XXXII Symposium on Bioinformatics and Computer-Aided Drug Discovery (BCADD-2026)

online, 21st October 2026

DELTA: Database for Extended Ligand-Target Analyses - a New Balanced Resource for AI Applications in Drug Discovery

Giulio Vistoli
Department of Pharmaceutical Sciences
University of Milan, via Mangiagalli, 25,
I-20133, Milan

giulio.vistoli@unimi.it

UNIVERSITÀ DEGLI STUDI DI MILANO
DIPARTIMENTO DI SCIENZE FARMACEUTICHE

Introduction

Reliable predictions of ligand-target affinity remains one of the **major challenges** in computational drug discovery.

Machine and deep learning approaches are increasingly utilized to support compound activity predictions; however, their performance is fundamentally **constrained by the data accuracy and the availability of purposely collected datasets**

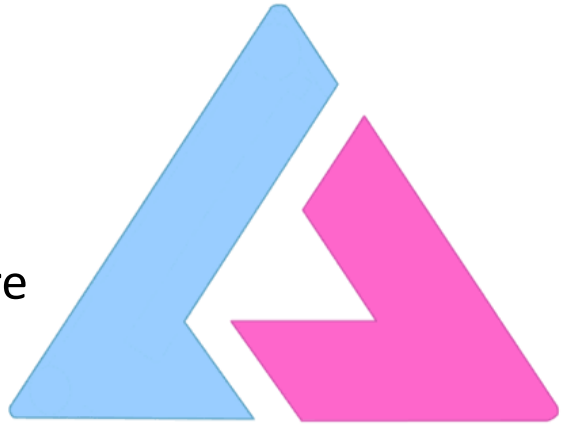
An ideal resource should include:

- ✓Only experimentally tested compounds
- ✓Balanced active and inactive compounds
- ✓Chemical space coverage
- ✓Wide range of targets
- ✓Ready-to-use 3D structures

UNIVERSITÀ DEGLI STUDI DI MILANO
DIPARTIMENTO DI SCIENZE FARMACEUTICHE

Database for Extended Ligand–Target Analyses (DELTA)

- Wide range of **therapeutically relevant protein targets**
- **Balanced** ligands' datasets
- **Binary** active/inactive classification
- **Curated & annotated** bioactivity data
- **Prepared** and **standardized** ligands' 3D structure
- **Ready-to-use** 3D protein structures



UNIVERSITÀ DEGLI STUDI DI MILANO
DIPARTIMENTO DI SCIENZE FARMACEUTICHE

Ligand Data Collection



Therapeutically relevant targets

Therapeutic Target Database (TTD)

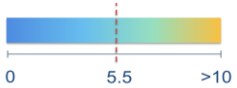
- Marketed (784)
- Clinical Trial (1227)
- Patented (180)



Experimentally tested compounds

Pubchem & ChEMBL

- Organic compounds
- MW < 1000 Da
- Activity = IC50, XC50, EC50, AC50, Ki, Kd or Potency



Activity threshold

pChEMBL=-log10(Activity)

≤5.5 → **Inactive**

>5.5 → **Active**



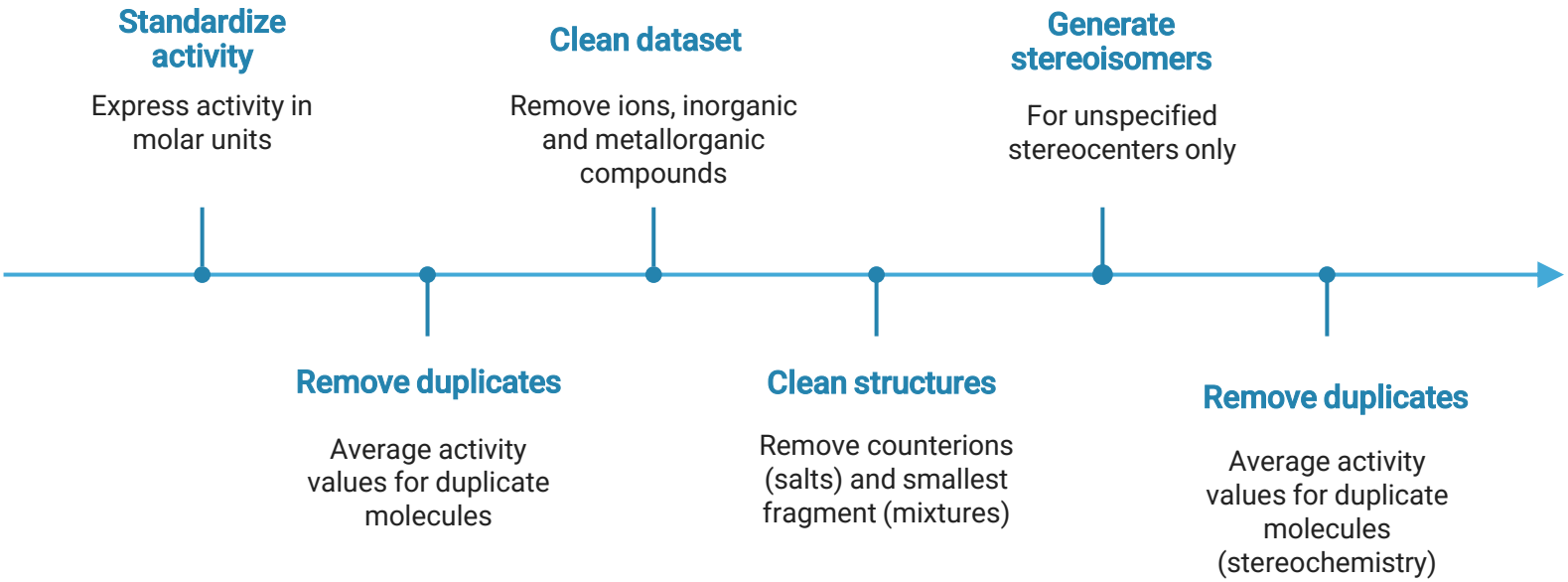
Criteria for target selection

Data availability

At least 100 active and 100 inactive molecules available

UNIVERSITÀ DEGLI STUDI DI MILANO
DIPARTIMENTO DI SCIENZE FARMACEUTICHE

Compounds Selection and Preparation



UNIVERSITÀ DEGLI STUDI DI MILANO
DIPARTIMENTO DI SCIENZE FARMACEUTICHE

Targets Data Collection and Preparation

Download structure from PDB

- Resolution
- Ligand availability
- Completeness

Model unavailable structures

- Homology model (coverage > 80%)
- Alphafold

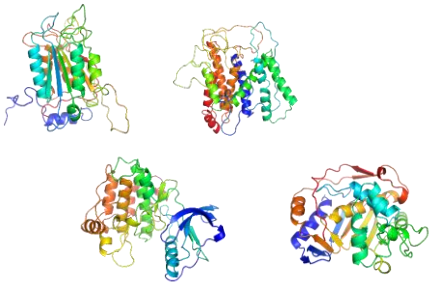
Refinement

- Model unresolved gaps and loops
- Add hydrogens

Minimization

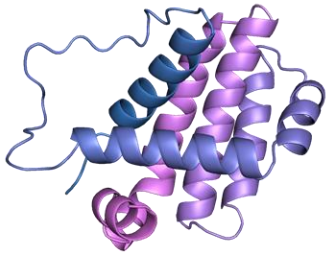
- Fixed backbone

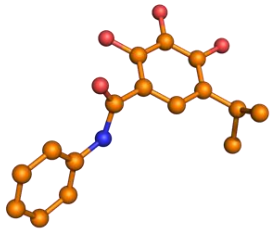
Final dataset



UNIVERSITÀ DEGLI STUDI DI MILANO
DIPARTIMENTO DI SCIENZE FARMACEUTICHE

DELTA in numbers







465

Targets

223,661

Ligands

307,438

Biological Data

Experimental structures = **415**

Homology models = **35**

Ab-initio models = **15**

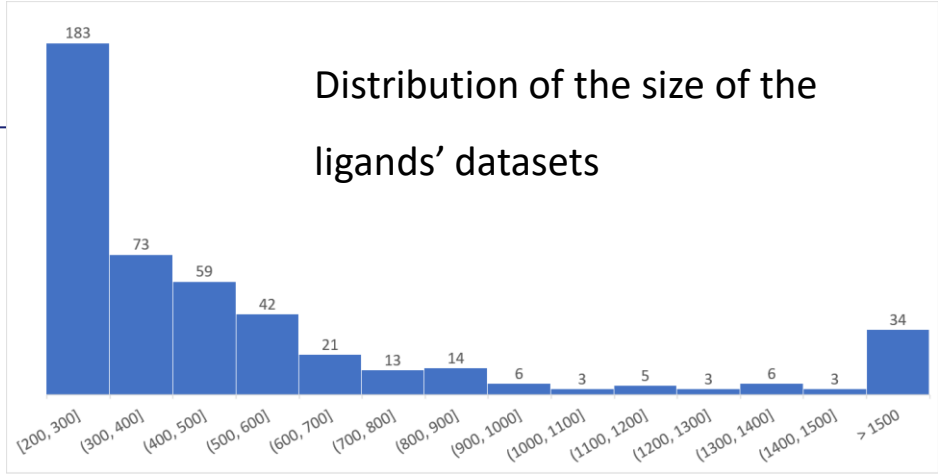
Targets for marketed drugs: **240**

Targets for drugs under clinical trials: **201**

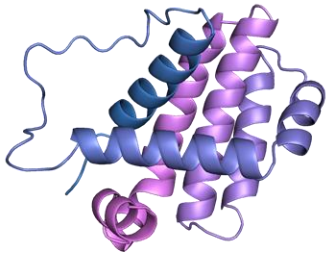
Targets with patented molecules: **24**

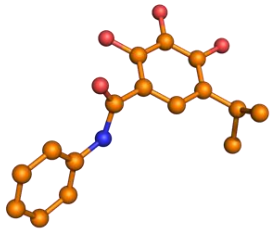
UNIVERSITÀ DEGLI STUDI DI MILANO
DIPARTIMENTO DI SCIENZE FARMACEUTICHE

DELTA in numbers



The size of datasets is not constant raging from 200 to 13,350 molecules







465

Targets

223,661

Ligands

307,438

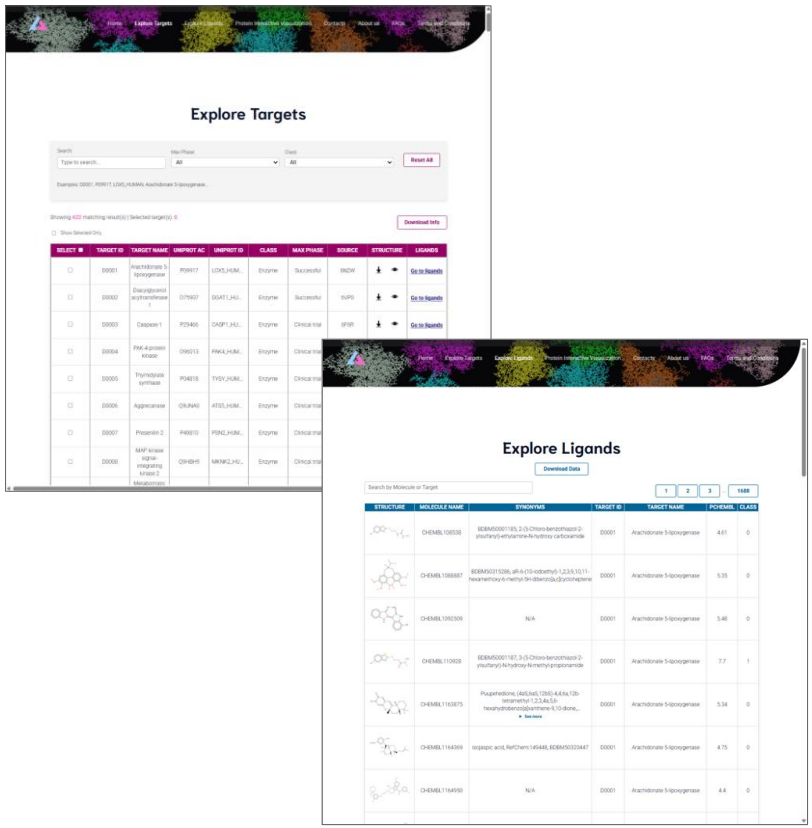
Biological Data

On average 660 molecules per dataset with an overlapping = 28%

UNIVERSITÀ DEGLI STUDI DI MILANO
DIPARTIMENTO DI SCIENZE FARMACEUTICHE

DELTA website https://delta.unimi.it/

The datasets, including **ligands** together with all **protein structures** and associated annotations, are freely available for consultation and download at:



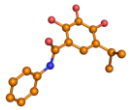
UNIVERSITÀ DEGLI STUDI DI MILANO
DIPARTIMENTO DI SCIENZE FARMACEUTICHE

Development of classification models

Input Datasets

465 balanced datasets with different size ranging from 200 to 13,350 molecules

Molecular Representation



Extended Connettivity Fingerprints (ECFP) 1024-bit, radius 2, chiral

119 **Physicochemical Descriptors** (RDKit) (e.g. MW, LogP, HBD/HBA…)

Data Partitioning



1. Random split (Stratified)

KFoldCV (K=10)

2. Scaffold split (Murcko)

GroupKFold (K=10)

3. Cluster split (Butina)

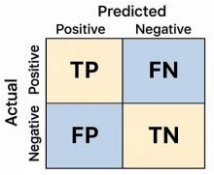
GroupKFold (K=10)

Classifiers & Optimization

Random Forest (RF)
Gradient Boosting (GB)
Support Vector Machine (SVM)
MultiLayer Perceptron (MLP)
K-Nearest Neighbors (KNN)

- Standard Implementation
- HyperParameter Tuning

Evaluation

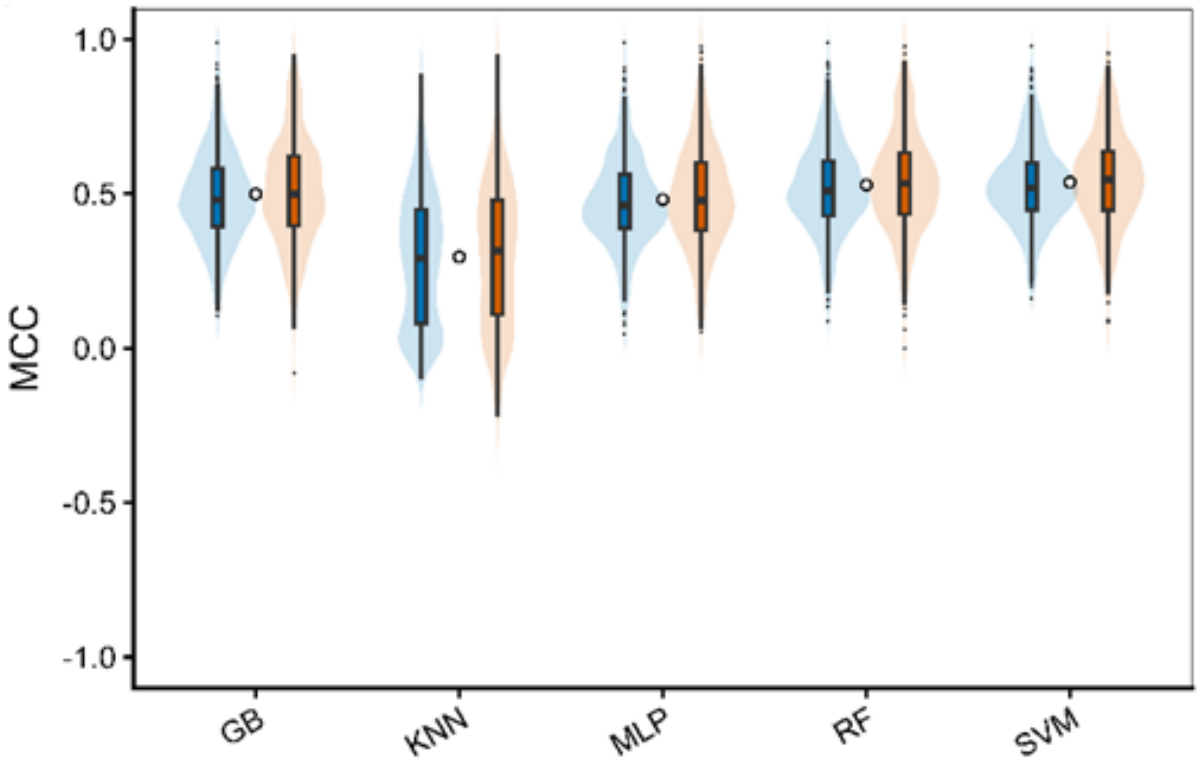


Used Metrics

Accuracy,
Precision,
Recall,
F1 score
MCC value

UNIVERSITÀ DEGLI STUDI DI MILANO
DIPARTIMENTO DI SCIENZE FARMACEUTICHE

Results

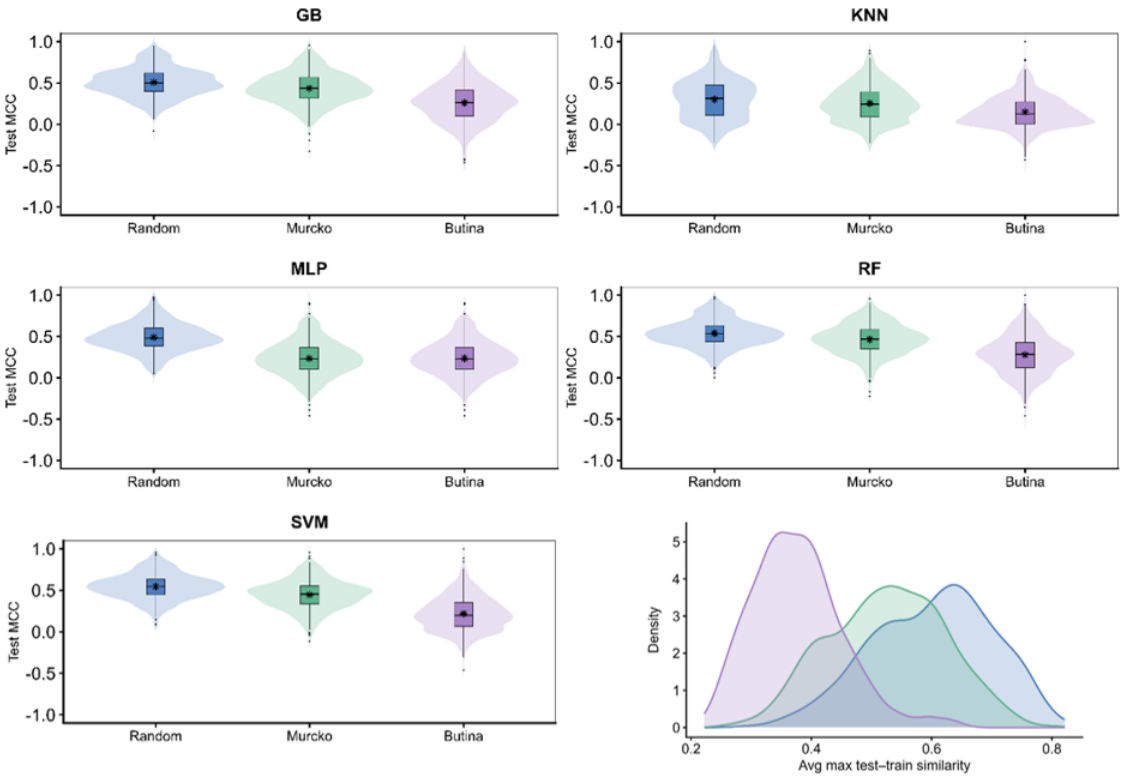


Blue = training set
Red = test set

- RF, GB, SVM & MLP → consistently strong performance
- KNN → lowest performance and highest variability
- Hyperparameter tuning → only marginal performance gains

UNIVERSITÀ DEGLI STUDI DI MILANO
DIPARTIMENTO DI SCIENZE FARMACEUTICHE

Performance Dependence on Splitting strategy

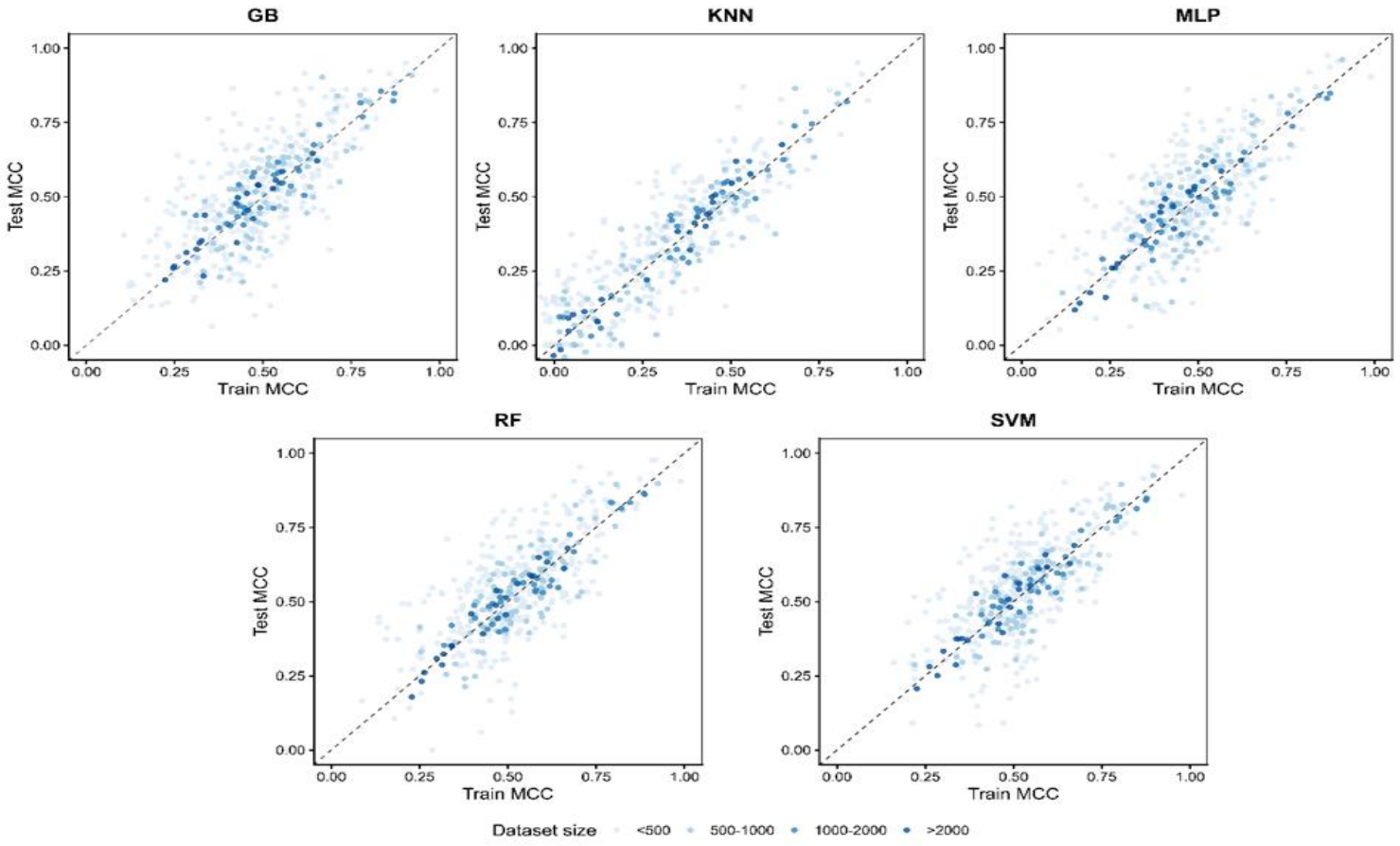


Split strategy determines model generalizability and depends on train test similarity:

- Random splitting: most optimistic performance
- Murcko scaffold splitting: reduced performance
- Butina cluster splitting: strongest distribution shift and largest performance loss.

UNIVERSITÀ DEGLI STUDI DI MILANO
DIPARTIMENTO DI SCIENZE FARMACEUTICHE

Dependence on Biological Function and Dataset Size

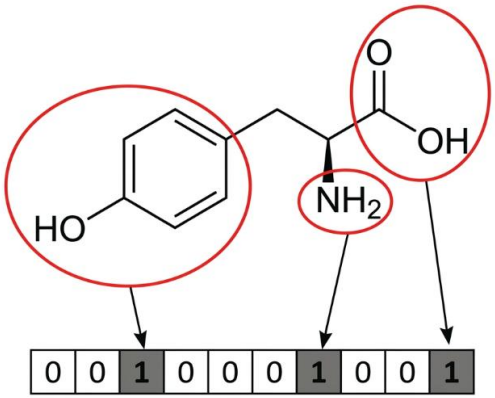


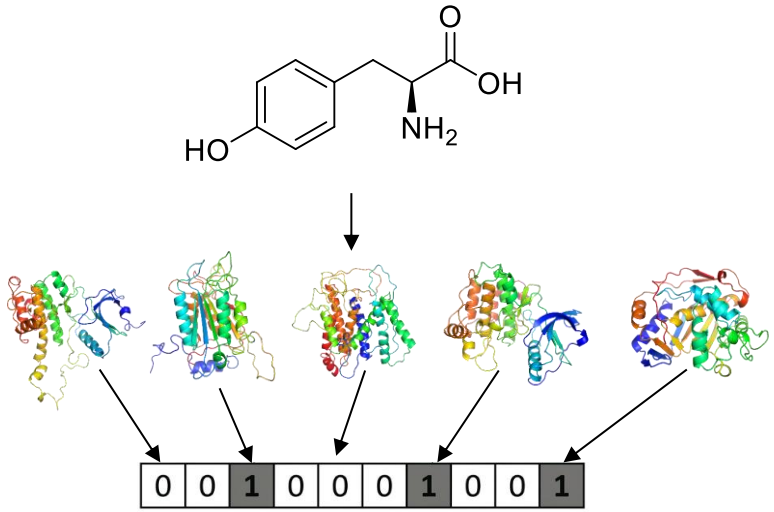
- Dataset size shows no correlation with model performance
- Positive correlation between train and test MCC values for larger datasets ascribable to greater stability

UNIVERSITÀ DEGLI STUDI DI MILANO
DIPARTIMENTO DI
SCIENZE FARMACEUTICHE

Development of a predicted affinity fingerprint

Molecular fingerprints are binary bit vectors that represent the chemical structure of the molecule. Each bit indicates the presence or absence of a specific chemical feature

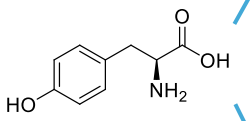




Affinity fingerprints are generated as vectors of a molecule's affinity profile across a panel of reference biological targets as predicted by classification models.

UNIVERSITÀ DEGLI STUDI DI MILANO
DIPARTIMENTO DI SCIENZE FARMACEUTICHE

DITO (DELTA Interpretable Target affinity Outcome)



Active/inactive classification



with already developed binary classification models for 465 DELTA targets

pChEMBL prediction



with newly developed regression models for 465 DELTA targets

DITO Binary version

D0001	D0002	D0003		D0463	D0464	D0465
0	1	0	...	0	0	0



DITO Continuous version

D0001	D0002	D0003		D0463	D0464	D0465
4.3	6.1	4.9	...	3.2	4.3	3.4

Both classification and regression models were developed by **Random Forest algorithm** (sklearn) using RDKit descriptors + ECFP (1024 bits, radius 2)

For regression models, average **r value** for test sets = 0.7 with average **RMSE** = 0.8

UNIVERSITÀ DEGLI STUDI DI MILANO
DIPARTIMENTO DI SCIENZE FARMACEUTICHE

DITO for cytotoxicity prediction

Cytotoxicity Data

HEK293 cell line
(AID 1345083)
93776 data points

Prepare data for machine learning

Data cleaning and class balancing
12296 data points

DITO calculation

Binary and continuous vector generation for the entire dataset

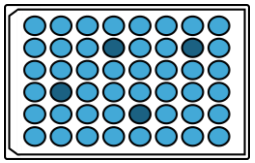
Expression filter

Remove features corresponding to proteins not expressed in HEK293

ML model training and validation

Evaluation of different algorithms in cross validation setting

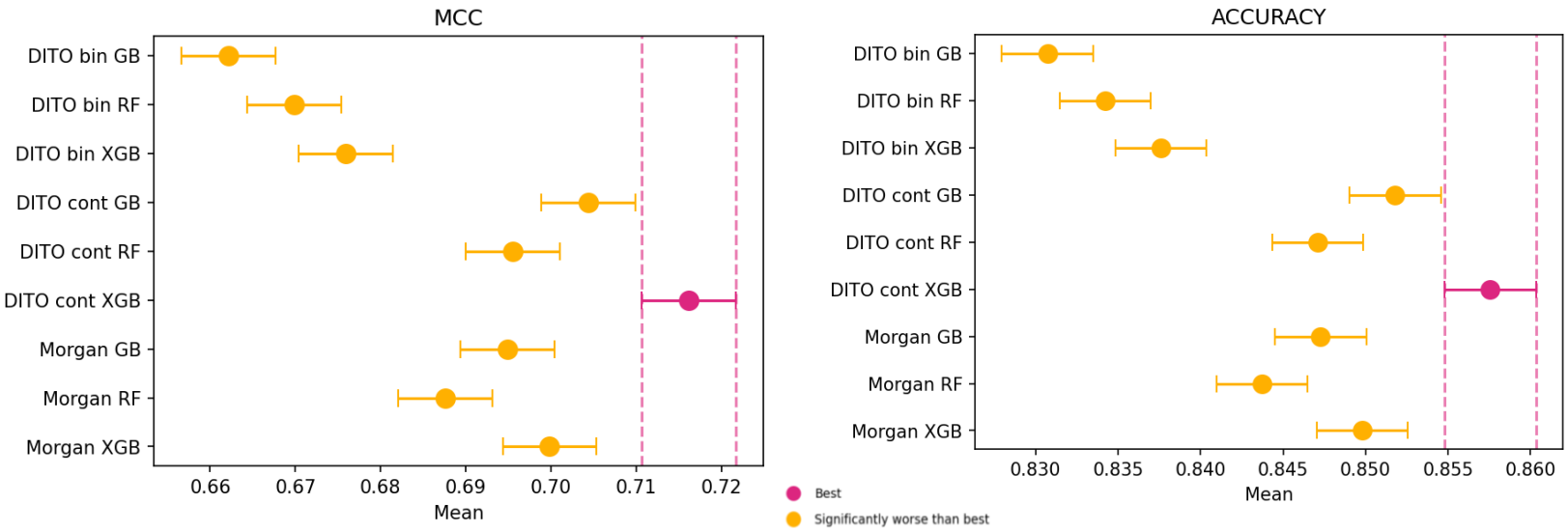
Final model



Cytotoxicity prediction on new molecules

UNIVERSITÀ DEGLI STUDI DI MILANO
DIPARTIMENTO DI SCIENZE FARMACEUTICHE

Results



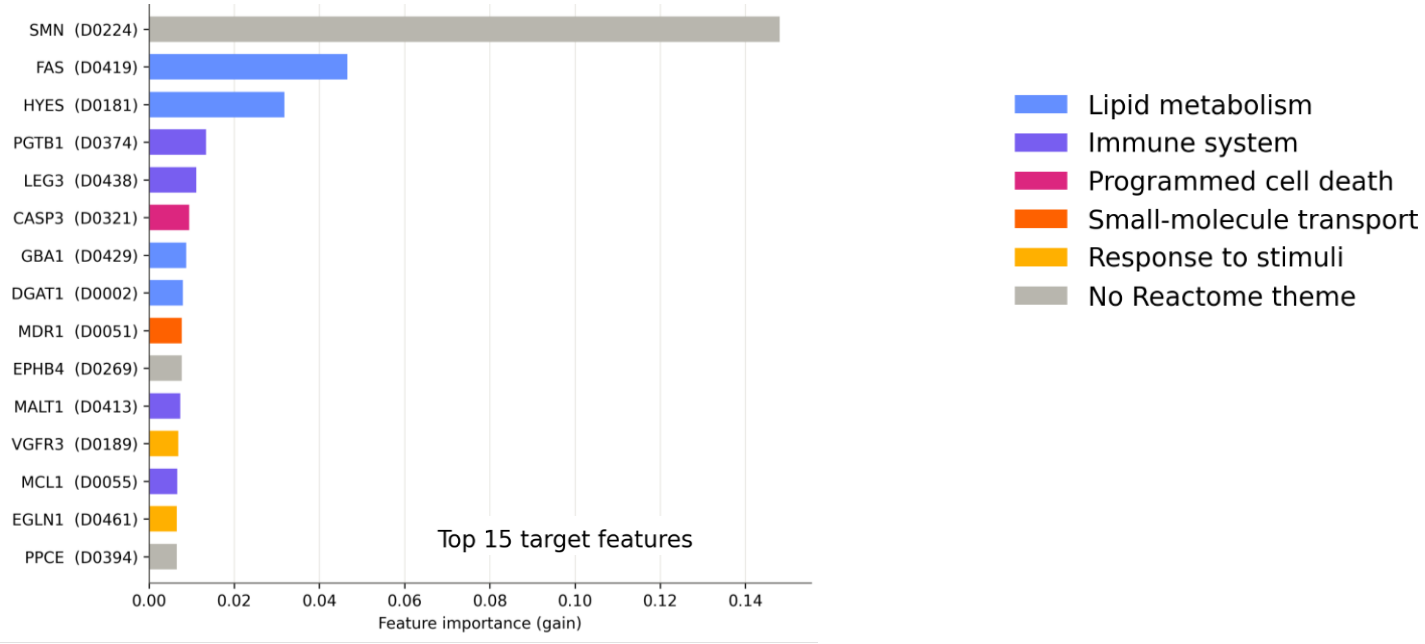
- Both continuous and binary DITOfp showed satisfying performance
- Continuous DITOfp performed better than the binary DITOfp and Morgan
- XGBoost proved to be the best algorithm in terms of MCC and accuracy

XGBoost was retrained on continuous DITOfp and then used to predict on a holdout test (20% of initial dataset)

Holdout test
MCC 0.73
Accuracy 0.86

UNIVERSITÀ DEGLI STUDI DI MILANO
DIPARTIMENTO DI SCIENZE FARMACEUTICHE

Model explainability



- Feature importance analysis can be used to highlight relevant targets in determining the studied cellular phenotype

- For cytotoxicity, top features fall into lipid metabolism and programmed cell death

UNIVERSITÀ DEGLI STUDI DI MILANO
DIPARTIMENTO DI SCIENZE FARMACEUTICHE

Conclusion

- DELTA provides a **curated, balanced and standardized ligand datasets** for more than 400 therapeutically-relevant human targets
- **Target-specific classifiers** were trained and validated across algorithms and splitting strategies in a reproducible workflow
- DELTA models were used to generate **DITO fingerprints**, a target space representation suitable for machine learning and directly interpretable
- **DITO fingerprints** can find meaningful application in **toxicity** predictions, **phenotyping** screening and cellular assays to reveal the involved **mechanisms** and targets' **networks**

UNIVERSITÀ DEGLI STUDI DI MILANO
DIPARTIMENTO DI SCIENZE FARMACEUTICHE



Acknowledgments



Drug Design Lab @ UNIMI
Prof. Alessandro Pedretti
Prof. Angelica Mazzolari
Dr. Arianna Pisati
Dr. Agnese Pozzi
Francesco Bilancia

Special thanks to
Dr. Andrea Zaliani

MUSA – Multilayered Urban Sustainability Action funded by the Ministry of University and Research under the National Recovery and Resilience Plan.

UNIVERSITÀ DEGLI STUDI DI MILANO
DIPARTIMENTO DI SCIENZE FARMACEUTICHE