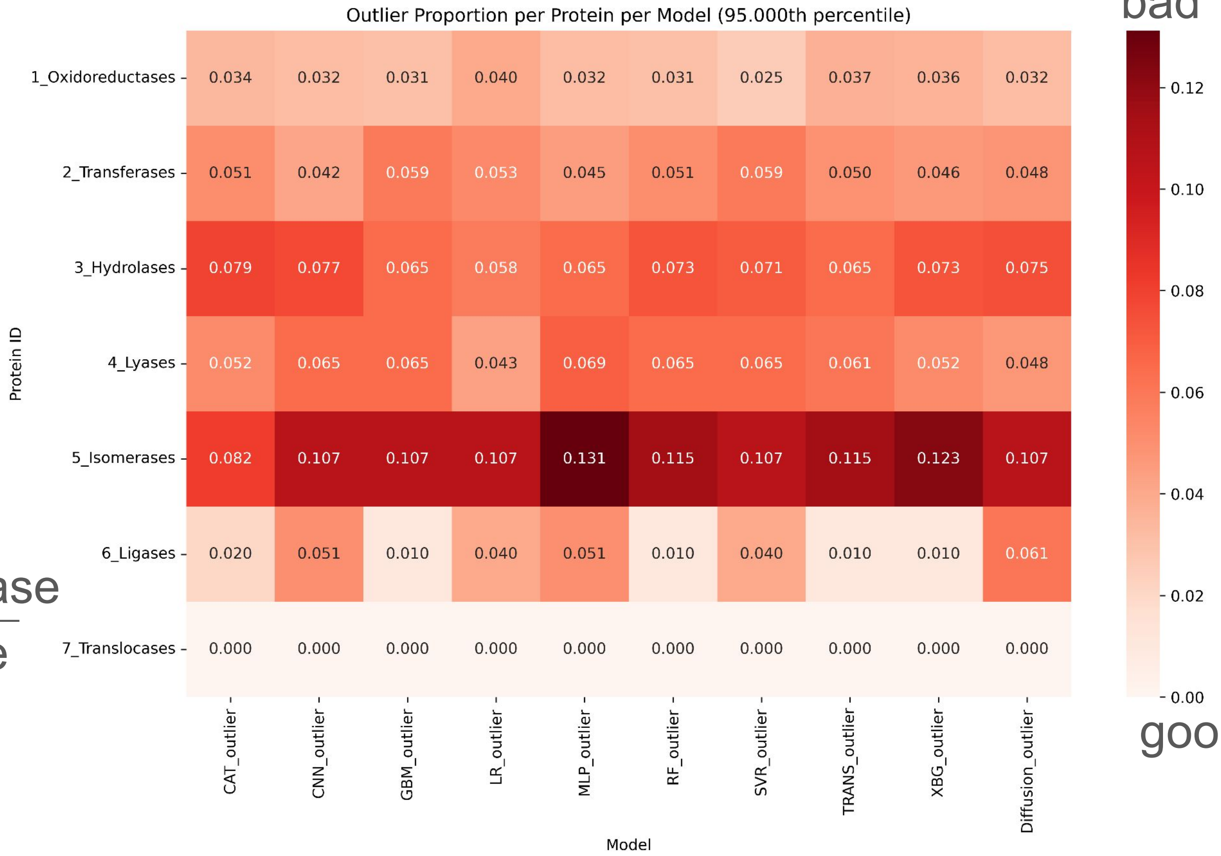
Bo-Wei

Results: Performance in isomerase need to be improved

- random split
- isomerase have most outlier

$$0.131 = \frac{\text{Marked as outlier in isomerase}}{\text{Total number in isomerase}}$$

bad

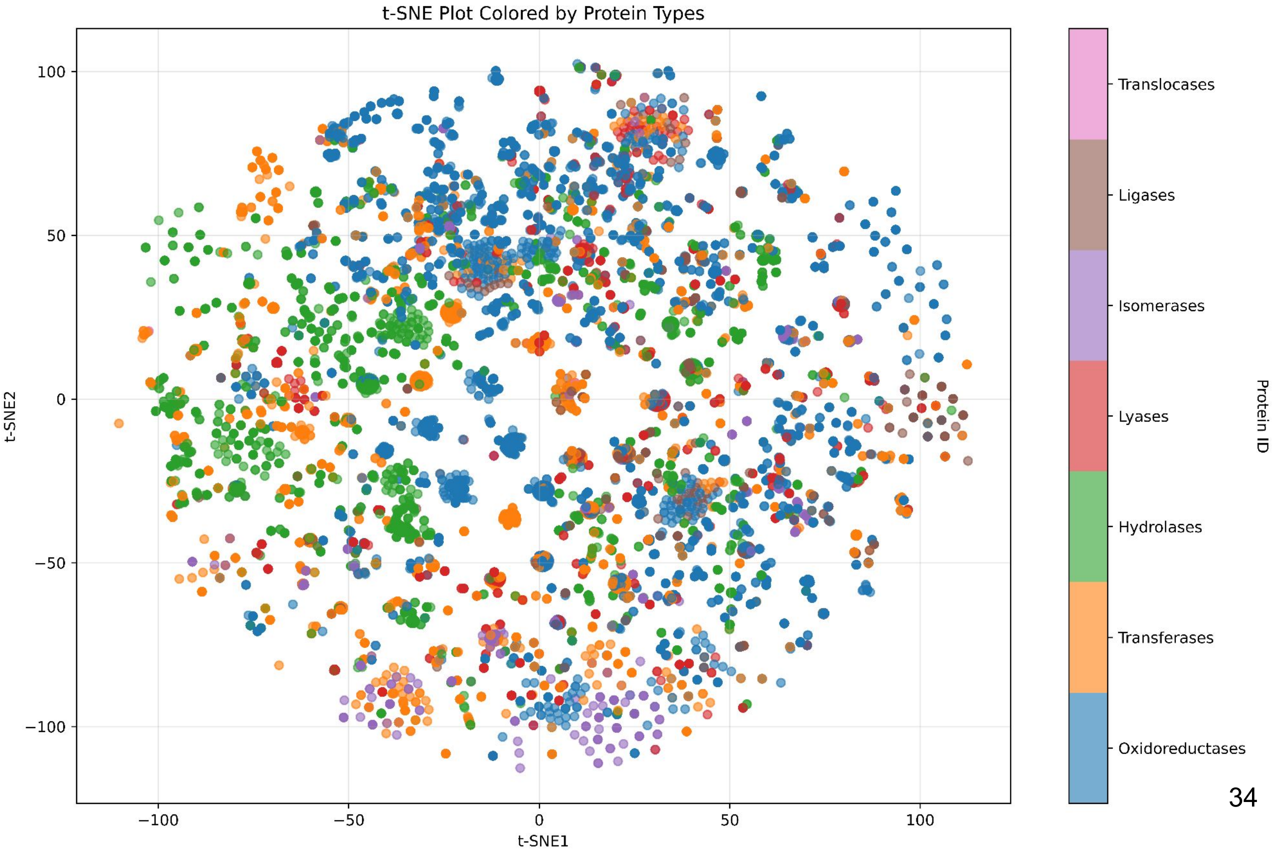


Protein ID	CAT_outlier	CNN_outlier	GBM_outlier	LR_outlier	MLP_outlier	RF_outlier	SVR_outlier	TRANS_outlier	XBG_outlier	Diffusion_outlier
1_Oxidoreductases	0.034	0.032	0.031	0.040	0.032	0.031	0.025	0.037	0.036	0.032
2_Transferases	0.051	0.042	0.059	0.053	0.045	0.051	0.059	0.050	0.046	0.048
3_Hydrolases	0.079	0.077	0.065	0.058	0.065	0.073	0.071	0.065	0.073	0.075
4_Lyases	0.052	0.065	0.065	0.043	0.069	0.065	0.065	0.061	0.052	0.048
5_Isomerases	0.082	0.107	0.107	0.107	0.131	0.115	0.107	0.115	0.123	0.107
6_Ligases	0.020	0.051	0.010	0.040	0.051	0.010	0.040	0.010	0.010	0.061
7_Translocases	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000

good

Results: t-SNE projection of enzyme features

- inhouse dataset



What’s new

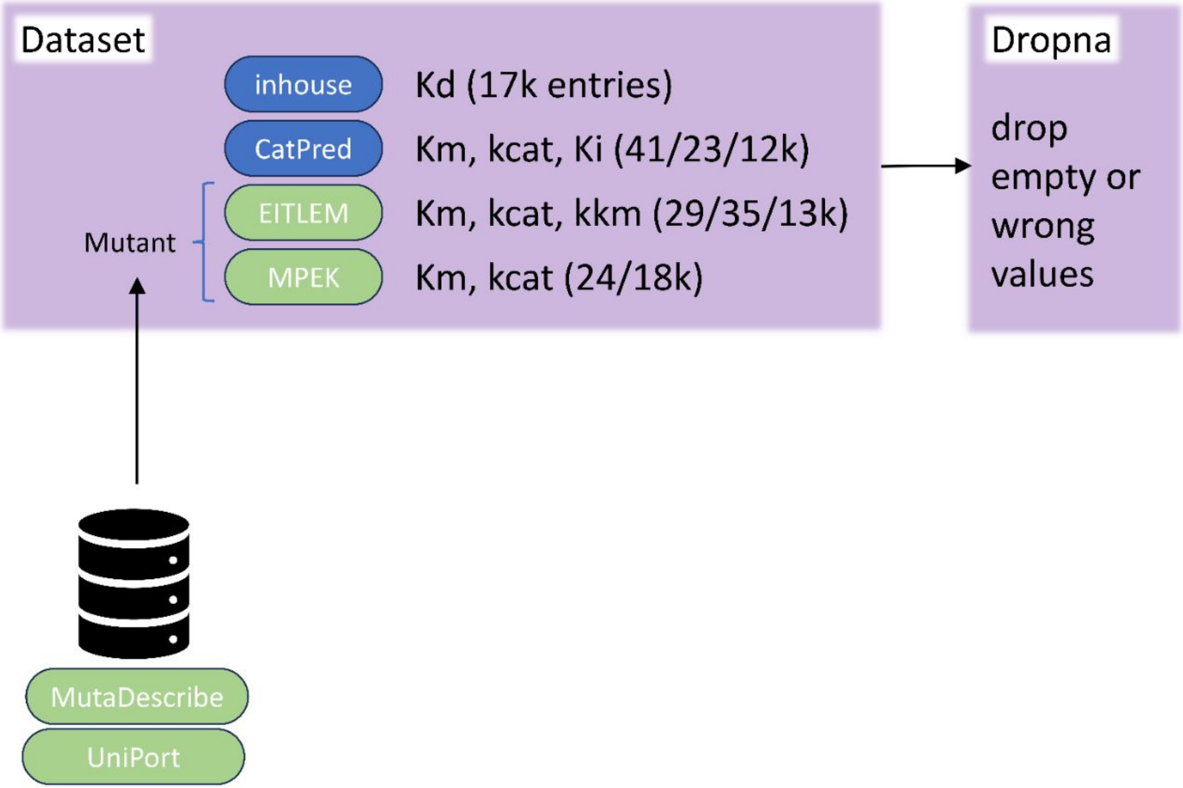
- new splitting methods
- mutant enzyme
- embedding method

Future Work

Future Work

- to retrain SOTA models using wild-type enzyme datasets (in-house and resplit Catpred)
- to develop the model for kinetic parameters prediction of mutant enzyme
- to investigate various protein embedding techniques to enhance the differentiation between wild-type and mutant enzymes in the feature space

Datasets



Dataset

inhouse Kd (17k entries)

CatPred Km, kcat, Ki (41/23/12k)

EITLEM Km, kcat, kkm (29/35/13k)

MPEK Km, kcat (24/18k)

Mutant

MutaDescribe

UniPort

Dropna

drop empty or wrong values

Next steps and timelines

Task and goals	Dataset and Code	Timelines	Person	Meeting?

Train Catpred/DLcat models on in-house dataset, catpred dataset	in-house datasets;
	captured datasets;
		By 08/30/2025	Pin-chi, Cheng	bi-weekly?

Train Catpred/DLcat models on mutant enzyme dataset (EITLEM, MPEK)	EITLEM datasets; MPEK datasets	By 09/15/2025	Pin-chi, Cheng	
Latent embedding comparison and visualization	Visualization of latent embeddings before (ESM2) and after training (model latent layer) for enzyme datasets	By 08/30/2025	Bo-wei	

Next steps and timelines

Task and goals	Dataset and Code	Timelines	Person	Meeting?

Descriptive Visualization, number of enzymes per class, tools
ggplot, or python, **high-quality figures**
	in-house datasets;
	captured datasets;
	EITLEM datasets;
	MPEK datasets
		By 08/30/2025	Bo-wei, Cheng	bi-weekly?

error analysis				

Reflection

Reflection: Pin-Chi

- “I've had a lot of worries in my life, most of which never happened.” — Mark Twain
- “You give a man a fish, he eats for a day; you teach him to fish, he eats forever.”
- A lot of… THANKS!!!!!

Bo-Wei

Reflection: Bowei

- Learned effective data cleaning & prep
- Built a full ML pipeline end-to-end
- Thank Chen for patient guidance