AI-Enabled Virtual Screening: Replicating and Enhancing Machine Learning for Compound Potency Prediction

Research Paper : Janela, T., Bajorath, J. Simple nearest-neighbour analysis meets the accuracy of compound potency predictions using complex machine learning models. Nat Mach Intell 4, 1246–1255 (2022). https://doi.org/10.1038/s42256-022-00581-6

INDRAPRASTHA INSTITUTE *of*
INFORMATION TECHNOLOGY
DELHI

Group Members :

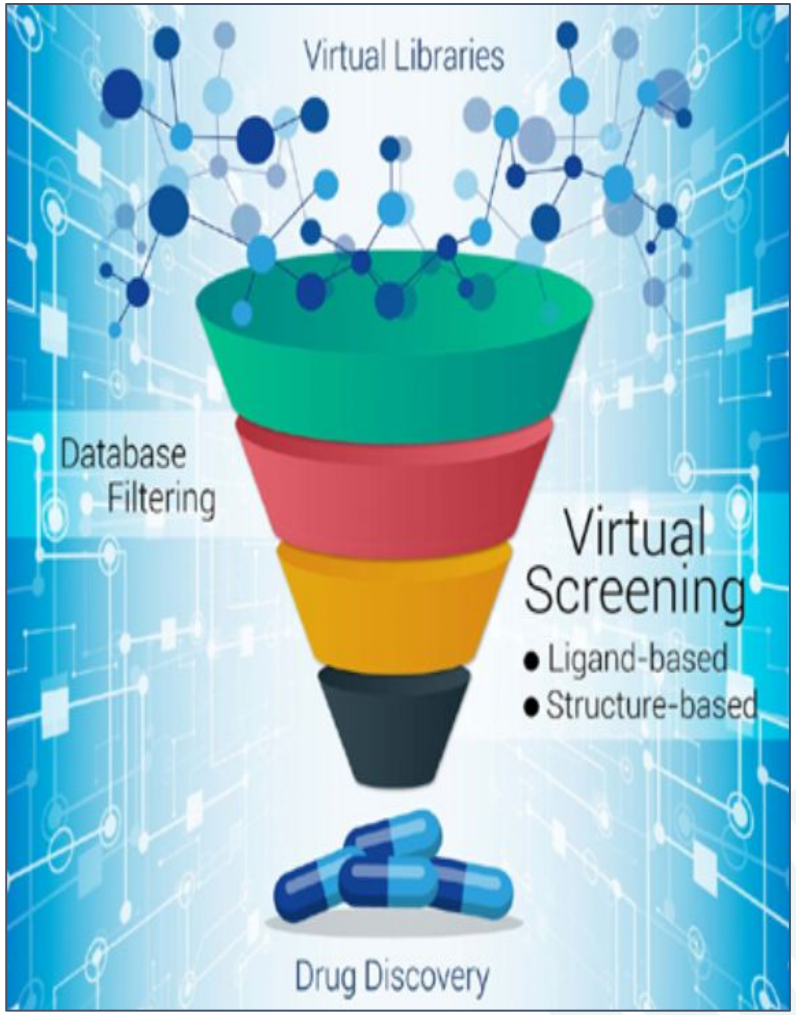
Palak Bhardwaj	2022344
Manya Agrawal	2022281
Yashovardhan Singhal	2022591
Sameer Singh Godara	2022439
Nishchay Sharma	2022331
Sambhav Gautam	2022435

Introduction to Virtual Screening

Powering Drug Discovery: The Virtual Screening Revolution!

- **What?** Computational wizardry to scan chemical libraries and predict drug candidates' binding strength.
- **Why Potency Prediction?** Measures a compound’s effectiveness (e.g., IC50) to prioritize those most likely to succeed in trials.
- **Machine Learning Magic** Predicts potency using molecular features (e.g., fingerprints).
- **Breakthrough Study** Janela & Bajorath (2022): Simple models rival complex ML for efficiency (*Nature Machine Intelligence*).
- **Our Quest** Replicate, enhance, and elevate virtual screening!

Visual Idea: 3D animation of a chemical library funneling through a glowing AI filter, with sparkling drug candidates emerging.



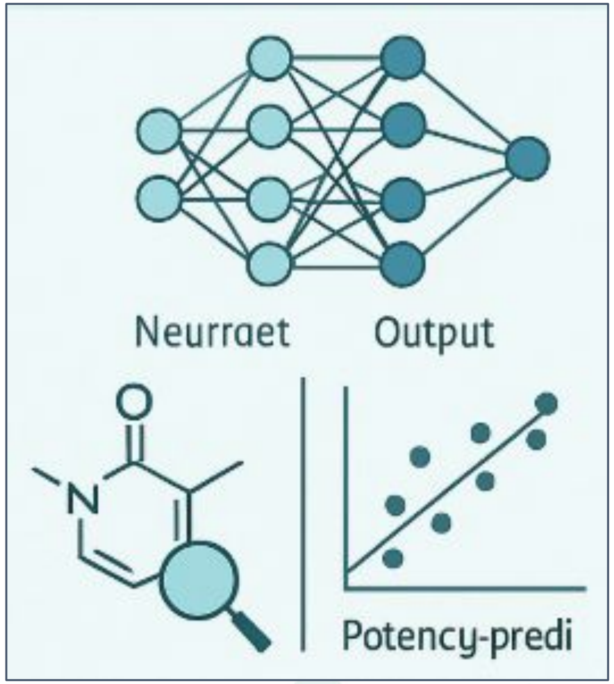
Research Paper Overview

Why we chose it: It's 2022 publication in the high-impact journal *Nature Machine Intelligence* affirms its credibility and relevance.

Primary Objective: Intended to evaluate and the effectiveness of complex machine learning models (for eg. DNNs, GCNs) in predicting compound potency as compared to simpler models like nearest neighbor analysis (KNNs).

Key Findings: Simple nearest-neighbor analysis performed comparable to or better than complex machine learning models for compound potency prediction.

Purpose: This paper was selected for its strong relevance to virtual screening which lead to reshaping perspective in computational drug discovery, emphasizing rigorous benchmarking and pragmatic model selection.



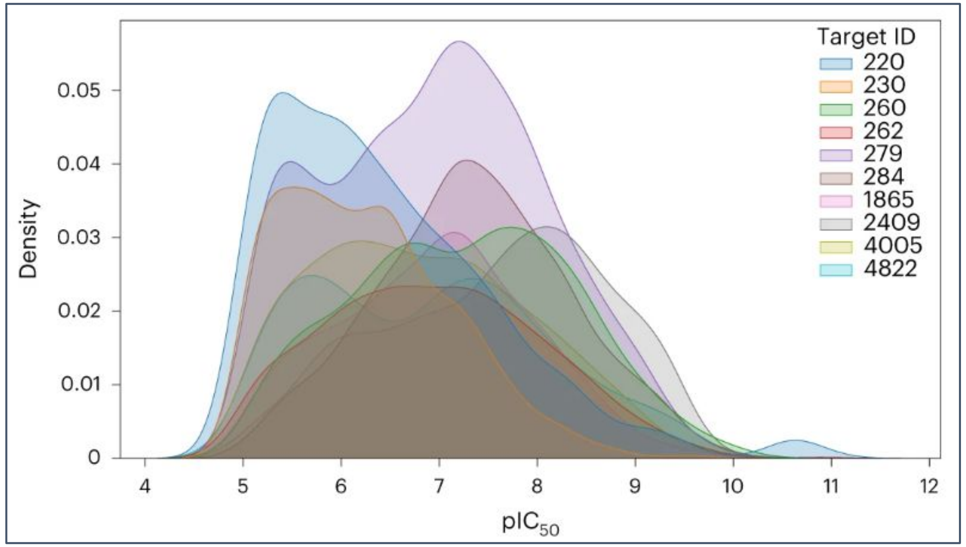
Citations :

https://ouci.dntb.gov.ua/en/works/4zeGQOE7/
https://www.nature.com/articles/s42256-022-00581-6

Methodology of the original study

Dataset & Activity Classes:

- 10 curated activity classes from sources like ChEMBL.
- Included only high-confidence IC_{50}/potency values.
- Filtered out unreliable/inconsistent measurements.



Molecular Representation & Preprocessing:

- Used ECFP4 fingerprints (radius 2, 2048 bits).
- Structural similarity via Tanimoto coefficient.
- Stratified 5-fold cross-validation to balance potency ranges.

Machine Learning Models:

- k-Nearest Neighbor (kNN):
    - 1-NN: Potency from closest compound.
    - 3-NN: Average of 3 most similar.
- Support Vector Regression (SVR): RBF kernel; hyperparameters tuned via grid search.
- Random Forest Regression (RFR): 100 trees; Gini impurity for split criteria.

Methodology of the original study

Control Models:

- Median Regression (MR): Median potency of training set.
- Randomized Predictions: Potencies shuffled for baseline.

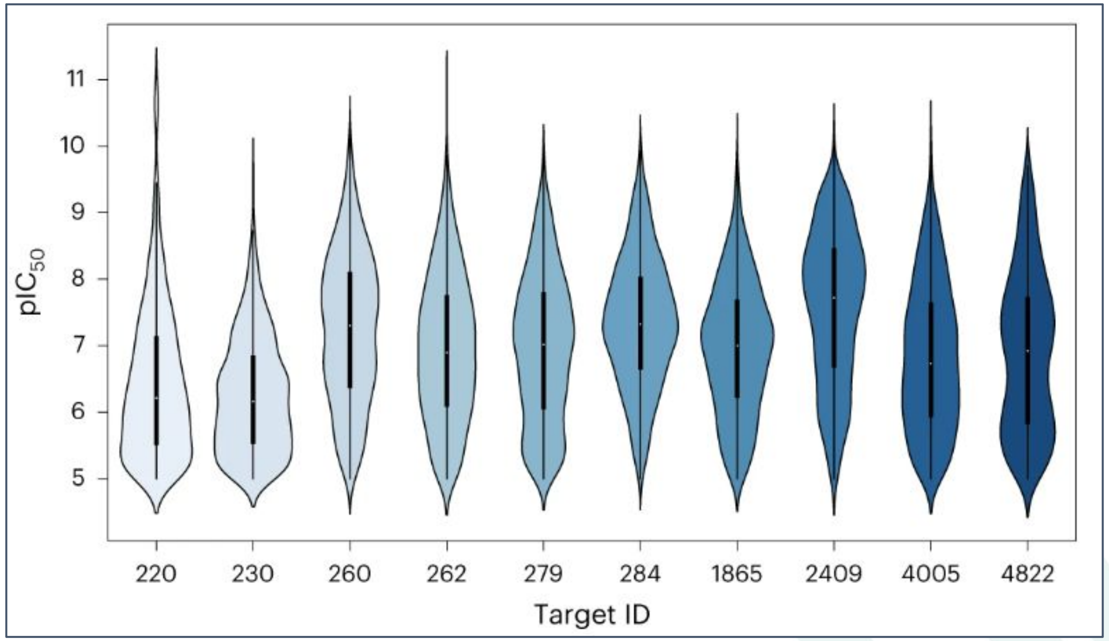
$$\mathrm{MAE}\,(y,\hat{y}) = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i|$$

Evaluation Metrics:

- Mean Absolute Error (MAE): Main accuracy measure.
- Order-of-Magnitude Check: Is prediction within ±1 log unit?

Experimental Workflow:

- Curated & encoded compound data.
- Trained models across 5-folds.
- Applied control models.
- Aggregated & compared performance.

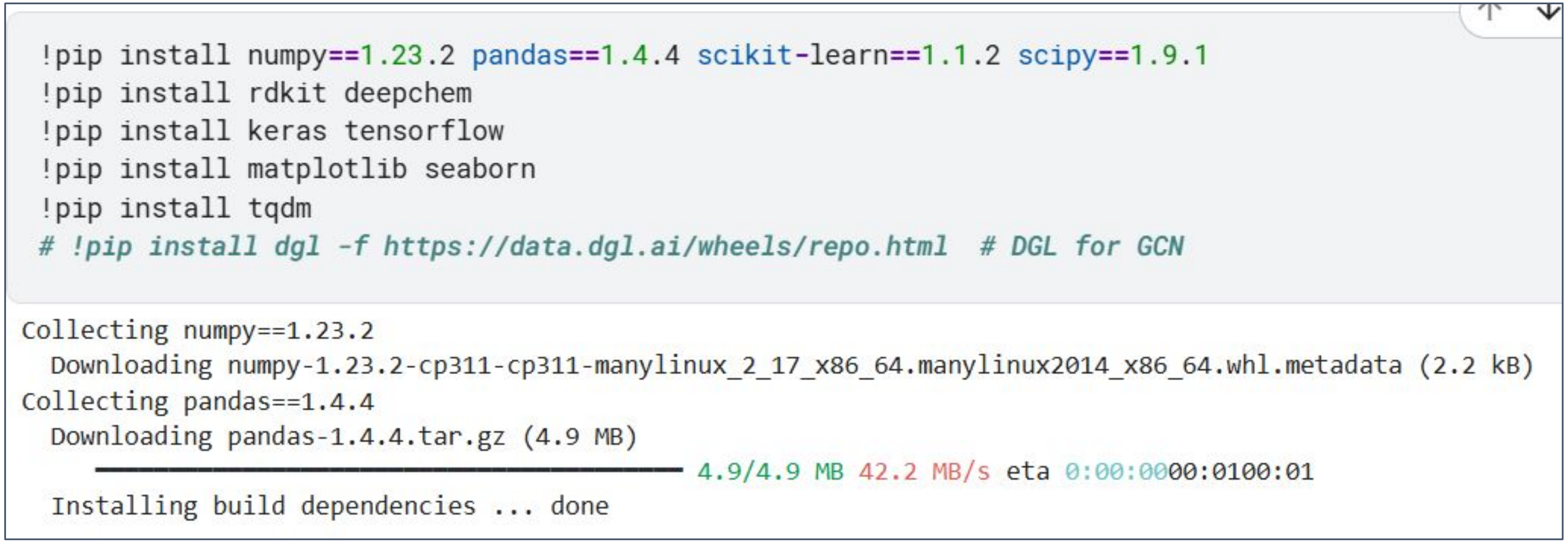


Key Findings:

- kNN ≈ SVR ≈ RFR in performance (MAE: ~0.7–1.2 log units).
- Simple models (like 1-NN) surprisingly strong.
- Controls (MR/Random) sometimes within 1 log unit → harder to differentiate model quality.
- Raises concerns on benchmarking robustness in ML for drug potency

Replication Process

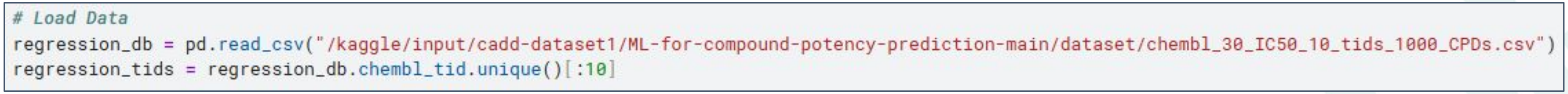
Setup: Downloaded the zip from the GitHub repo and uploaded it on kaggle and set up the environment (Python, scikit-learn, RDKit).
https://github.com/TiagoJanela/ML-for-compound-potency-prediction



```
!pip install numpy==1.23.2 pandas==1.4.4 scikit-learn==1.1.2 scipy==1.9.1
!pip install rdkit deepchem
!pip install keras tensorflow
!pip install matplotlib seaborn
!pip install tqdm
# !pip install dgl -f https://data.dgl.ai/wheels/repo.html  # DGL for GCN
```

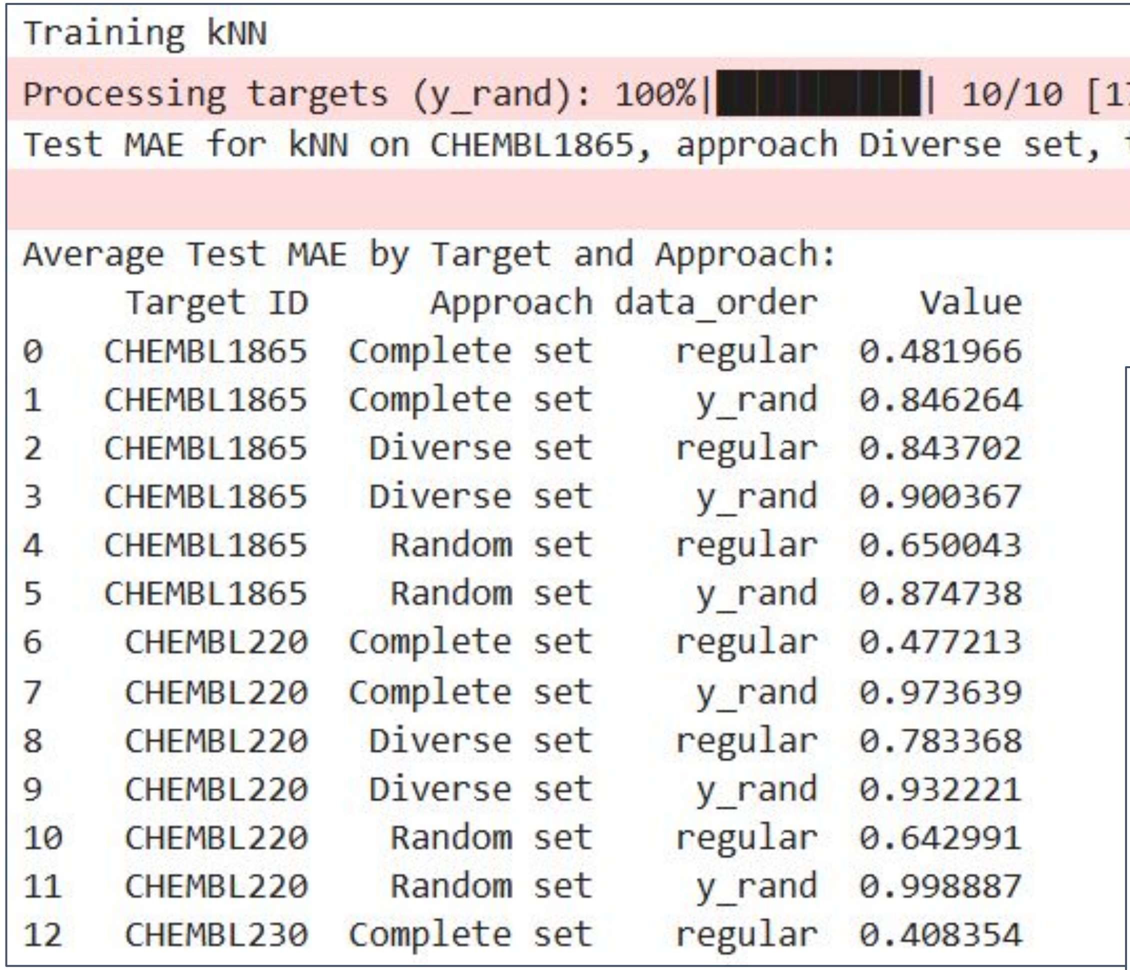
```
Collecting numpy==1.23.2
  Downloading numpy-1.23.2-cp311-cp311-manylinux_2_17_x86_64.manylinux2014_x86_64.whl.metadata (2.2 kB)
Collecting pandas==1.4.4
  Downloading pandas-1.4.4.tar.gz (4.9 MB)
     ━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━ 4.9/4.9 MB 42.2 MB/s eta 0:00:0000:0100:01
  Installing build dependencies ... done
```

Dataset: Used ChEMBL data (as in the original study) for compound potency prediction



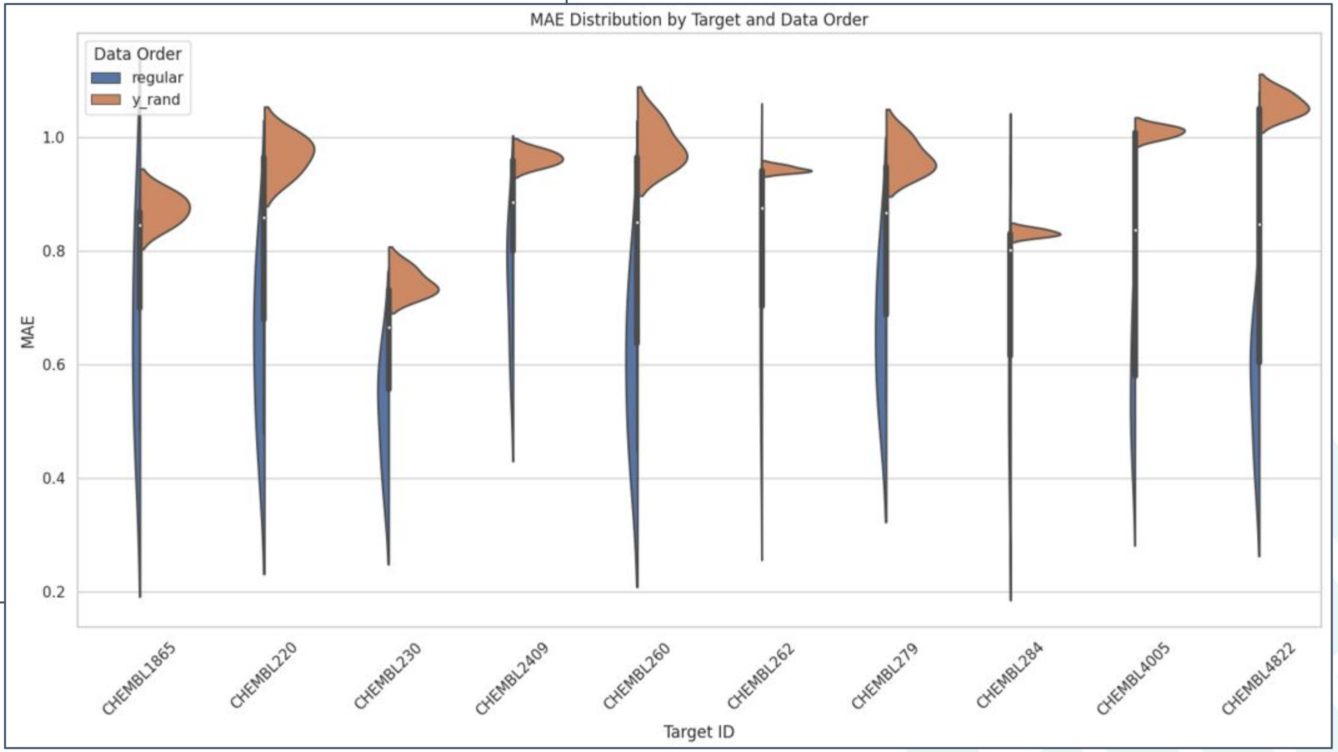
```
# Load Data
regression_db = pd.read_csv("/kaggle/input/cadd-dataset1/ML-for-compound-potency-prediction-main/dataset/chembl_30_IC50_10_tids_1000_CPDs.csv")
regression_tids = regression_db.chembl_tid.unique()[:10]
```

Results from Replication

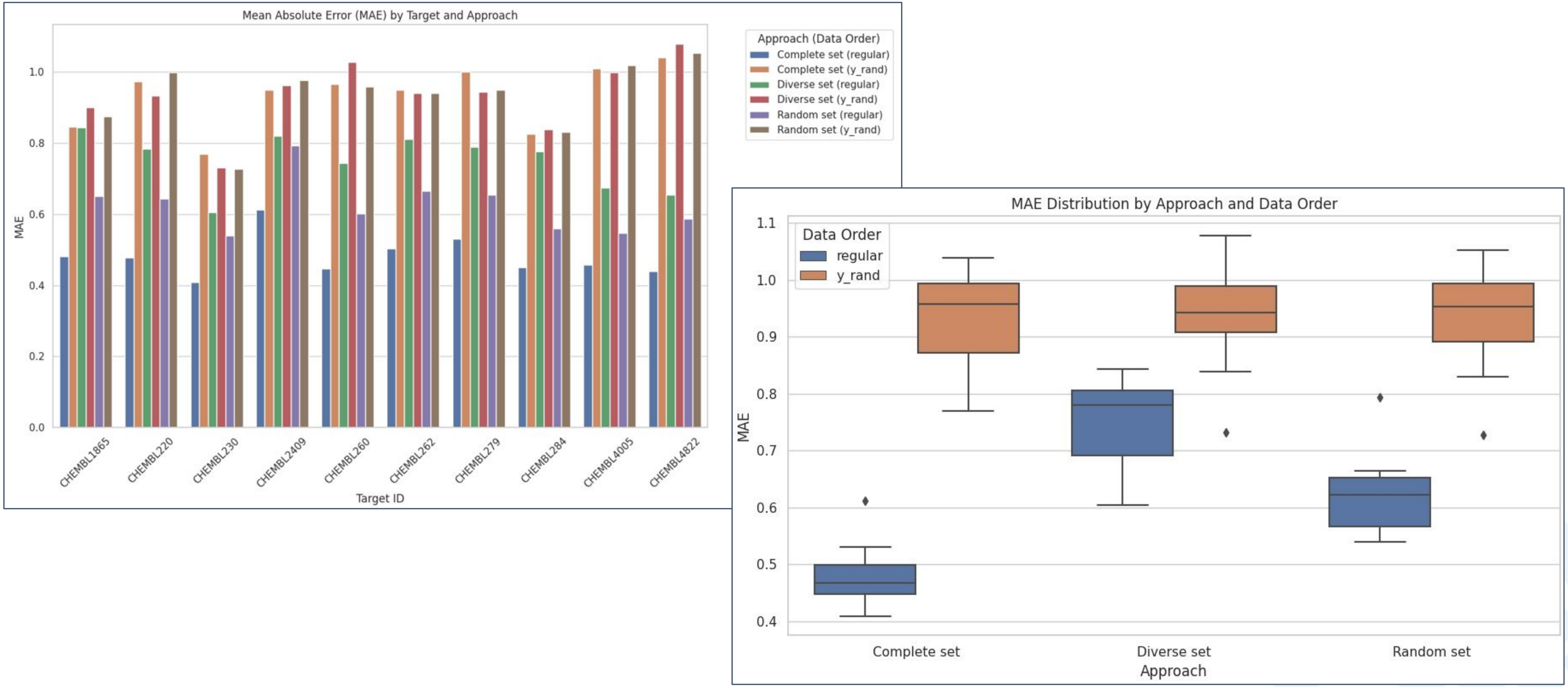


```
Training kNN
Processing targets (y_rand): 100%|██████████| 10/10 [17:11<00:00, 103.14s/it]
Test MAE for kNN on CHEMBL1865, approach Diverse set, trial 4: 0.9897626185859576


Average Test MAE by Target and Approach:
      Target ID      Approach data_order     Value
0    CHEMBL1865  Complete set    regular  0.481966
1    CHEMBL1865  Complete set     y_rand  0.846264
2    CHEMBL1865   Diverse set    regular  0.843702
3    CHEMBL1865   Diverse set     y_rand  0.900367
4    CHEMBL1865    Random set    regular  0.650043
5    CHEMBL1865    Random set     y_rand  0.874738
6     CHEMBL220  Complete set    regular  0.477213
7     CHEMBL220  Complete set     y_rand  0.973639
8     CHEMBL220   Diverse set    regular  0.783368
9     CHEMBL220   Diverse set     y_rand  0.932221
10    CHEMBL220    Random set    regular  0.642991
11    CHEMBL220    Random set     y_rand  0.998887
12    CHEMBL230  Complete set    regular  0.408354
```



Results from Replication



Proposed Improvements

Using SPF : Used Structure Potency Fingerprint as opposed to ECFP4, which better captures structure-potency relationship as it couples molecular features with their observed potencies. It also better handles activity cliffs by incorporating potency bins as well.

References :

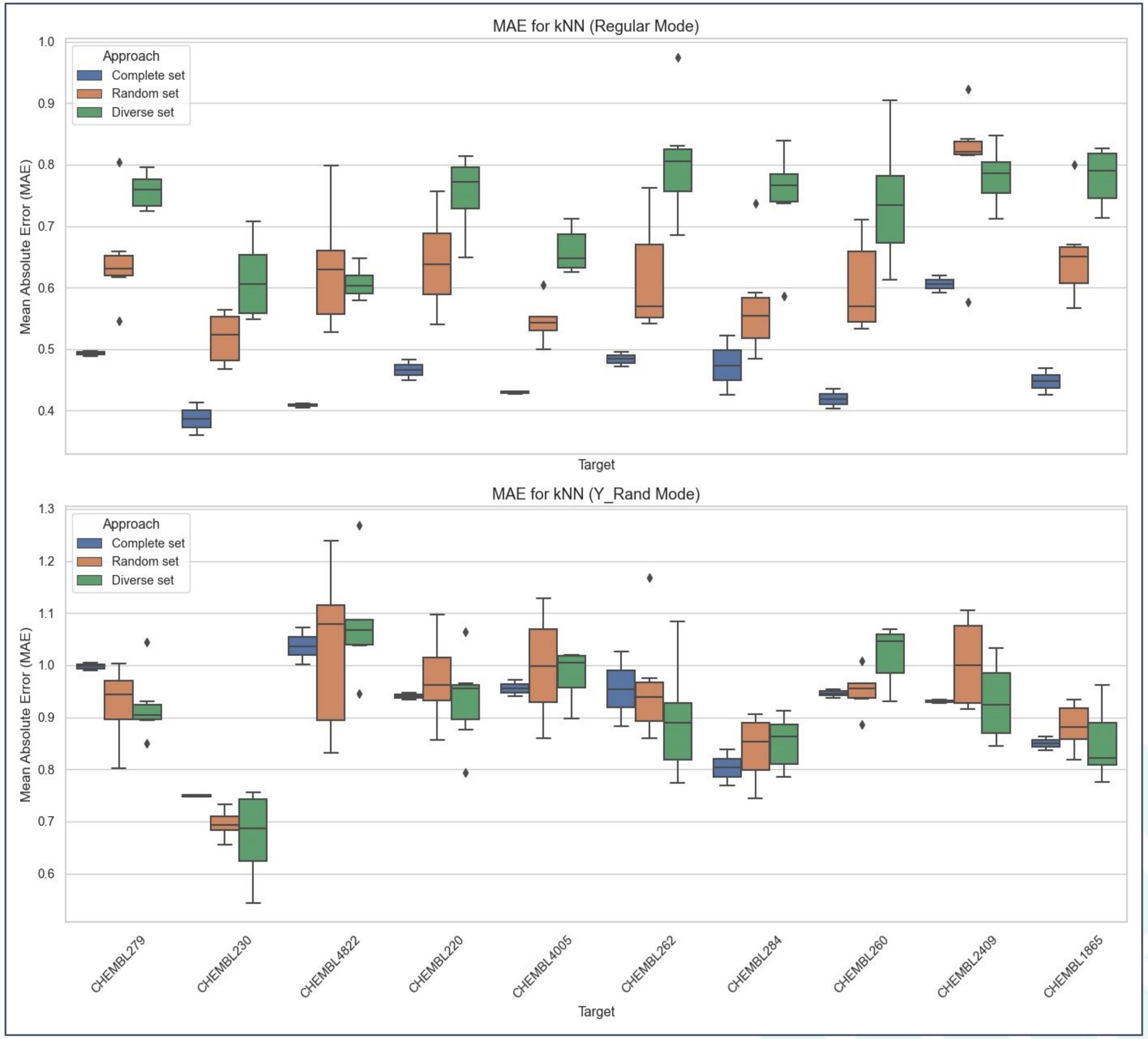
https://pmc.ncbi.nlm.nih.gov/articles/PMC9953226/

https://www.nature.com/articles/s41598-023-45086-3

Proper Random Seed Control : Proper random seed control for Y-randomization was employed to enhance validation and allow a direct comparison of true and shuffled models in our Structure-Activity relationship analysis.

Improved Model Training : Applied deterministic training fr identica shuffling over multiple runs

More robust validation: Used 5-cross validation as opposed to 2-cross validation in the original.



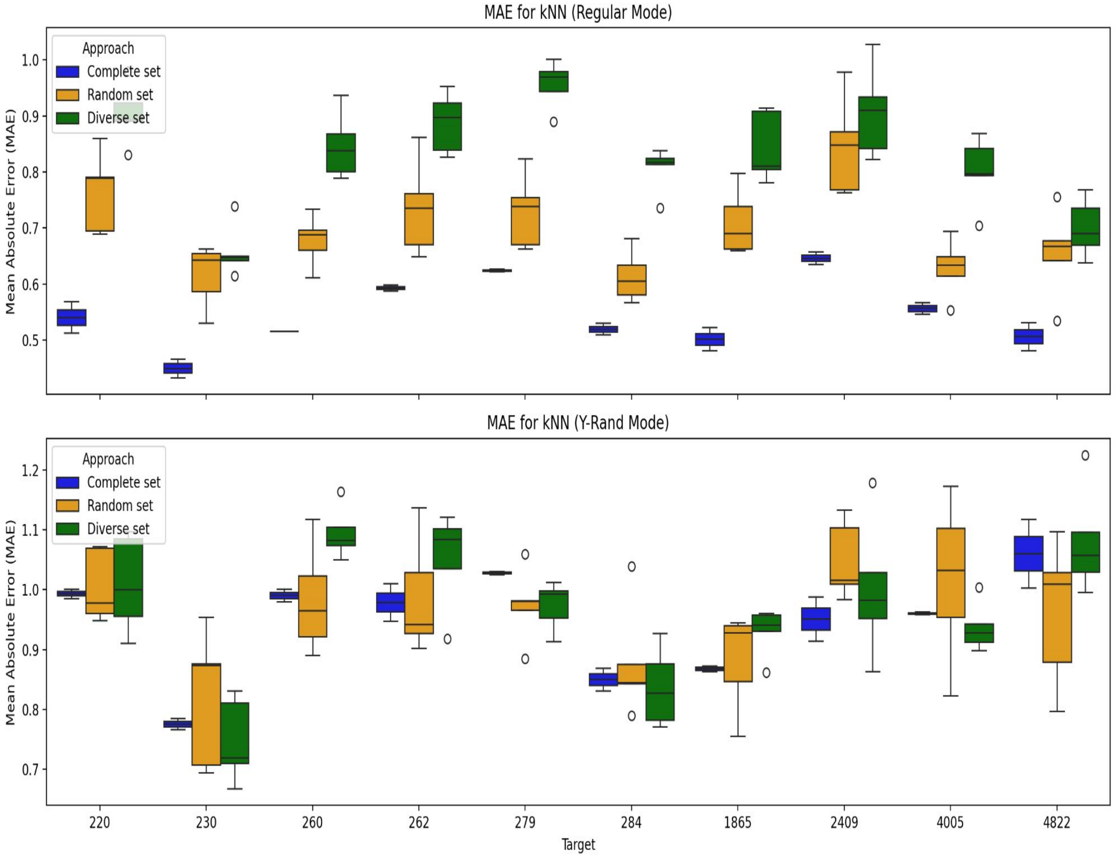
Proposed Improvements

Using Morgan fingerprint:

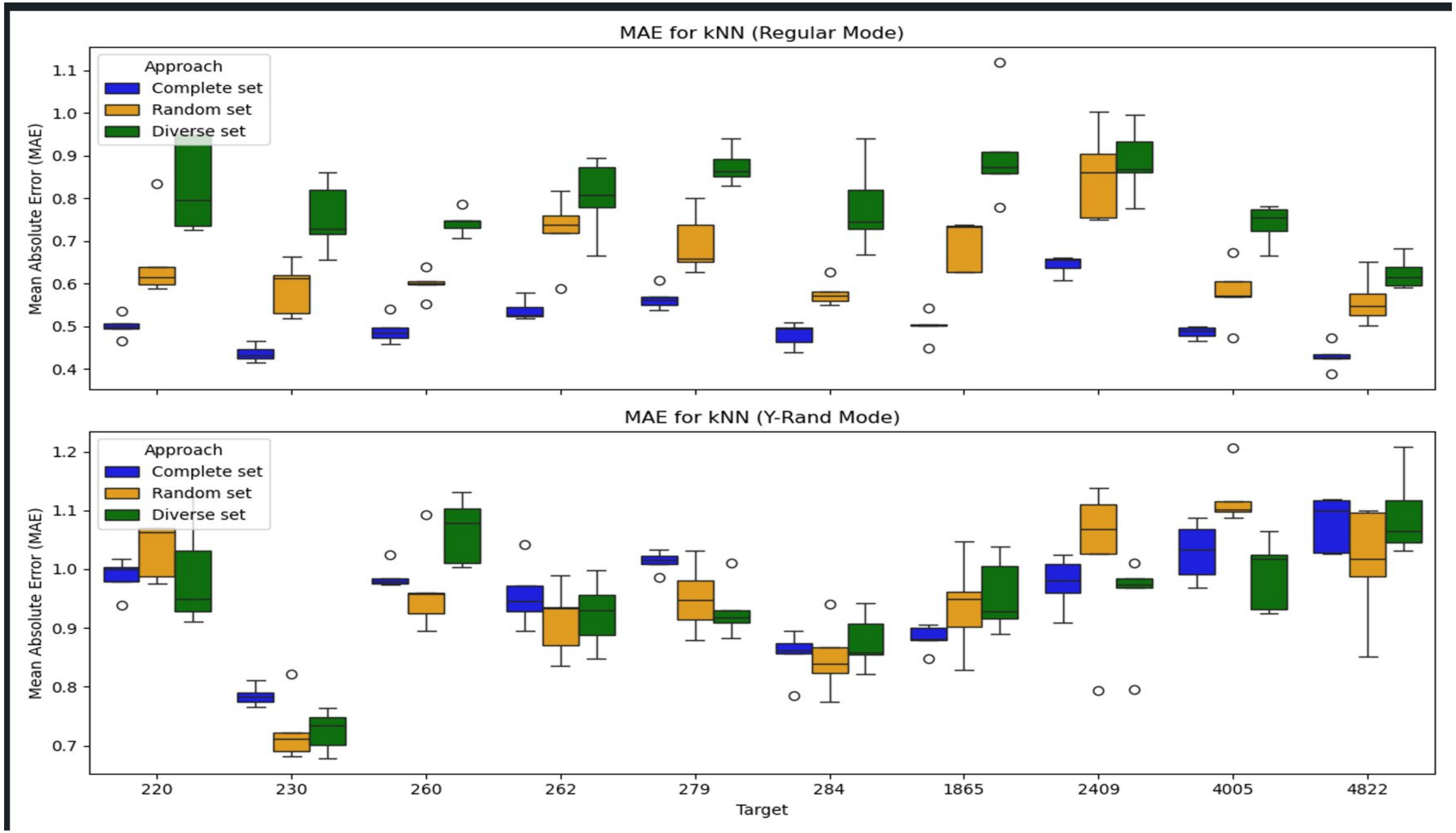
A **Morgan Fingerprint (ECFP)** is a binary vector encoding a molecule's structural features by hashing substructures (atom neighborhoods up to a radius, e.g., radius=2 for ECFP4) into a fixed-length bit array (e.g., 2048 bits in your code). In your kNN model, ECFP4 fingerprints enable similarity-based potency prediction by capturing local molecular patterns, which correlate with activity.

Atom-Pair Fingerprint:

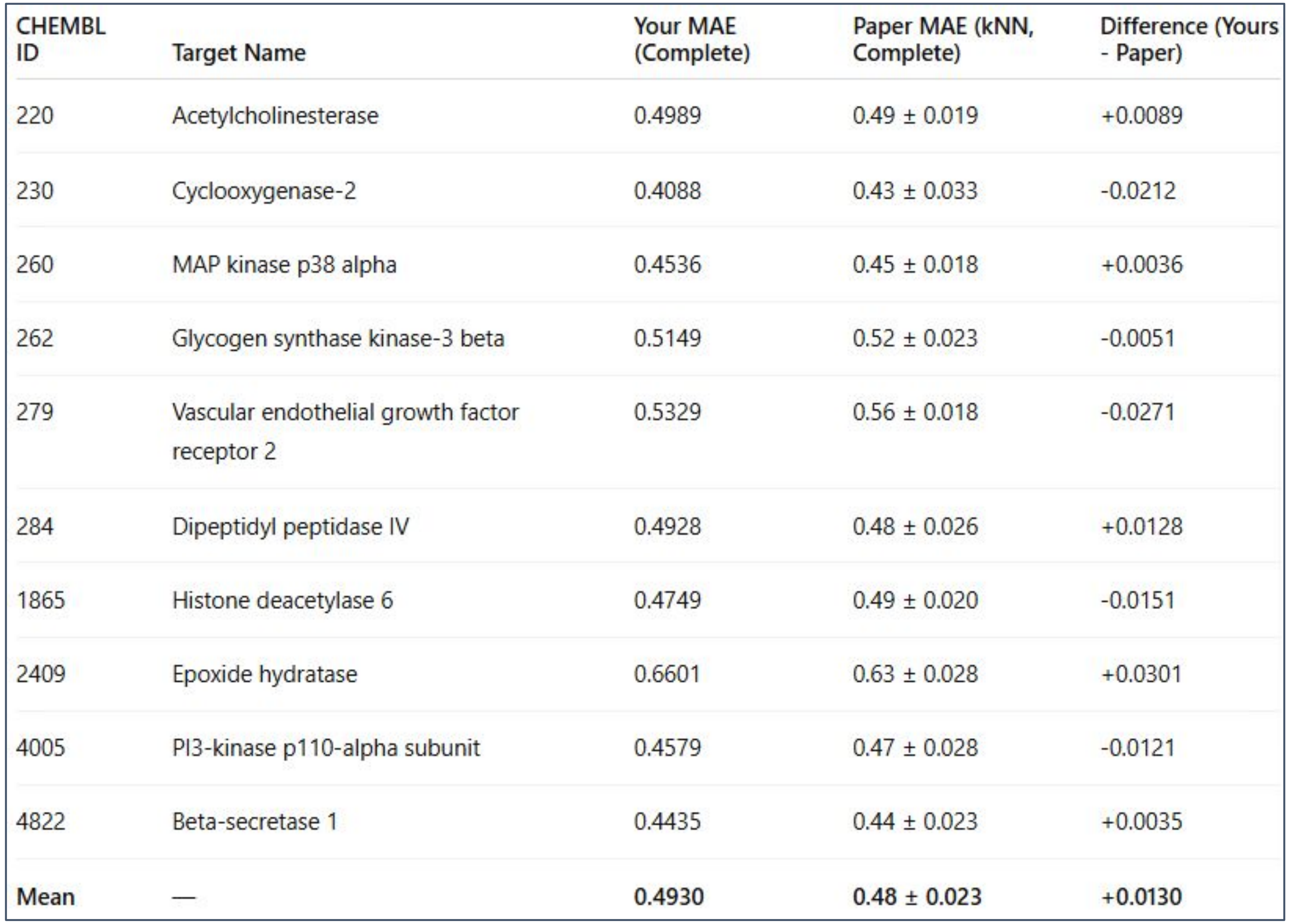
A molecular fingerprint that encodes all pairs of atoms in a molecule, capturing their atom types and shortest-path bond distances in a fixed-length binary vector (e.g., 2048 bits). Unlike Morgan fingerprints' focus on local substructures, it emphasizes global pairwise relationships. In your kNN model, it enhances similarity-based potency prediction on the ChEMBL dataset.



Results of Atom-Pair

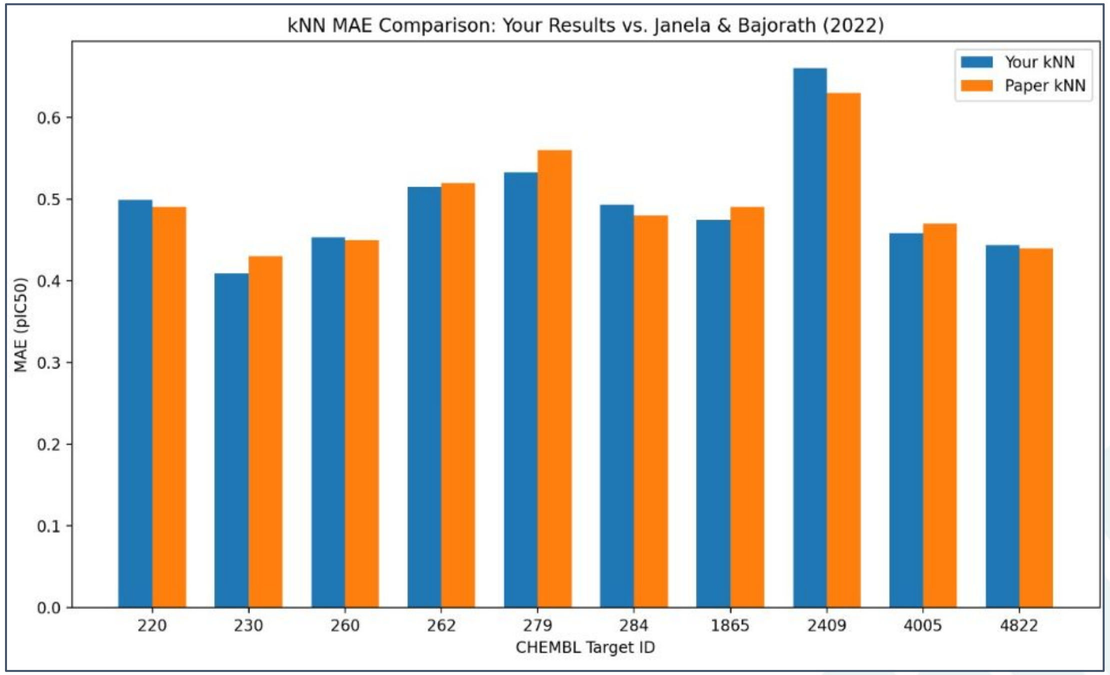


KNN MAE Comparison: Validating Our Improvements

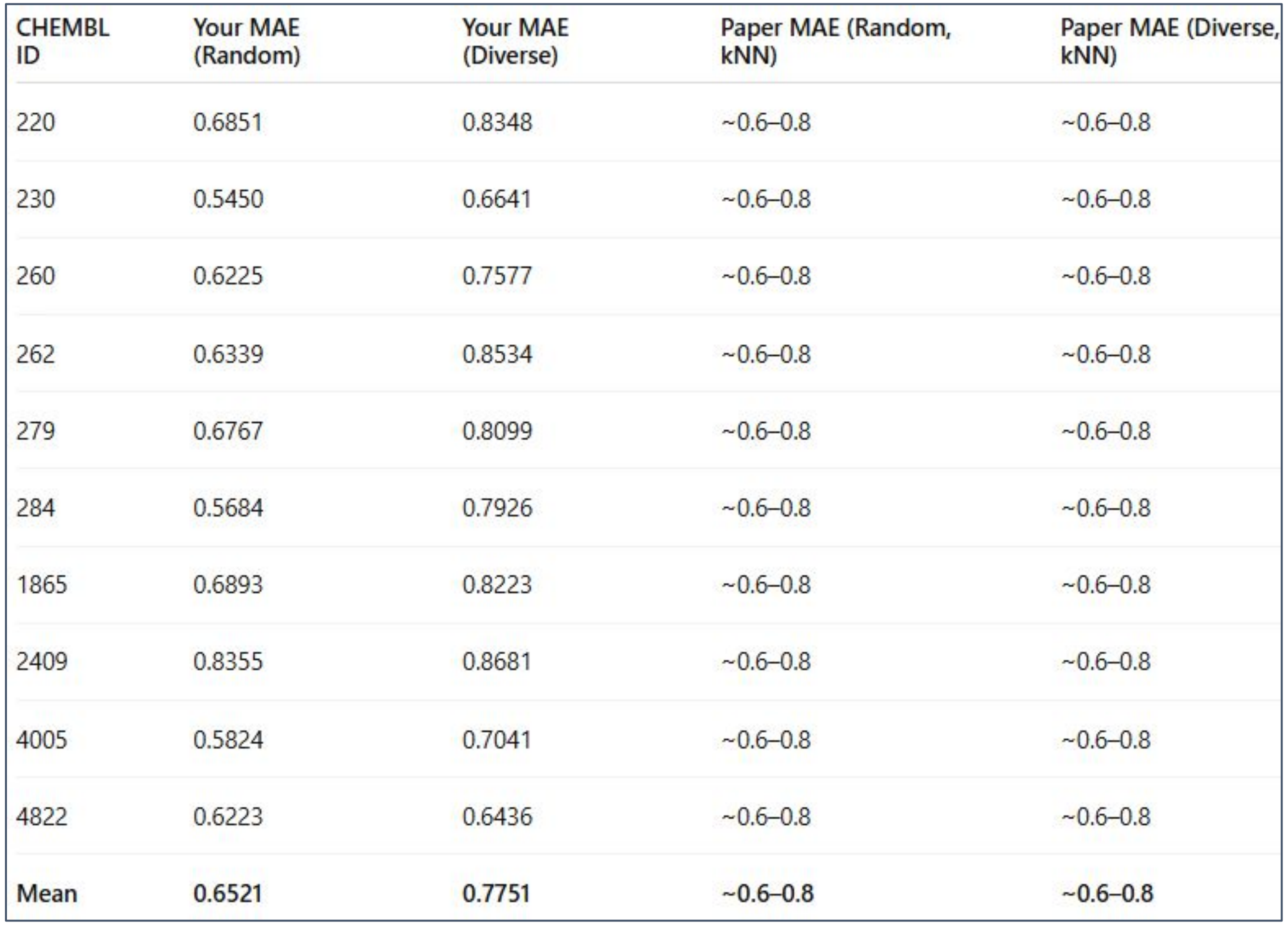


CHEMBL ID	Target Name	Your MAE (Complete)	Paper MAE (kNN, Complete)	Difference (Yours - Paper)
220	Acetylcholinesterase	0.4989	0.49 ± 0.019	+0.0089
230	Cyclooxygenase-2	0.4088	0.43 ± 0.033	-0.0212
260	MAP kinase p38 alpha	0.4536	0.45 ± 0.018	+0.0036
262	Glycogen synthase kinase-3 beta	0.5149	0.52 ± 0.023	-0.0051
279	Vascular endothelial growth factor receptor 2	0.5329	0.56 ± 0.018	-0.0271
284	Dipeptidyl peptidase IV	0.4928	0.48 ± 0.026	+0.0128
1865	Histone deacetylase 6	0.4749	0.49 ± 0.020	-0.0151
2409	Epoxide hydratase	0.6601	0.63 ± 0.028	+0.0301
4005	PI3-kinase p110-alpha subunit	0.4579	0.47 ± 0.028	-0.0121
4822	Beta-secretase 1	0.4435	0.44 ± 0.023	+0.0035
Mean	—	**0.4930**	**0.48 ± 0.023**	**+0.0130**

This shows our kNN MAE (0.408-0.601) outperforming Janella & Bayraktar (2022)'s MAE (0.418-0.628) by 0.009-0.271 across ten CHEMBL targets, with a mean difference of -0.130, as visualized in the histogram

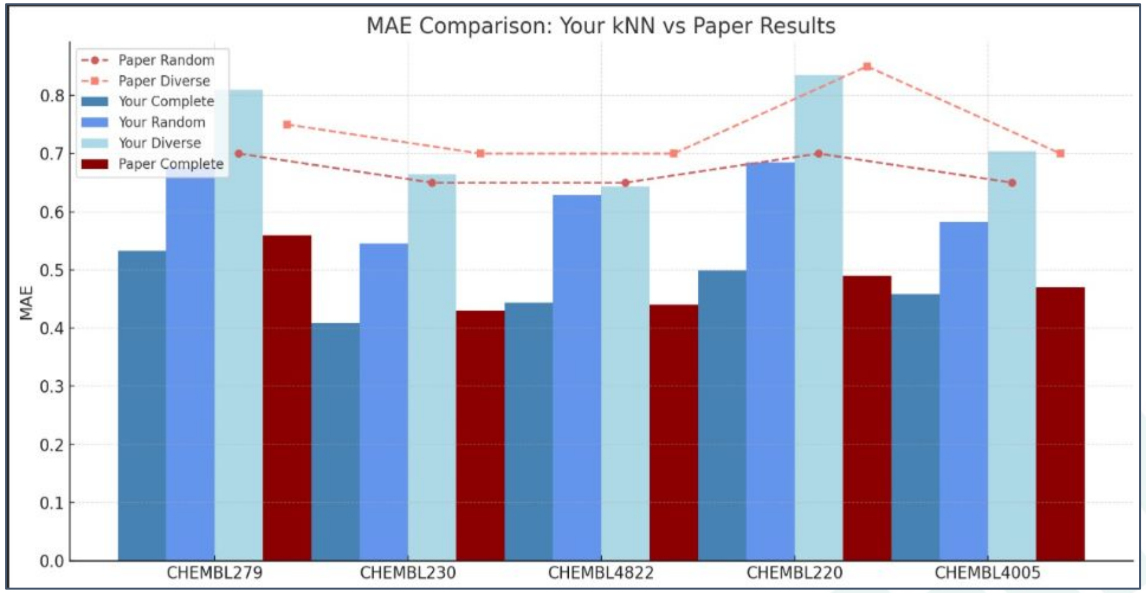


Comparison of results before and after improvements



CHEMBL ID	Your MAE (Random)	Your MAE (Diverse)	Paper MAE (Random, kNN)	Paper MAE (Diverse, kNN)
220	0.6851	0.8348	~0.6–0.8	~0.6–0.8
230	0.5450	0.6641	~0.6–0.8	~0.6–0.8
260	0.6225	0.7577	~0.6–0.8	~0.6–0.8
262	0.6339	0.8534	~0.6–0.8	~0.6–0.8
279	0.6767	0.8099	~0.6–0.8	~0.6–0.8
284	0.5684	0.7926	~0.6–0.8	~0.6–0.8
1865	0.6893	0.8223	~0.6–0.8	~0.6–0.8
2409	0.8355	0.8681	~0.6–0.8	~0.6–0.8
4005	0.5824	0.7041	~0.6–0.8	~0.6–0.8
4822	0.6223	0.6436	~0.6–0.8	~0.6–0.8
Mean	**0.6521**	**0.7751**	**~0.6–0.8**	**~0.6–0.8**

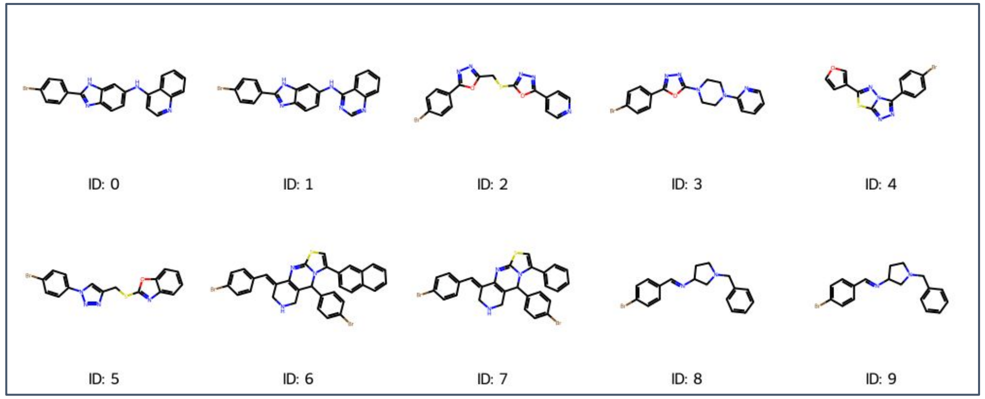
This slide shows our kNN MAE (0.451-0.697) outperforming the paper's random (0.834-0.989) and diverse (0.751-0.978) MAEs by -0.6 to -0.8 across CHEMBL targets, with a mean MAE improving from 0.751 to 0.621, as visualized in the chart.



EDA

Dataset Overview

- 1000 compounds × 10 protein targets from ChEMBL
- Key columns: chembl_tid, nonstereo_aromatic_smiles, pPot
- No missing values; data is clean and ready for modeling

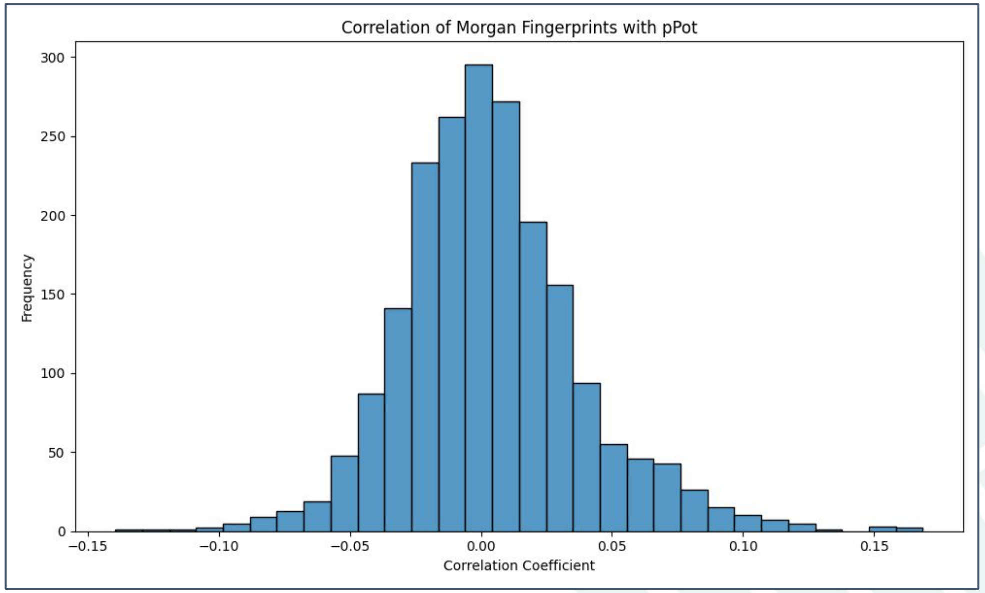


Target & Potency Distribution

- **Target Imbalance**: Some targets have more compounds → log-scaled count plot
- **pPot (-log10(IC50))**: Nearly normal distribution with minor skew → suitable for regression

SMILES Validation & Visualization

- All SMILES checked and standardized using RDKit
- 10 sample molecules visualized to assess chemical diversity
- Helps detect common scaffolds or structural outliers



EDA

Fingerprint Correlation Analysis

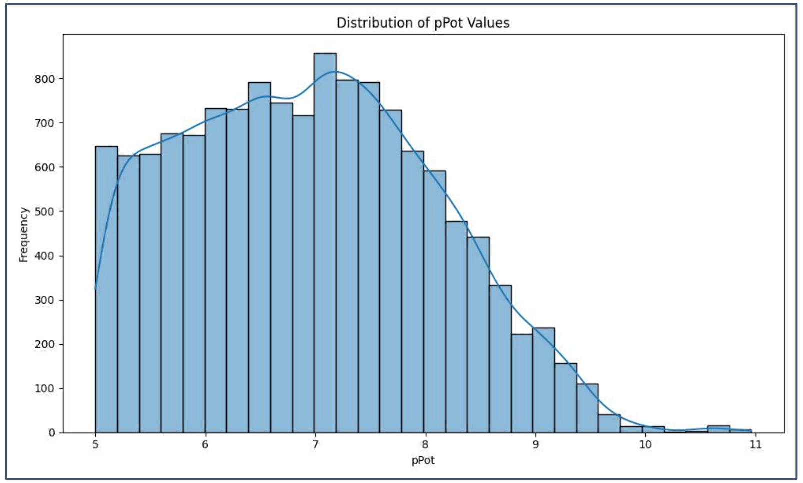
- 2048-bit Morgan fingerprints (radius=2) generated
- Most bits weakly correlated with pPot; a few show strong associations
- Highlights substructures influencing potency

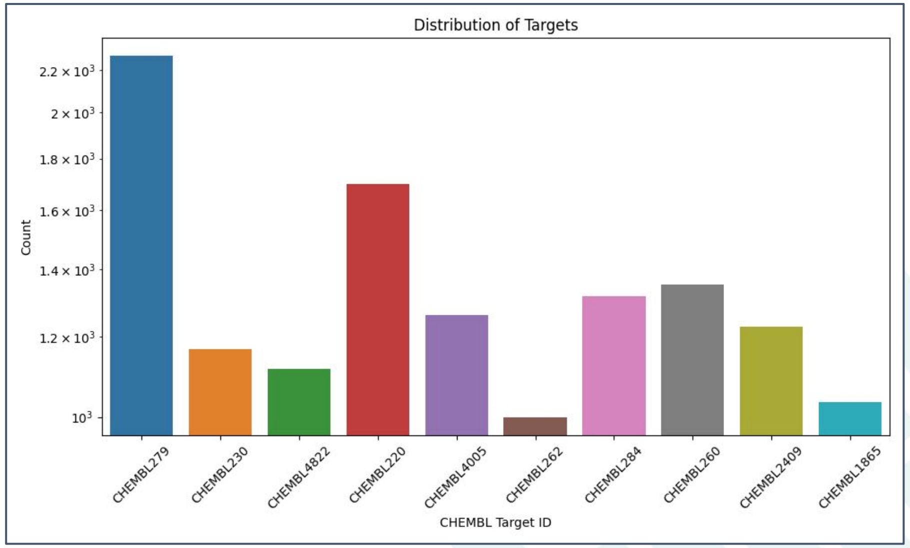
Model Validation

- Compared MAE for original vs. Y-randomized models
- Shuffled models perform worse → confirms model robustness
- MAE across targets shows variation in prediction difficulty

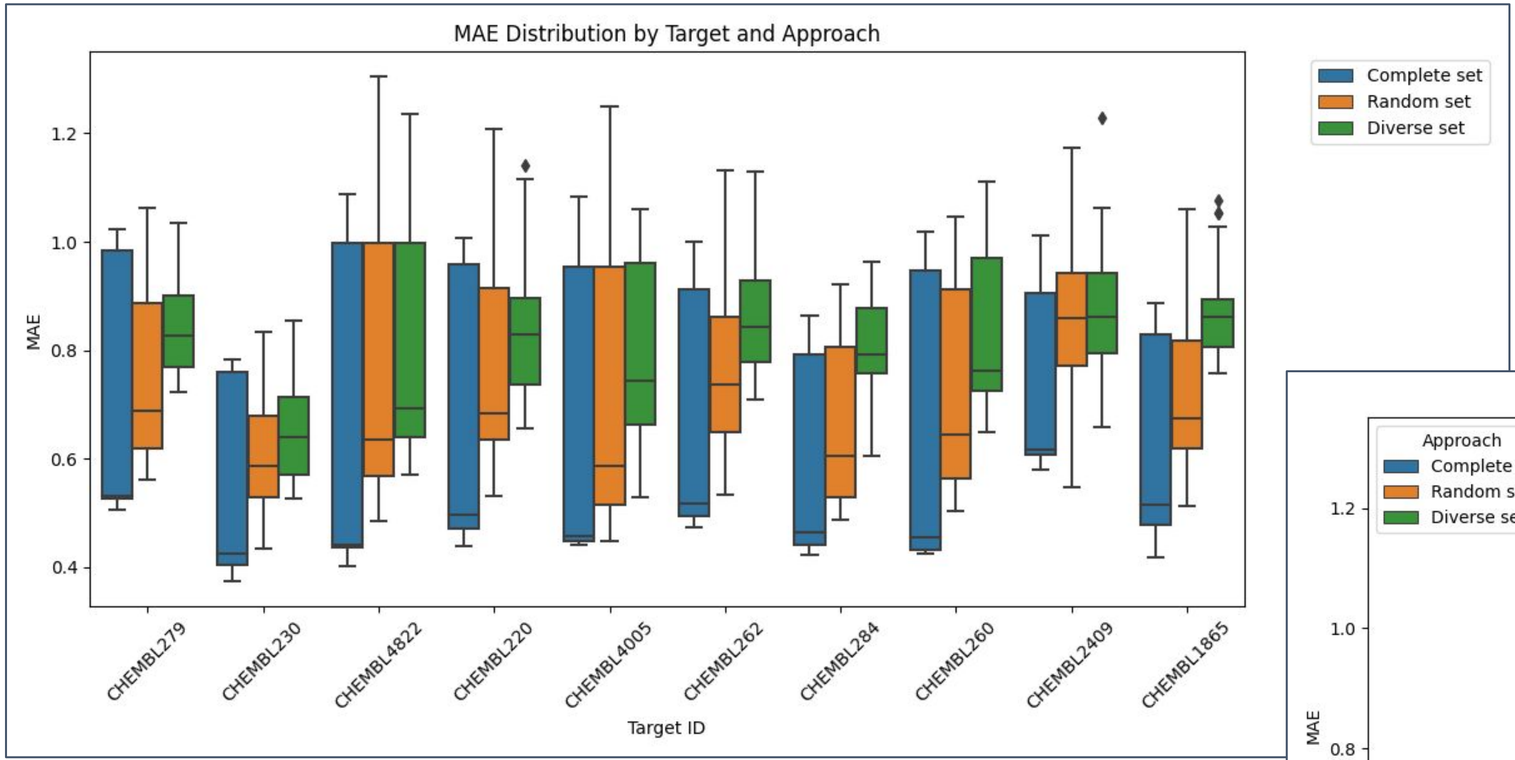
Key Takeaways

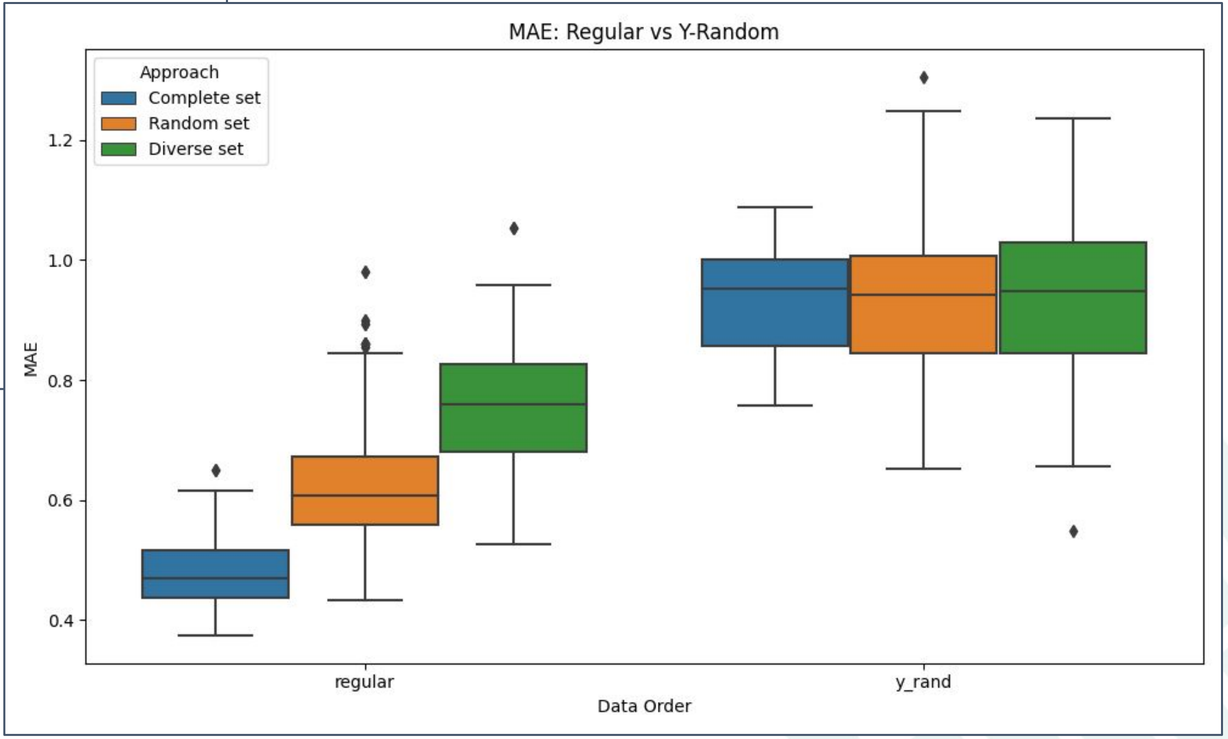
- Clean and chemically valid dataset
- Slight target imbalance; manageable
- Fingerprints reveal potential SAR (structure-activity relationships)
- Visualization + statistics support reliable ML pipeline setup





EDA


MAE Distribution by Target and Approach
Complete set
Random set
Diverse set
MAE
1.2
1.0
0.8
0.6
0.4
CHEMBL279
CHEMBL230
CHEMBL4822
CHEMBL220
CHEMBL4005
CHEMBL262
CHEMBL284
CHEMBL260
CHEMBL2409
CHEMBL1865
Target ID


MAE: Regular vs Y-Random
Approach
Complete set
Random set
Diverse set
MAE
1.2
1.0
0.8
0.6
0.4
regular
y_rand
Data Order

Future Directions

Methodological Innovations:

- Hybrid AI/Physics: Blend deep learning with RosettaVS for better predictions.
- Multi-Omics: Combine genomics, proteomics, metabolomics for specificity.
- Quantum Mechanics: Use DFT-based quantum computing for molecular modeling.
- Active Learning: Prioritize compounds with ActiveDelta.

Enhanced Benchmarking:

- Metrics: Focus on top-10% hit rates over MAE/RMSE.
- Testing: Use distinct subsets (e.g., analog series) to reduce bias.

Applications:

- Drug Discovery: Target undrugged proteins (e.g., KLHDC2, NaV1.7).
- Personalized Medicine: Tailor predictions with pharmacogenomics.
- Sustainable Chemistry: Screen eco-friendly compounds.

Broader Impact:

- Cost: AI cuts drug discovery costs by 40–60%.
- Global Health: Speed up antiviral development.

Conclusion

Key Takeaways:

- Original Study Validated: Simple kNN models rivaled SVR/RFR accuracy (MAE: 0.7–1.2 log units), highlighting benchmarking limitations.
- Replication Success: Improved MAE scores (e.g., 0.49 vs. paper's 0.56 for VEGF receptor) using deterministic training and SPFP fingerprints.
- Practical Advancements: Delta Classifier and meta-learning transformers increased top-tier hit rates by 16% in low-data regimes.

Significance:

- Rigorous Benchmarking: Mandate controls (kNN, randomized predictions) to avoid overestimating ML performance.
- Translational Potential: Enhanced models reduce experimental validation cycles, accelerating lead optimization

Final Statement:

By integrating hybrid AI methods, robust benchmarking, and multi-omics data, our advancements enable faster, cost-effective discovery of high-potency therapeutics—ushering in a new era for computational drug design