Master's thesis

EMCIP

An Ensemble Model For Cdr1 Inhibitor Prediction

Student: The-Chuong Trinh

Supervisor: Prof. Ahcène Boumendjel

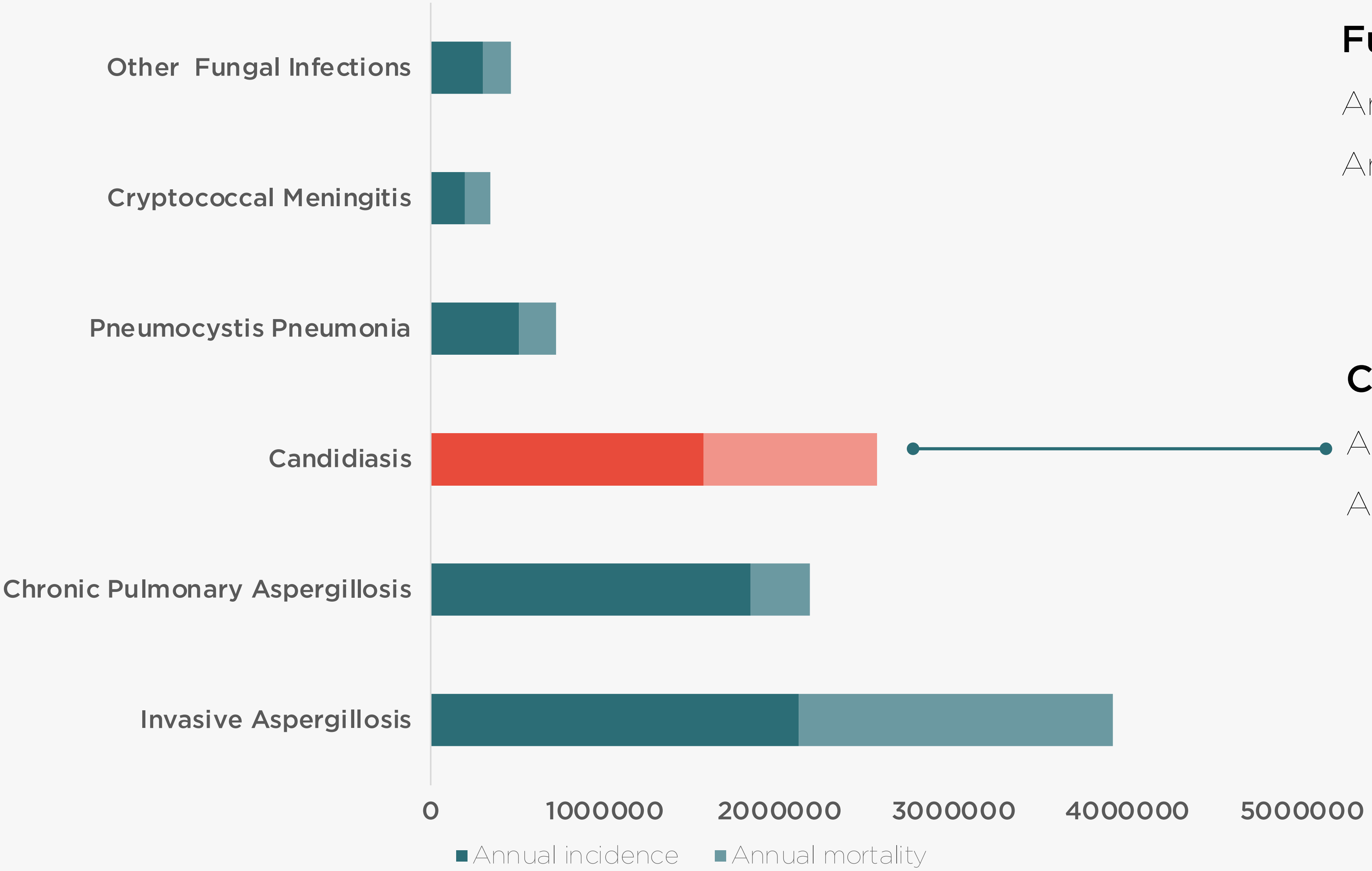
lrb Laboratoire Radiopharmaceutiques Bioclinique

UGA Université Grenoble Alpes

Inserm La science pour la santé From science to health

1. INTRODUCTION

FUNGAL INFECTION



Fungal infection

Annual incidence: 6.5 million cases

Annual mortality: 3.8 million cases

Candidiasis

Annual incidence: 1.6 million cases

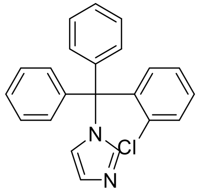
Annual mortality: 995,000 cases

David W Denning, *The Lancet Infectious Disease*, 2024

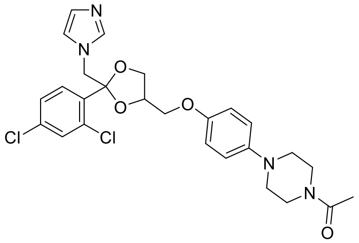
CDR1 INHIBITORS

Azoles

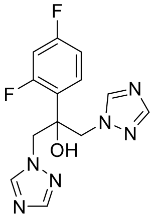
The most widely used for treating and preventing Candida infections



Clotrimazole

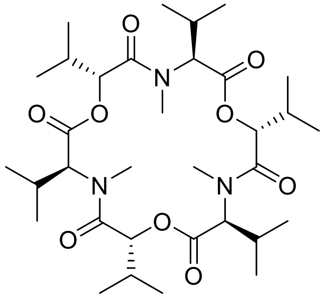


Ketoconazole

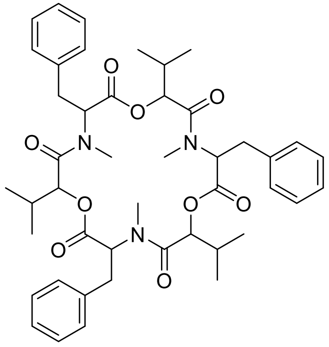


Fluconazole

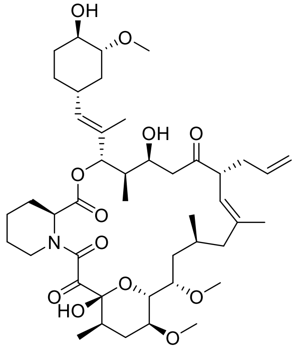
Cdr1 inhibitors



Enniatin B



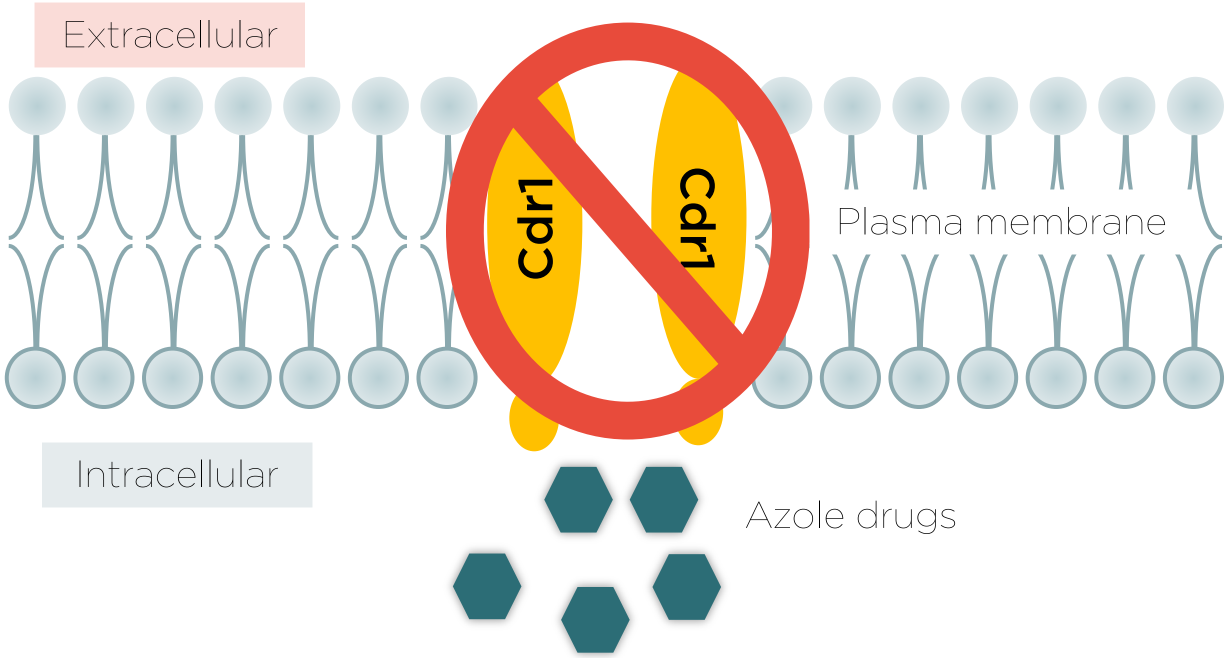
Beauvericin



Fk506

Cdr1 efflux pump

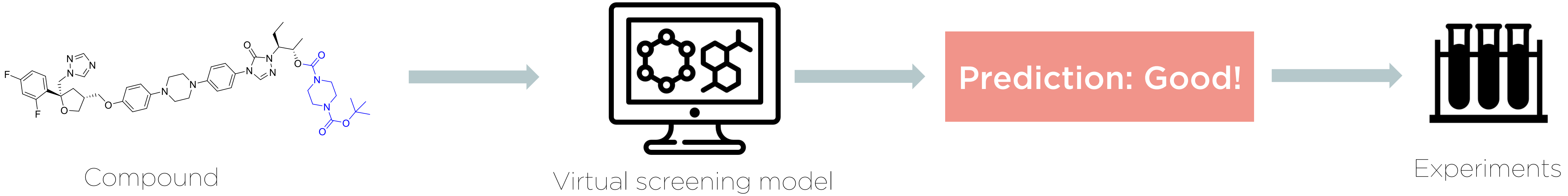
One key resistance mechanism of Candida



VIRTUAL SCREENING

Virtual screening

Assess whether a compound is a good drug using computation models (Walter et al., 1998)



Compound

Virtual screening model

Prediction: Good!

Experiments

Pros

- Much faster than experimental screening in wet labs
- Test 10^8 compounds within a day
- Much cheaper than experimental screening

WORK PACKAGES

1

DATA CURATION

Collect and curate relevant compounds for Cdr1 transporters.

2

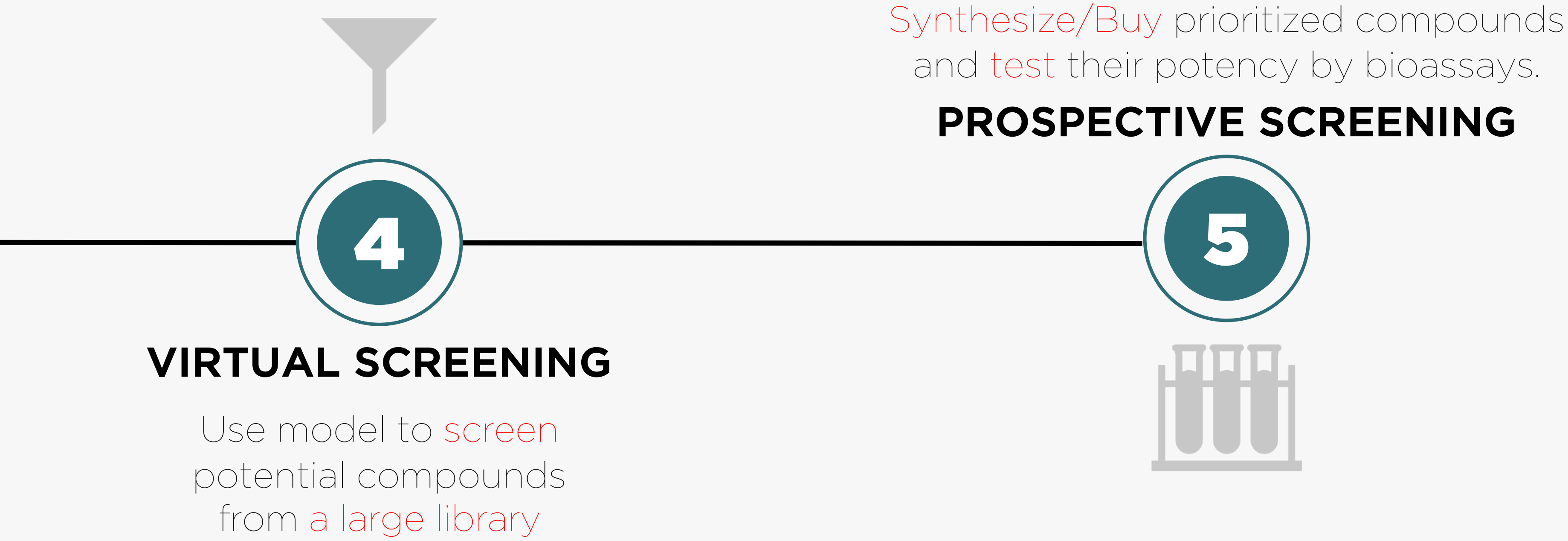
MOLECULAR REPRESENTATION

Discover and design optimal molecular representation schemes for our models.

3

AI MODEL DEVELOPEMENT

Discover optimals machine learning algorithms and train a deep learning model to predict Cdr1 inhibitors



VIRTUAL SCREENING

Use model to screen potential compounds from a large library

PROSPECTIVE SCREENING

Synthesize/Buy prioritized compounds and test their potency by bioassays.

2. METHODS

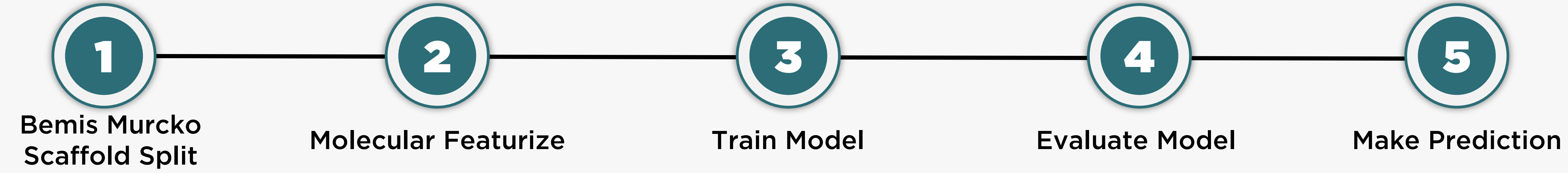
QSAR

Quantitative structure-activity relationship

QSAR is a mathematical model showing relationship

between **biological activity** and **molecular properties.**

$$\text{Bio\_activity} = f(D_1, D_2, D_3, \ldots, D_n)$$



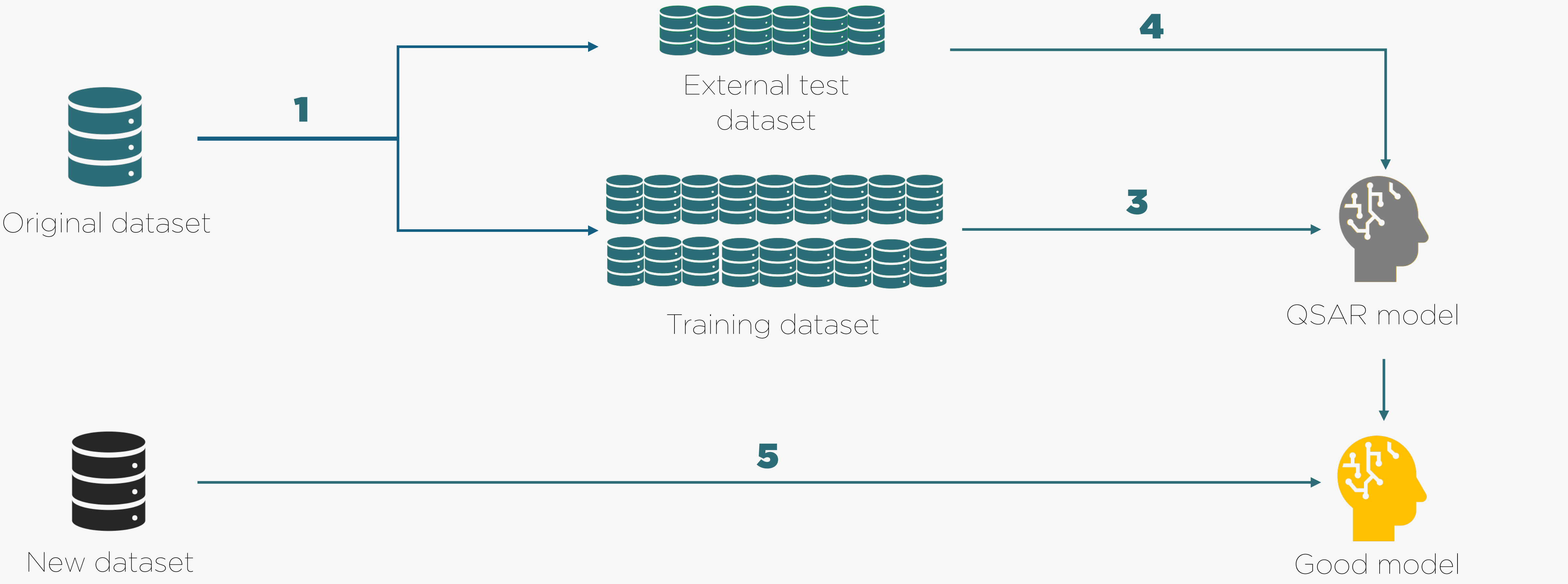
1 Bemis Murcko Scaffold Split

2 Molecular Featurize

3 Train Model

4 Evaluate Model

5 Make Prediction



Original dataset

External test dataset

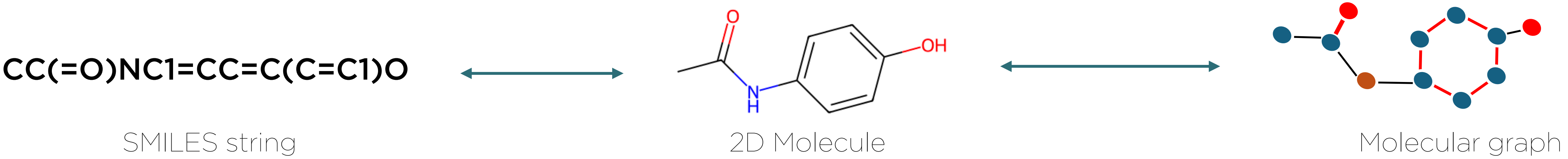
Training dataset

QSAR model

New dataset

Good model

MOLECULAR REPRESENTATION



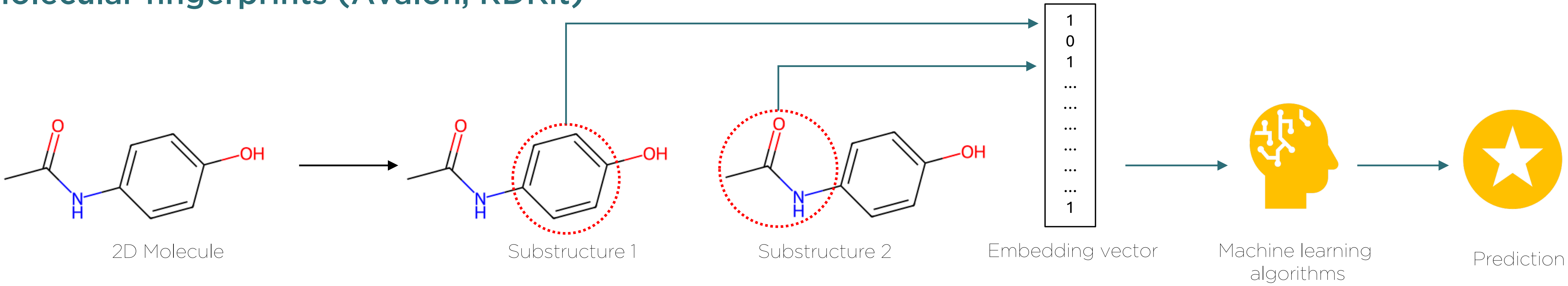
CC(=O)NC1=CC=C(C=C1)O

SMILES string

2D Molecule

Molecular graph

Molecular fingerprints (Avalon, RDKit)



2D Molecule

Substructure 1

Substructure 2

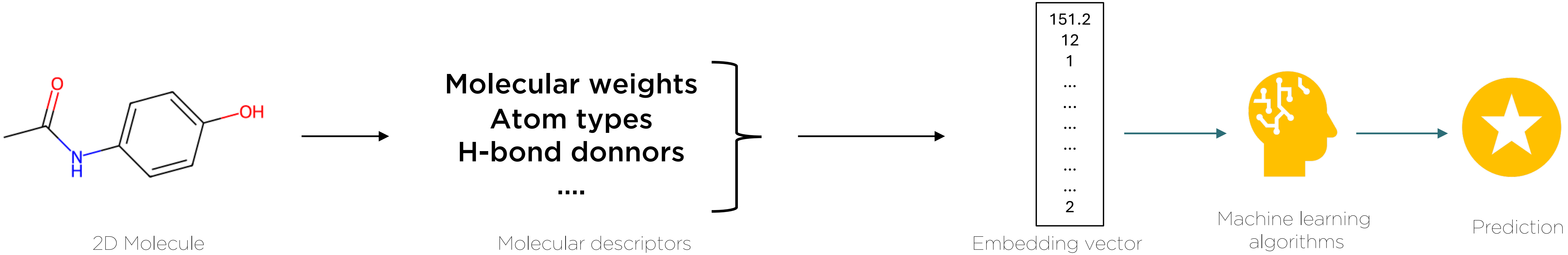
1
0
1
…
…
…
…
…
…
1

Embedding vector

Machine learning algorithms

Prediction

Molecular descriptors (Mordred)



Molecular weights
Atom types
H-bond donnors
....

2D Molecule

Molecular descriptors

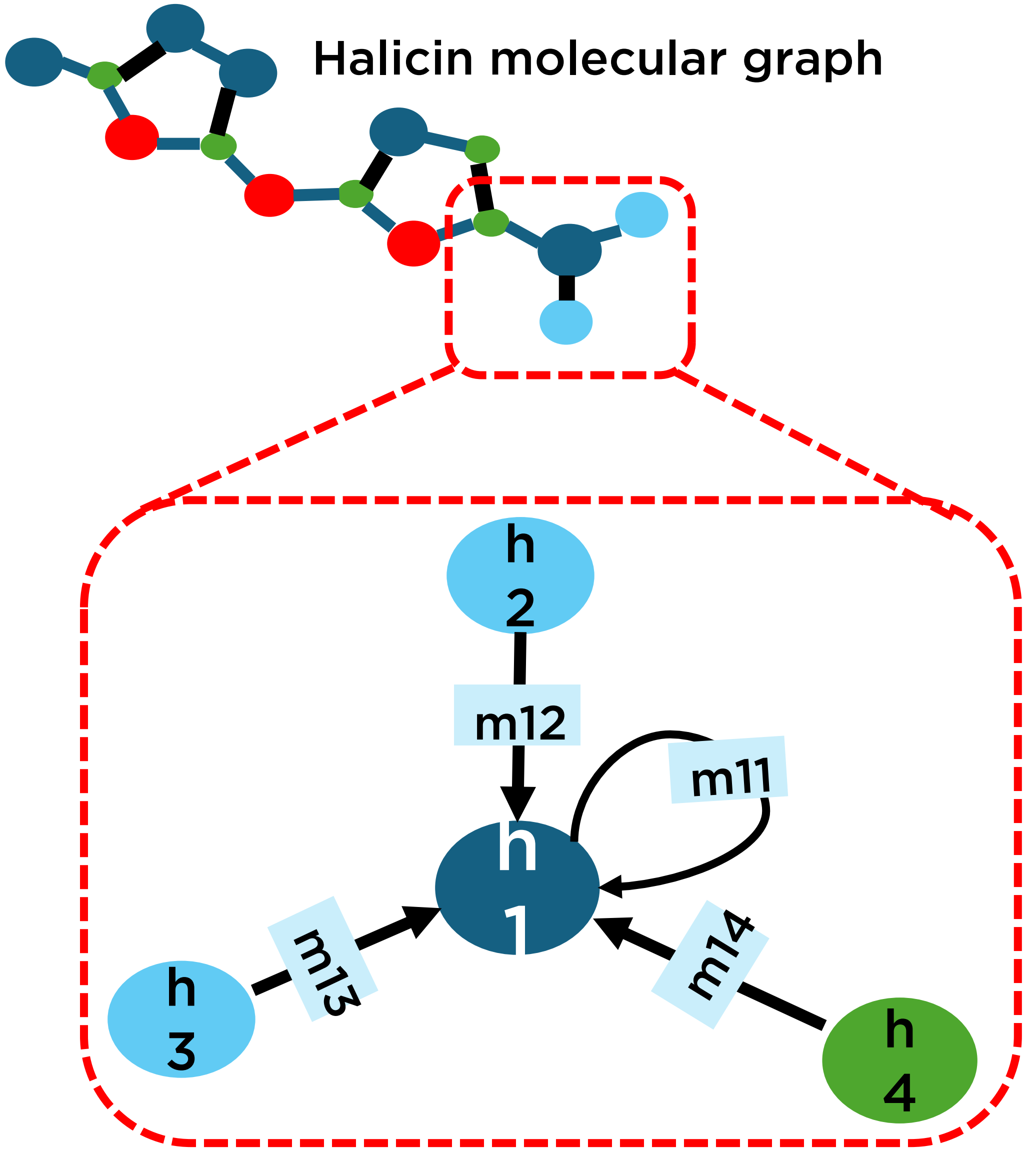
151.2
12
1
…
…
…
…
…
…
2

Embedding vector

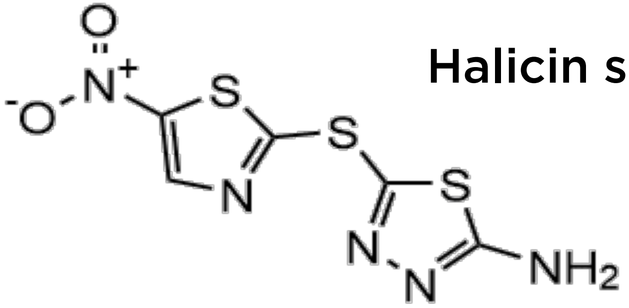
Machine learning algorithms

Prediction

MESSAGE PASSING



Halicin molecular graph



Halicin structures

- Message passing mechanism

$$m_v^{t+1} = \sum_{w \epsilon N(v)} M_t(h_v^t, h_w^t, e_{vw}) \quad (1)$$

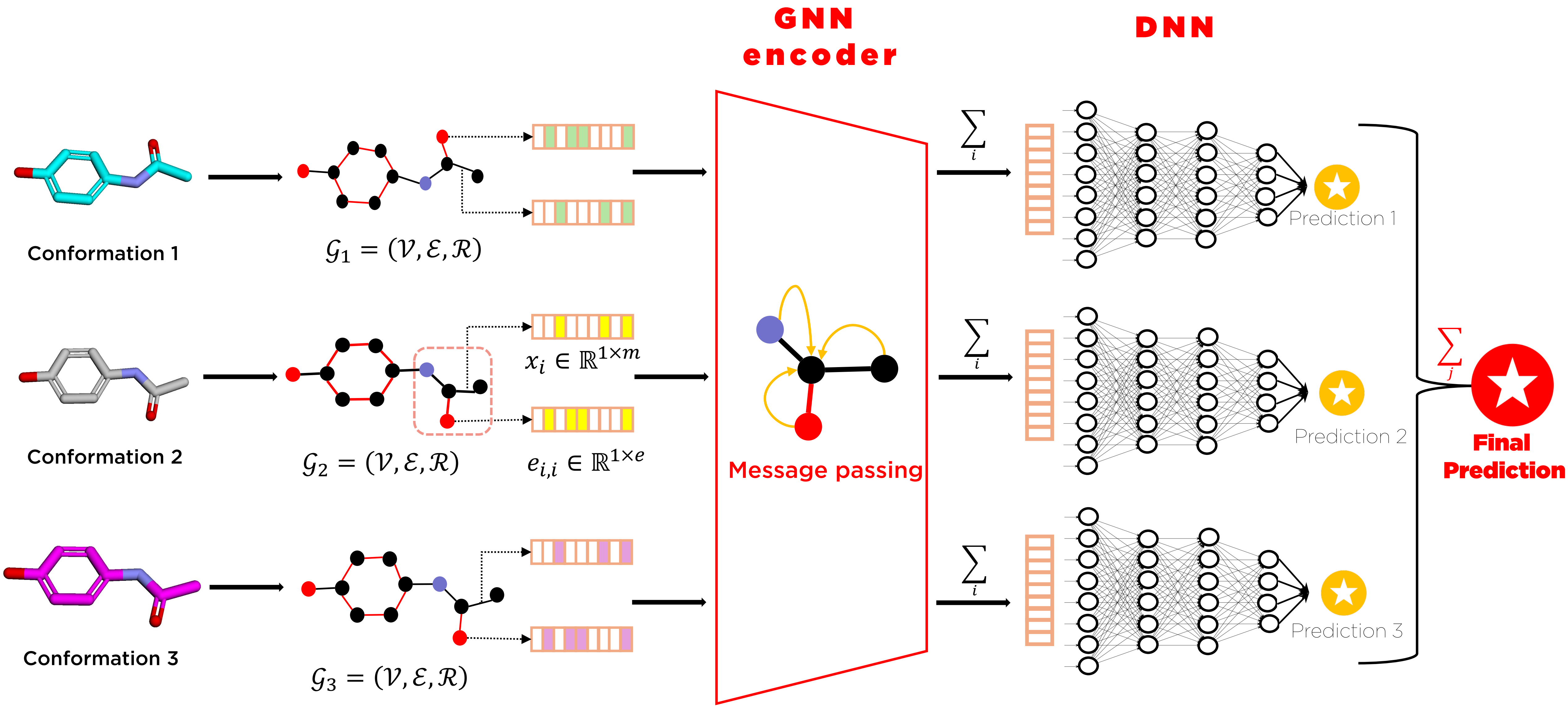
$$h_v^{t+1} = U_t(h_v^t, m_v^{t+1}) \quad (2)$$

$$h_G = \mathrm{R}(\{h_v^T | \, v \epsilon G\}) \quad (3)$$

- **Node features:** atomic number, number of bonds for each atom, formal charge, chirality, number of bonded hydrogens, hybridization, aromaticity, atomic mass.

- **Edge features:** bond type (single/double/triple/aromatic), conjugation, ring membership, stereochemistry.

Multi-instance 3D Graph neural network



GNN encoder

DNN

Conformation 1

$\mathcal{G}_1 = (\mathcal{V}, \mathcal{E}, \mathcal{R})$

Conformation 2

$\mathcal{G}_2 = (\mathcal{V}, \mathcal{E}, \mathcal{R})$

$x_i \in \mathbb{R}^{1\times m}$

$e_{i,i} \in \mathbb{R}^{1\times e}$

Conformation 3

$\mathcal{G}_3 = (\mathcal{V}, \mathcal{E}, \mathcal{R})$

Message passing

$\sum_i$

Prediction 1

Prediction 2

Prediction 3

$\sum_j$

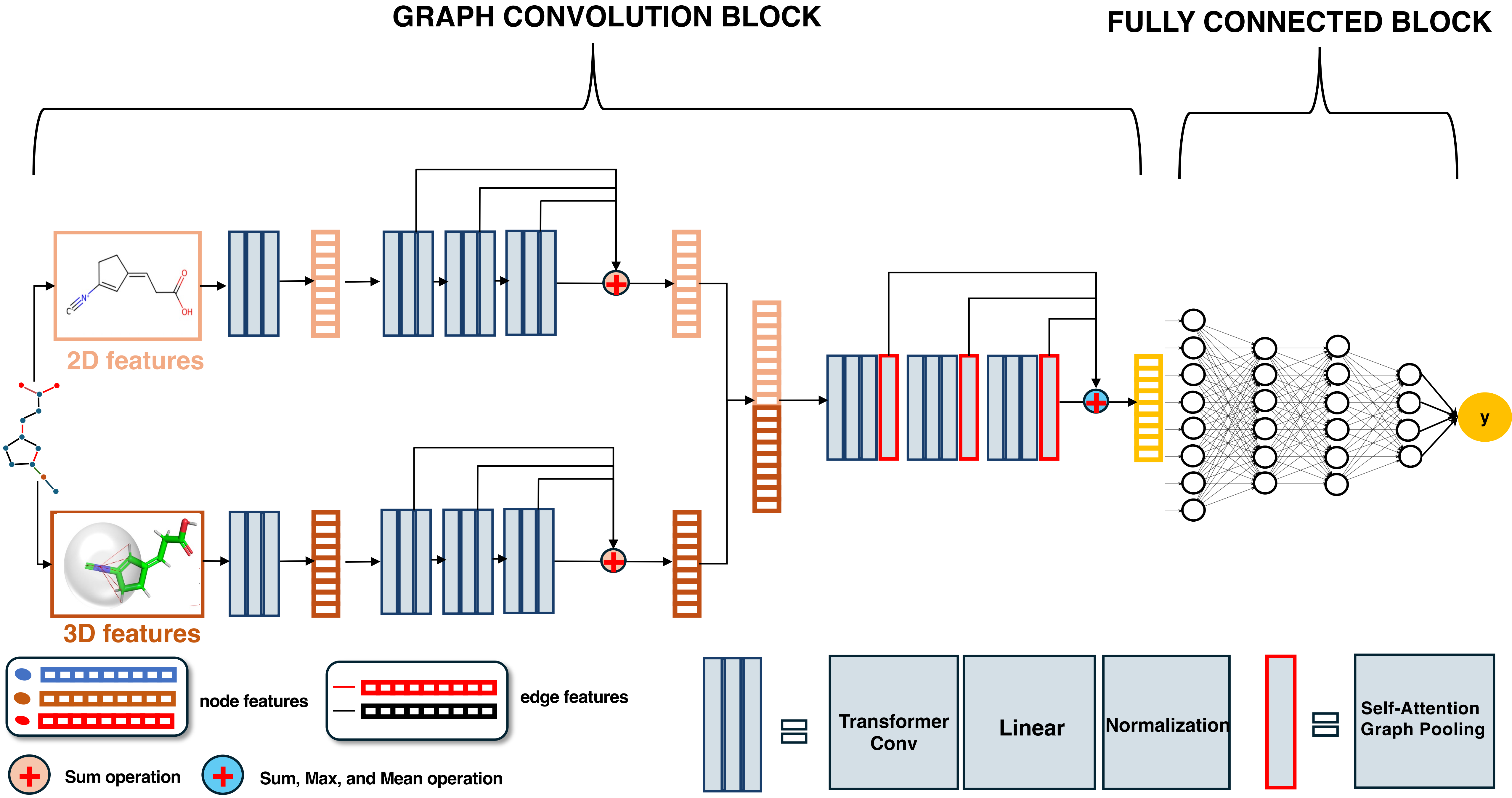
Final Prediction

GNN: Graph neural network

$\sum_i$ = Sum || Max || Mean operation

DNN: Deep neural network

$\sum_j$ = Mean operation


GRAPH CONVOLUTION BLOCK
FULLY CONNECTED BLOCK
2D features
3D features
y
node features
edge features
Sum operation
Sum, Max, and Mean operation
Transformer Conv
Linear
Normalization
Self-Attention Graph Pooling

EVALUATION METRICS

Confusion matrix

		Prediction	
		Active	**Inactive**
Ground truth	**Active**	True positive (TP)	False negative (FN)
	Inactive	False positive (FP)	True negative (TN)

Precision

$$\frac{\mathrm{TP}}{\mathrm{TP}+\mathrm{FP}}$$

Recall

$$\frac{\mathrm{TP}}{\mathrm{TP}+\mathrm{FN}}$$

Specificity

$$\frac{\mathrm{TN}}{\mathrm{FP}+\mathrm{FN}}$$

False positive rate (FPR)

$$\frac{\mathrm{FP}}{\mathrm{FP}+\mathrm{FN}}$$

1. Average precision

The area under Precision Recall curve

2. F1-score

The harmonic mean of Precision and Recall

3. ROC-AUC

The area under the

Receiver operating characteristic (ROC) curve

4. Balanced accuracy

The average between Recall and Specificity

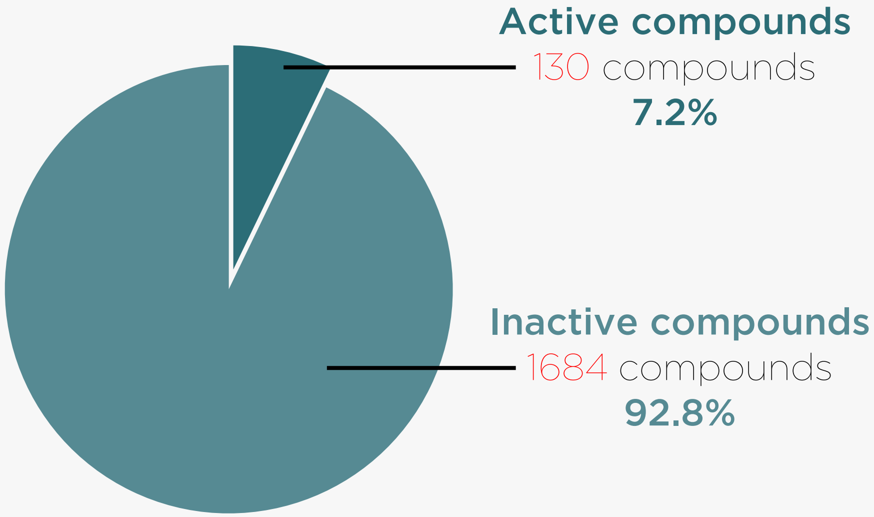
3. RESULTS

DATASET

Sources: Public repository (PubChem, ChEMBL), Literatures, Chemical patents (US11174267B2)

Partition method: Bemis Murcko scaffold and Local outlier factor

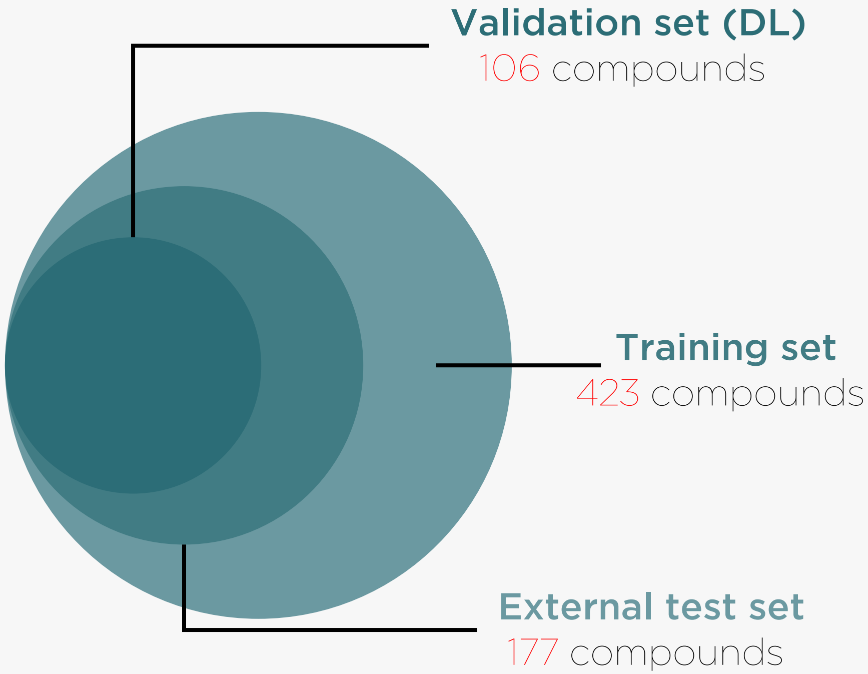
Active/Inactive ratio in each subset: 1/4.5



Active compounds
130 compounds
7.2%

Inactive compounds
1684 compounds
92.8%

Composition in the original dataset

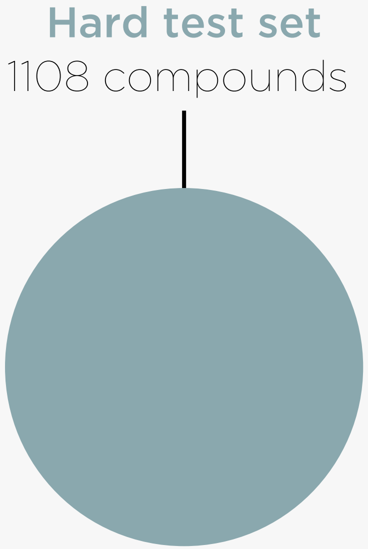


Validation set (DL)
106 compounds

Training set
423 compounds

External test set
177 compounds

Composition in the each subset



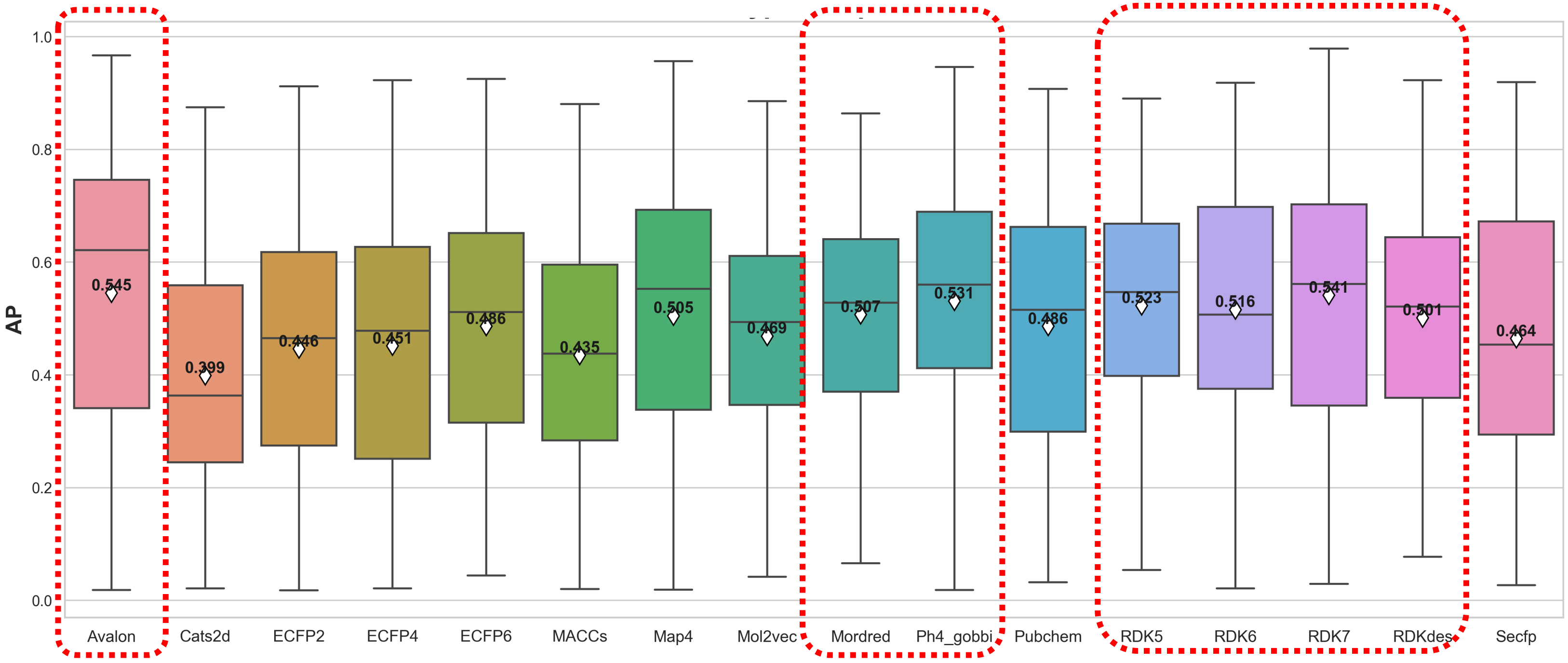
Hard test set
1108 compounds

Inactive compounds outside the chemical space (Outliers)

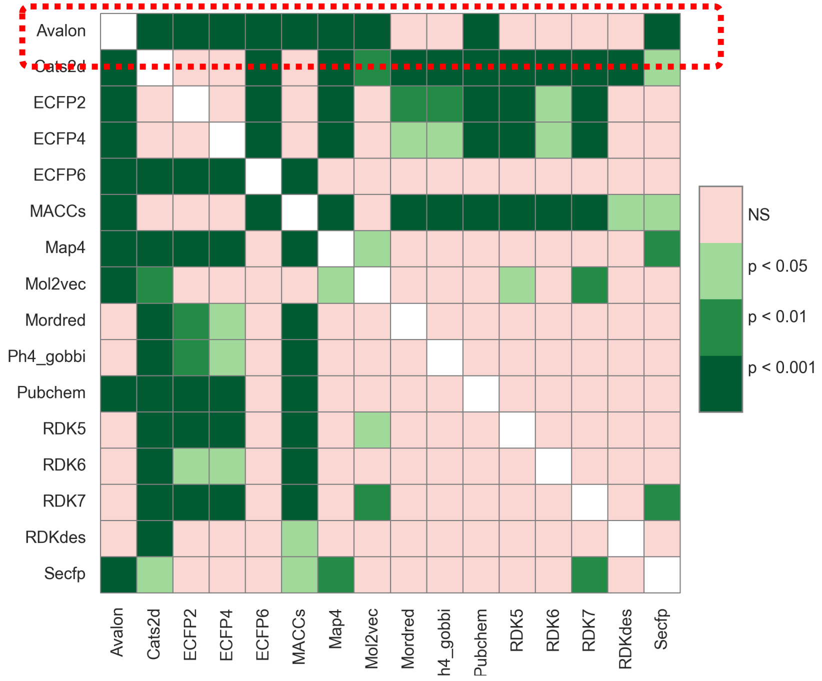
MOLECULAR REPRESENTATION

Meta-analysis

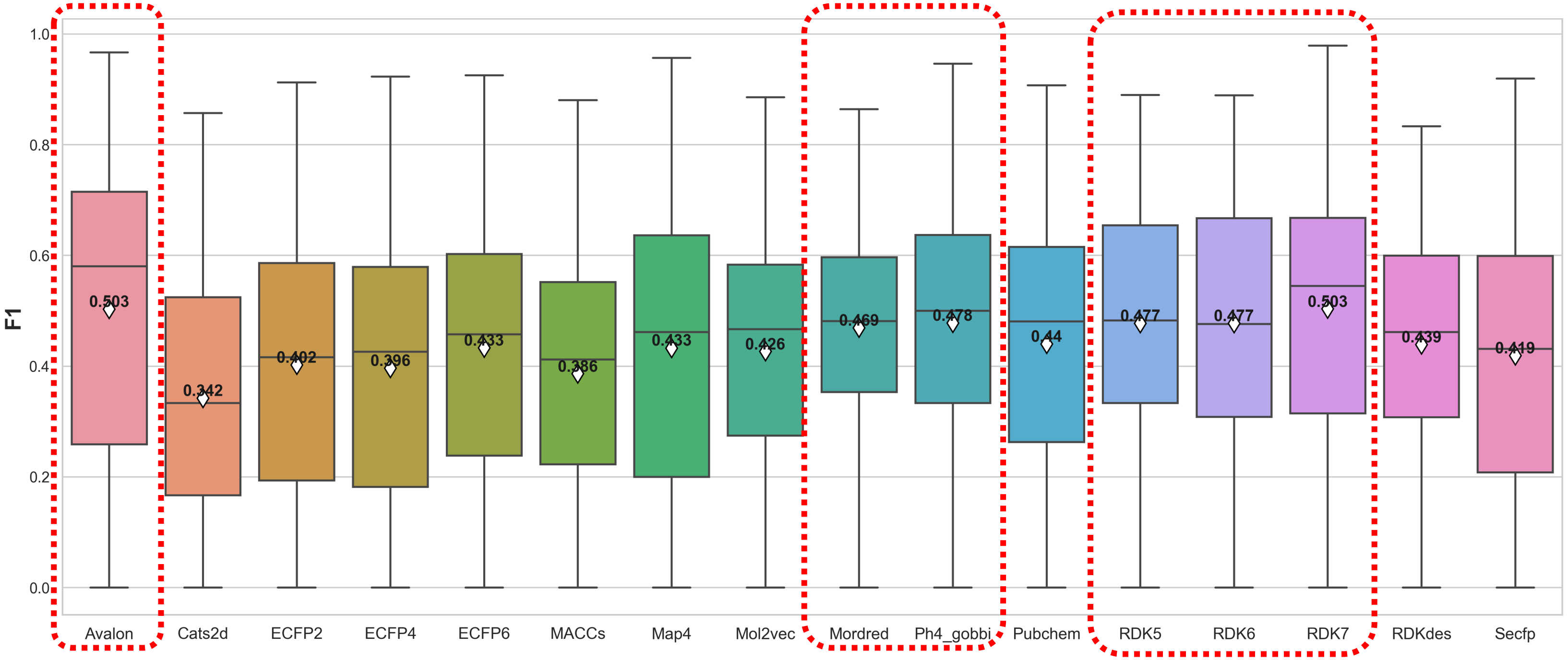
16 types of molecular representations



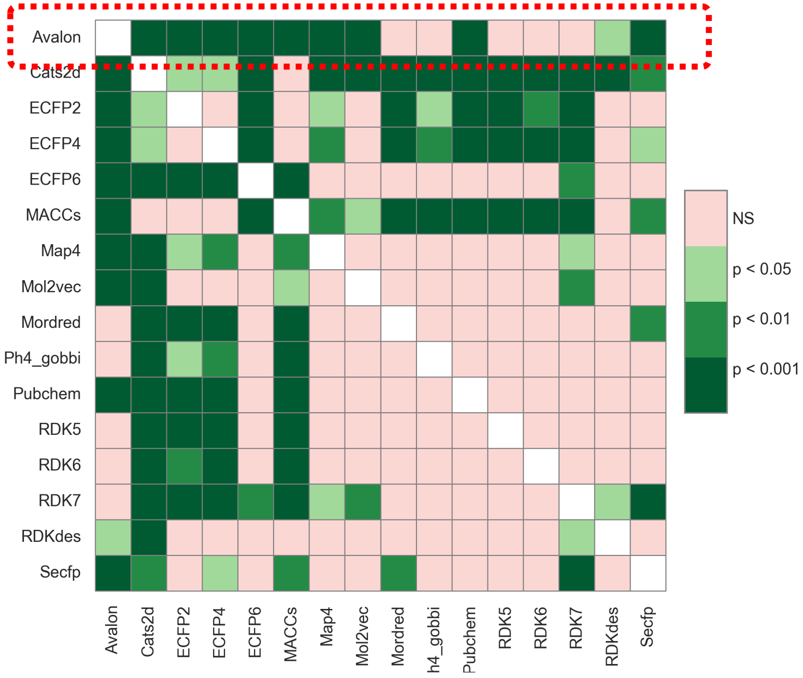
Boxplots comparing the BM 10-fold cross validation results based on average precision



Heat map illustrating the results of Wilcoxon signed-rank tests based on average precision



Boxplots comparing the BM 10-fold cross validation results based on F1-score



Heat map illustrating the results of Wilcoxon signed-rank tests based on F1-score

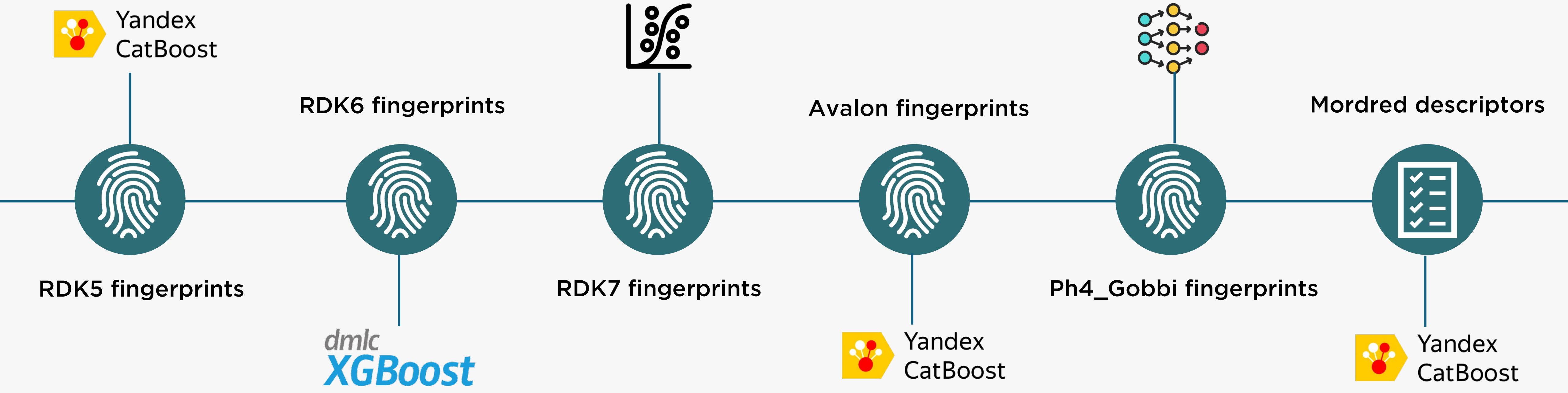
Molecular fingerprints: RDK5, RDK6, RDK7, Avalon, Gobbi Pharmacophore fingerprints

Molecular descriptors: Mordred descriptors

MOLECULAR REPRESENTATION

Model selection

10 machine learning algorithms: Logistic regression, K-nearest neighbors, Support vector machine, Random forest, Extra tree, AdaBoost, Gradient boosting, XGBoost, CatBoost, and Multi-layer perceptron



Yandex CatBoost

RDK5 fingerprints

RDK6 fingerprints

dmlc XGBoost

RDK7 fingerprints

Avalon fingerprints

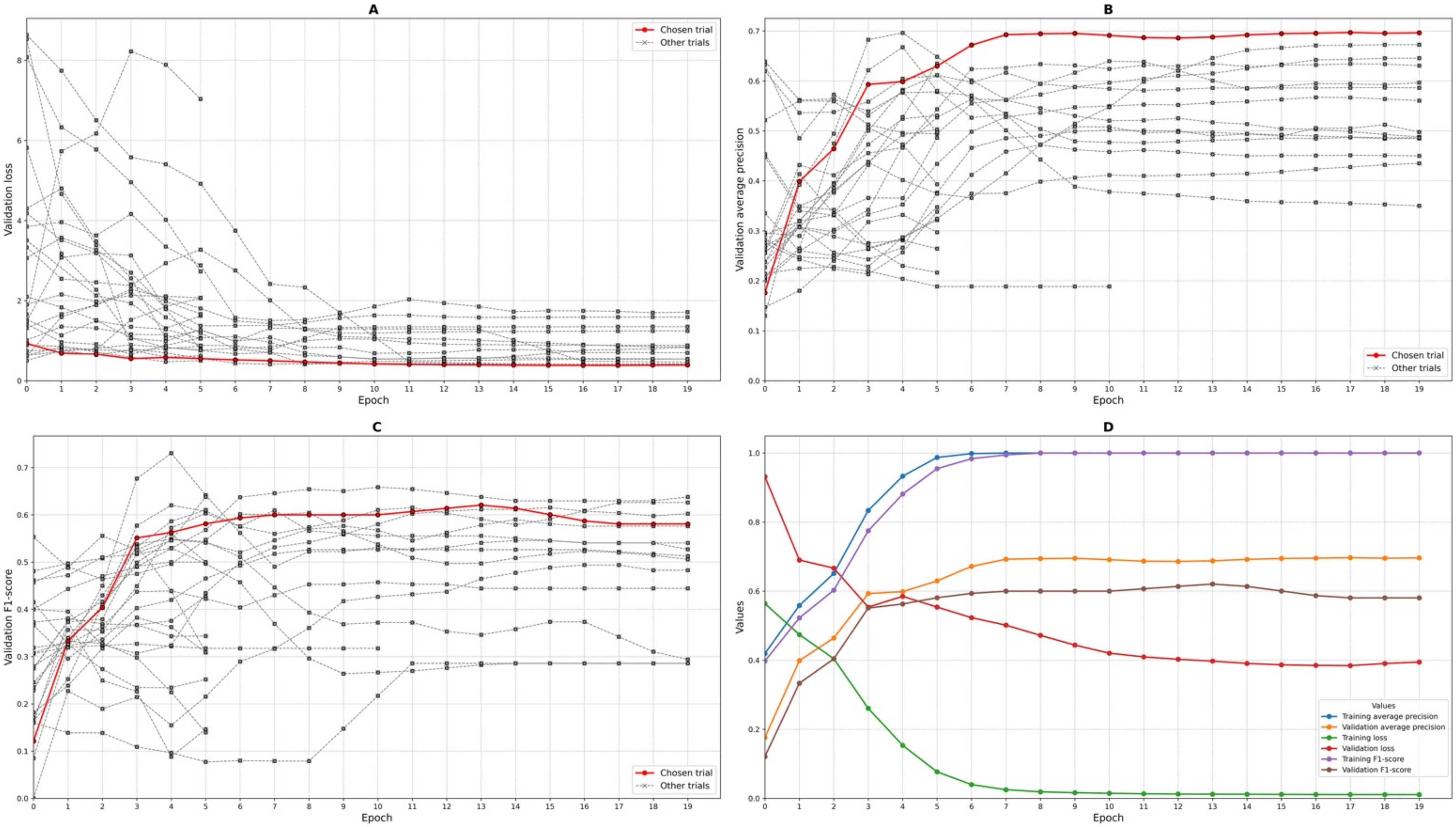
Yandex CatBoost

Ph4_Gobbi fingerprints

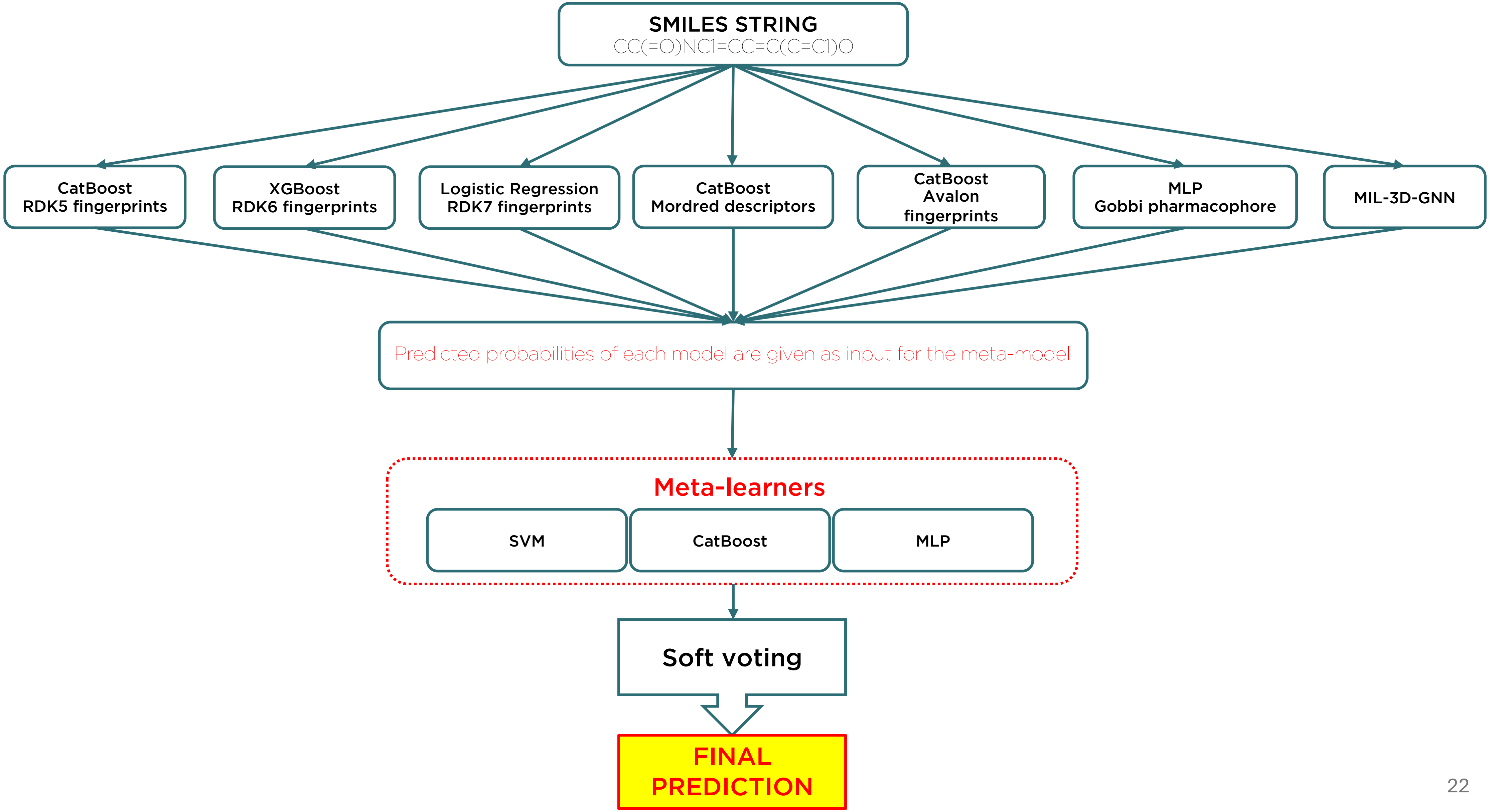
Mordred descriptors

Yandex CatBoost

TUNNING HYPERPARAMETERS



ENSEMBLE MODEL



SMILES STRING
CC(=O)NC1=CC=C(C=C1)O

CatBoost RDK5 fingerprints

XGBoost RDK6 fingerprints

Logistic Regression RDK7 fingerprints

CatBoost Mordred descriptors

CatBoost Avalon fingerprints

MLP Gobbi pharmacophore

MIL-3D-GNN

Predicted probabilities of each model are given as input for the meta-model

Meta-learners

SVM

CatBoost

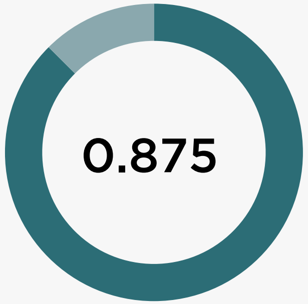
MLP

Soft voting

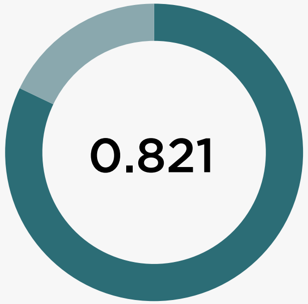
FINAL PREDICTION

MODEL PERFORMANCE

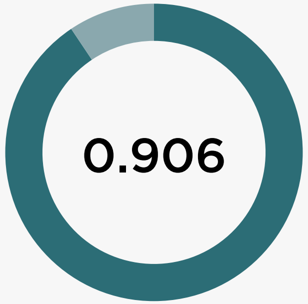
Validation set



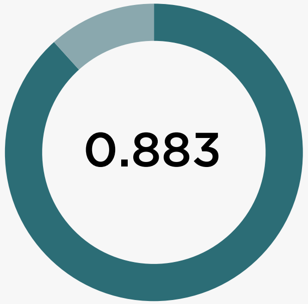
Average precision



F1-score

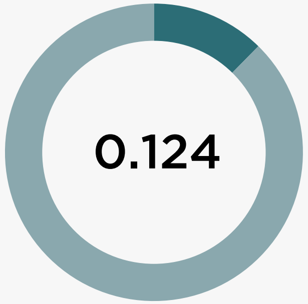


ROC-AUC



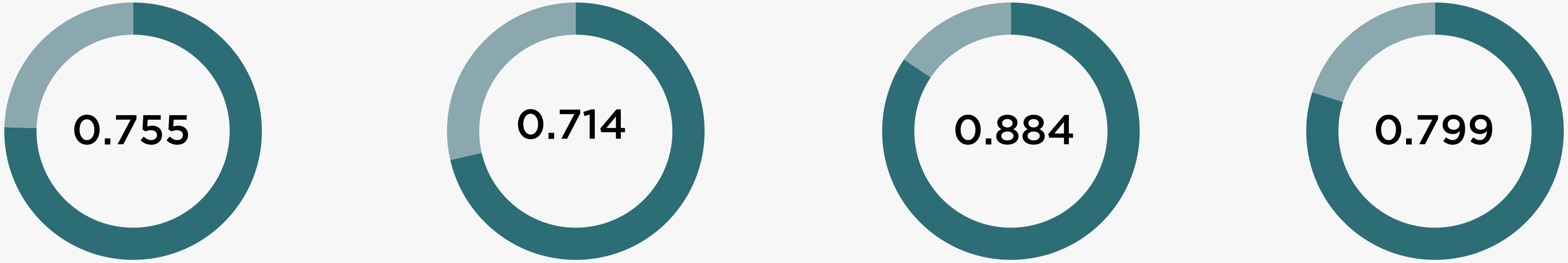
Balanced accuracy

Hard test set



False positve rate

External test set



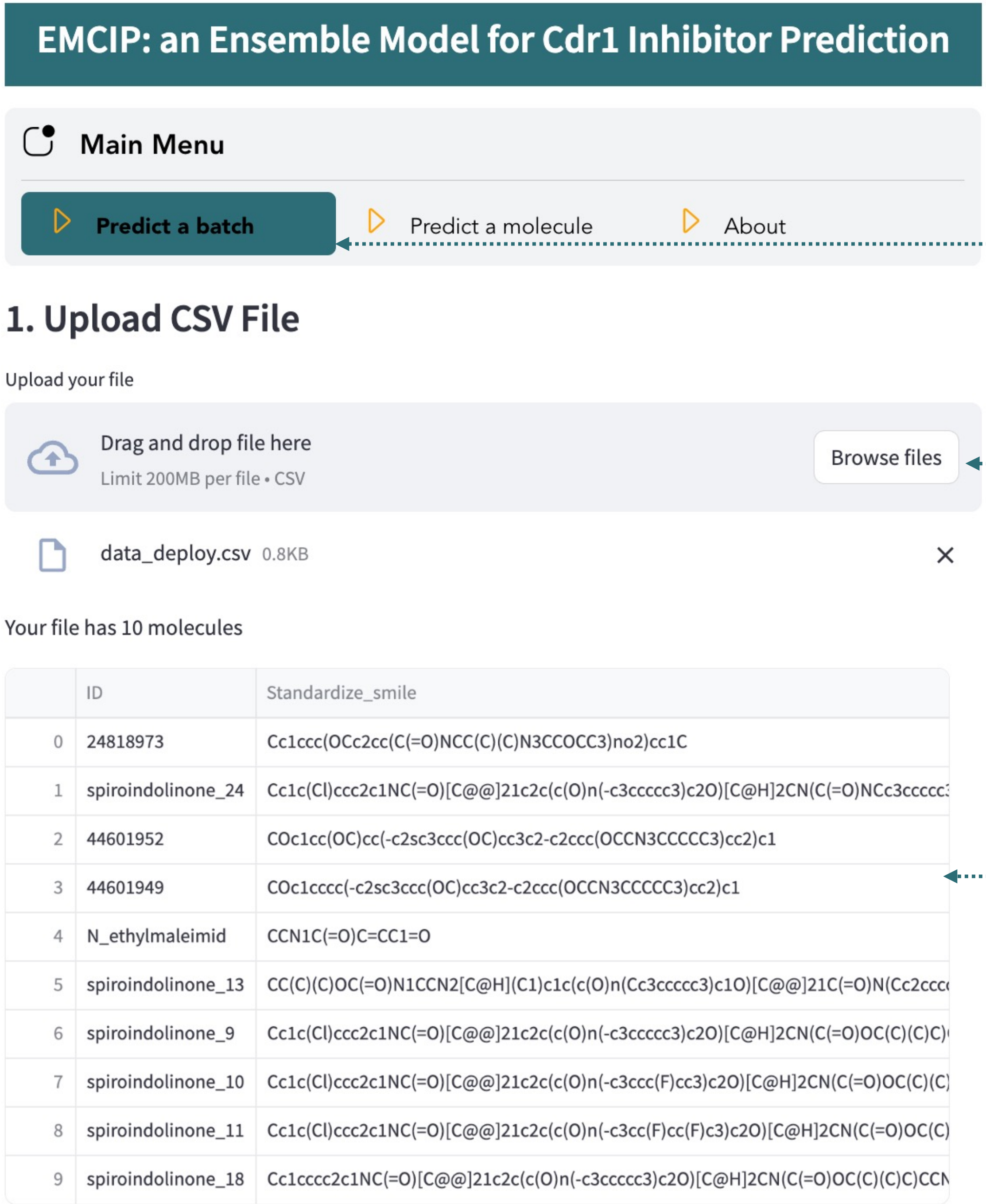
Average precision F1-score ROC-AUC Balanced accuracy

The generalizability of the EMCIP model and its effectiveness on unseen data

DEPLOYMENT

scikit learn PyTorch PyTorch geometric Streamlit

EMCIP GUI



EMCIP: an Ensemble Model for Cdr1 Inhibitor Prediction

1. Upload CSV File

Upload your file

Your file has 10 molecules

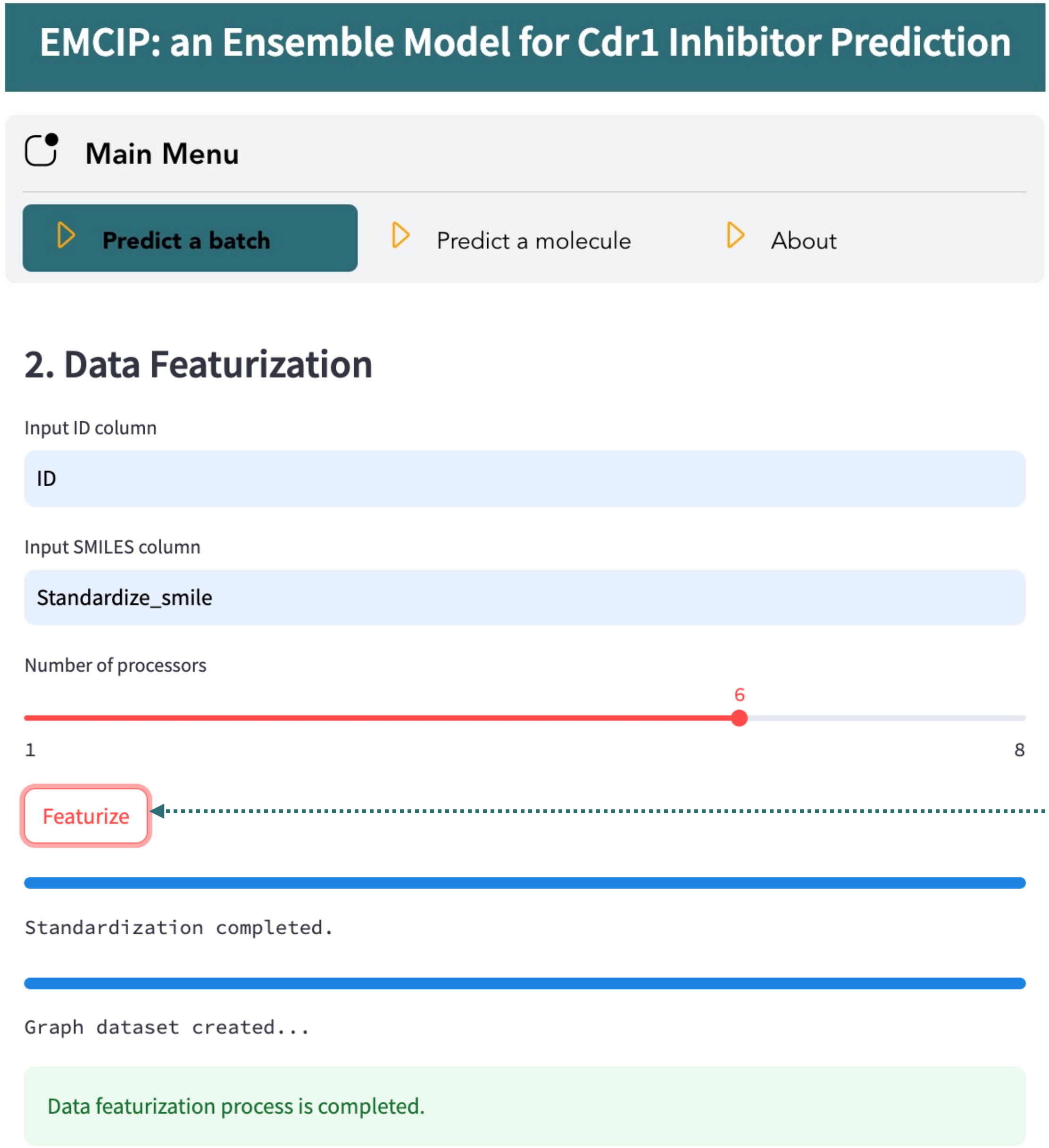
	ID	Standardize_smile
0	24818973	Cc1ccc(OCc2cc(C(=O)NCC(C)(C)N3CCOCC3)no2)cc1C
1	spiroindolinone_24	Cc1c(Cl)ccc2c1NC(=O)[C@@]21c2c(c(O)n(-c3ccccc3)c2O)[C@H]2CN(C(=O)NCc3ccccc
2	44601952	COc1cc(OC)cc(-c2sc3ccc(OC)cc3c2-c2ccc(OCCN3CCCCC3)cc2)c1
3	44601949	COc1cccc(-c2sc3ccc(OC)cc3c2-c2ccc(OCCN3CCCCC3)cc2)c1
4	N_ethylmaleimid	CCN1C(=O)C=CC1=O
5	spiroindolinone_13	CC(C)(C)OC(=O)N1CCN2[C@H](C1)c1c(c(O)n(Cc3ccccc3)c1O)[C@@]21C(=O)N(Cc2ccc
6	spiroindolinone_9	Cc1c(Cl)ccc2c1NC(=O)[C@@]21c2c(c(O)n(-c3ccccc3)c2O)[C@H]2CN(C(=O)OC(C)(C)C)
7	spiroindolinone_10	Cc1c(Cl)ccc2c1NC(=O)[C@@]21c2c(c(O)n(-c3ccc(F)cc3)c2O)[C@H]2CN(C(=O)OC(C)(C)
8	spiroindolinone_11	Cc1c(Cl)ccc2c1NC(=O)[C@@]21c2c(c(O)n(-c3cc(F)cc(F)c3)c2O)[C@H]2CN(C(=O)OC(C)
9	spiroindolinone_18	Cc1cccc2c1NC(=O)[C@@]21c2c(c(O)n(-c3ccccc3)c2O)[C@H]2CN(C(=O)OC(C)(C)C)CCN

Batch prediction

Upload a csv file

SMILES and ID table

EMCIP GUI



2. Data Featurization

Input ID column: ID

Input SMILES column: Standardize_smile

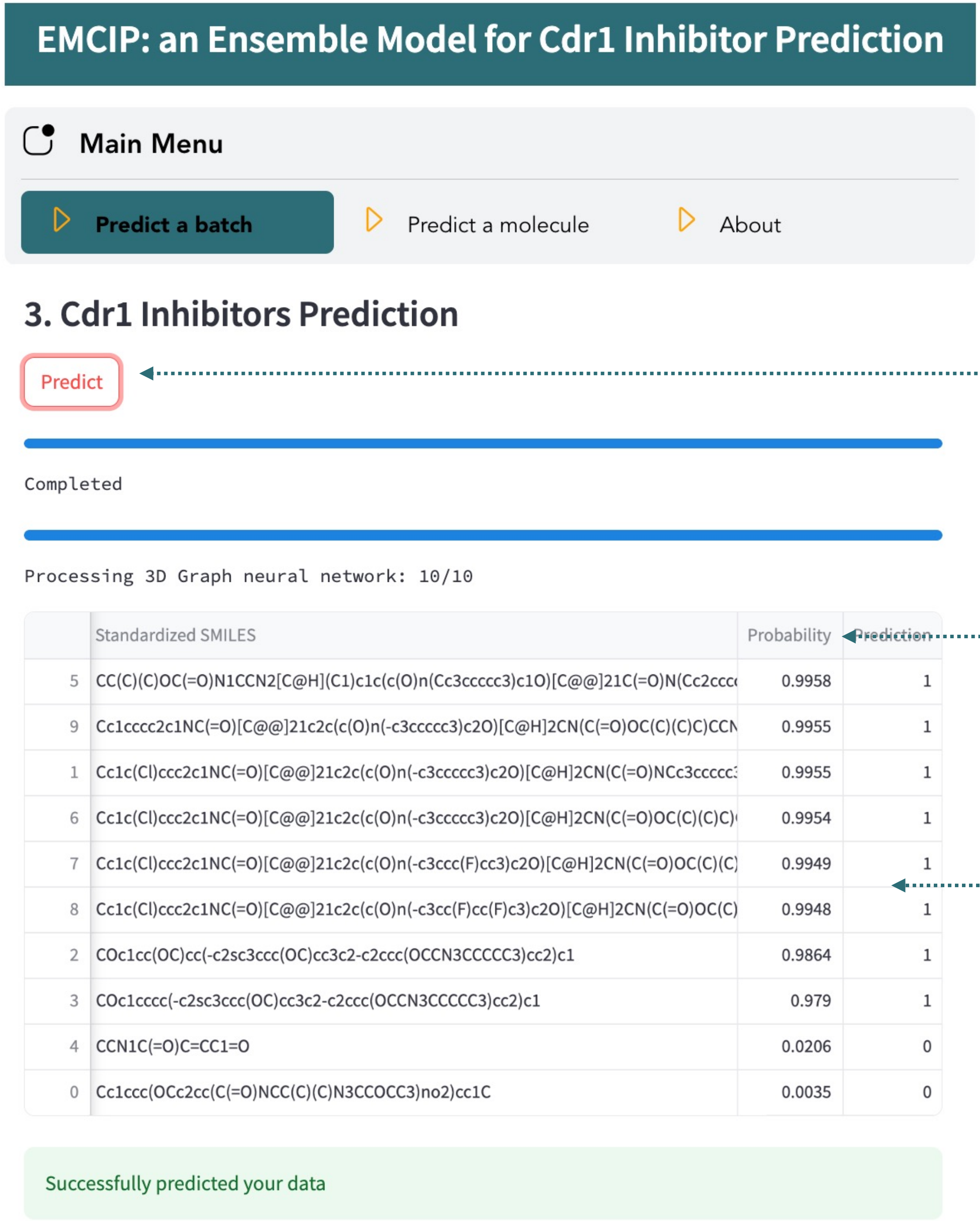
Number of processors: 6 (1–8)

```
Standardization completed.

Graph dataset created...
```

Data featurization process is completed.

Featurize molecules

EMCIP GUI



3. Cdr1 Inhibitors Prediction

```
Completed

Processing 3D Graph neural network: 10/10
```

	Standardized SMILES	Probability	Prediction
5	CC(C)(C)OC(=O)N1CCN2[C@H](C1)c1c(c(O)n(Cc3ccccc3)c1O)[C@@]21C(=O)N(Cc2ccc	0.9958	1
9	Cc1cccc2c1NC(=O)[C@@]21c2c(c(O)n(-c3ccccc3)c2O)[C@H]2CN(C(=O)OC(C)(C)C)CCN	0.9955	1
1	Cc1c(Cl)ccc2c1NC(=O)[C@@]21c2c(c(O)n(-c3ccccc3)c2O)[C@H]2CN(C(=O)NCc3cccc	0.9955	1
6	Cc1c(Cl)ccc2c1NC(=O)[C@@]21c2c(c(O)n(-c3ccccc3)c2O)[C@H]2CN(C(=O)OC(C)(C)C)	0.9954	1
7	Cc1c(Cl)ccc2c1NC(=O)[C@@]21c2c(c(O)n(-c3ccc(F)cc3)c2O)[C@H]2CN(C(=O)OC(C)(C)	0.9949	1
8	Cc1c(Cl)ccc2c1NC(=O)[C@@]21c2c(c(O)n(-c3cc(F)cc(F)c3)c2O)[C@H]2CN(C(=O)OC(C)	0.9948	1
2	COc1cc(OC)cc(-c2sc3ccc(OC)cc3c2-c2ccc(OCCN3CCCCC3)cc2)c1	0.9864	1
3	COc1cccc(-c2sc3ccc(OC)cc3c2-c2ccc(OCCN3CCCCC3)cc2)c1	0.979	1
4	CCN1C(=O)C=CC1=O	0.0206	0
0	Cc1ccc(OCc2cc(C(=O)NCC(C)(C)N3CCOCC3)no2)cc1C	0.0035	0

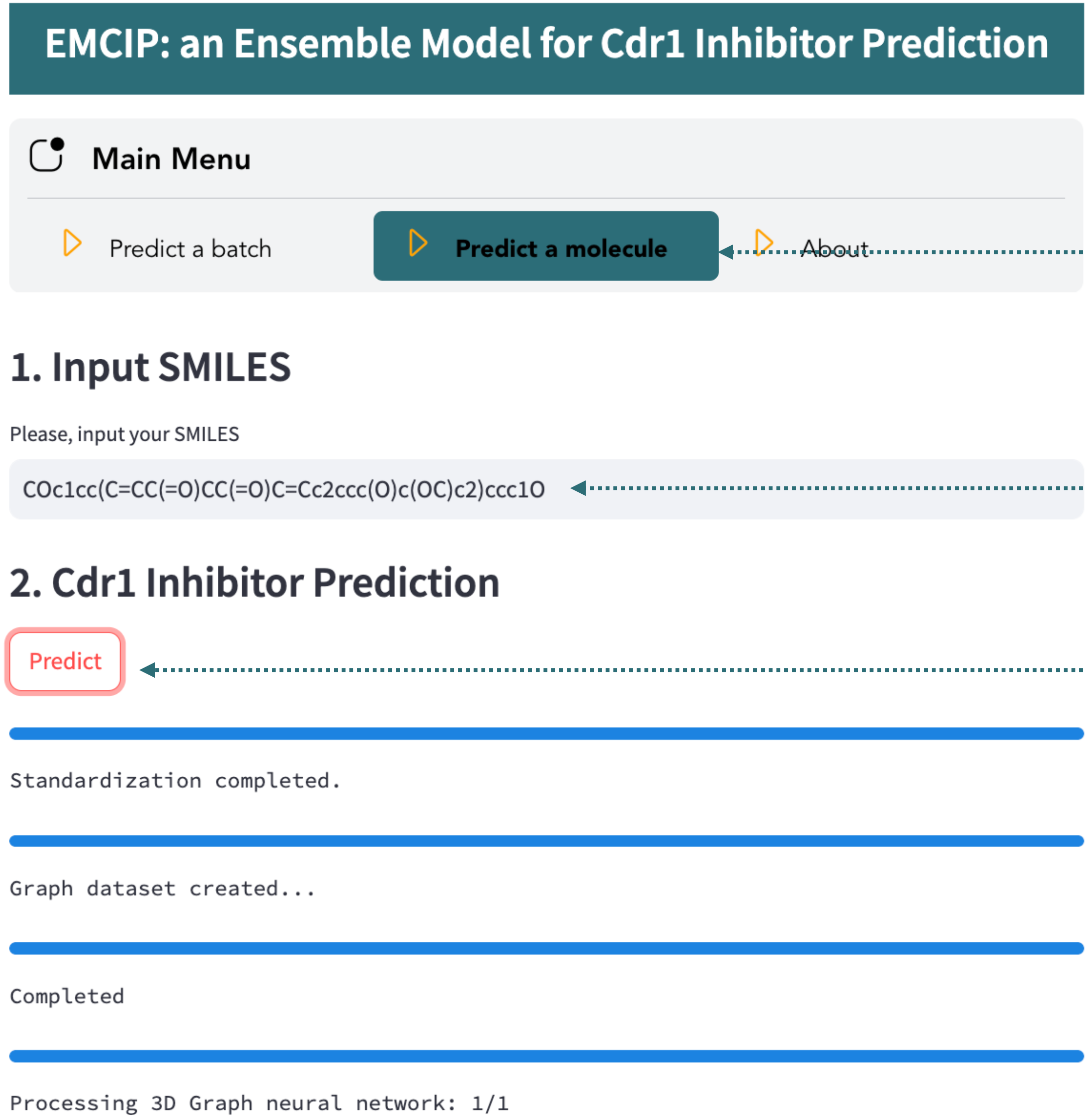
Successfully predicted your data

Predict molecules

Probability of being a Cdr1 inhibitor

Result table

EMCIP GUI



EMCIP: an Ensemble Model for Cdr1 Inhibitor Prediction

1. Input SMILES

Please, input your SMILES

COc1cc(C=CC(=O)CC(=O)C=Cc2ccc(O)c(OC)c2)ccc1O

2. Cdr1 Inhibitor Prediction

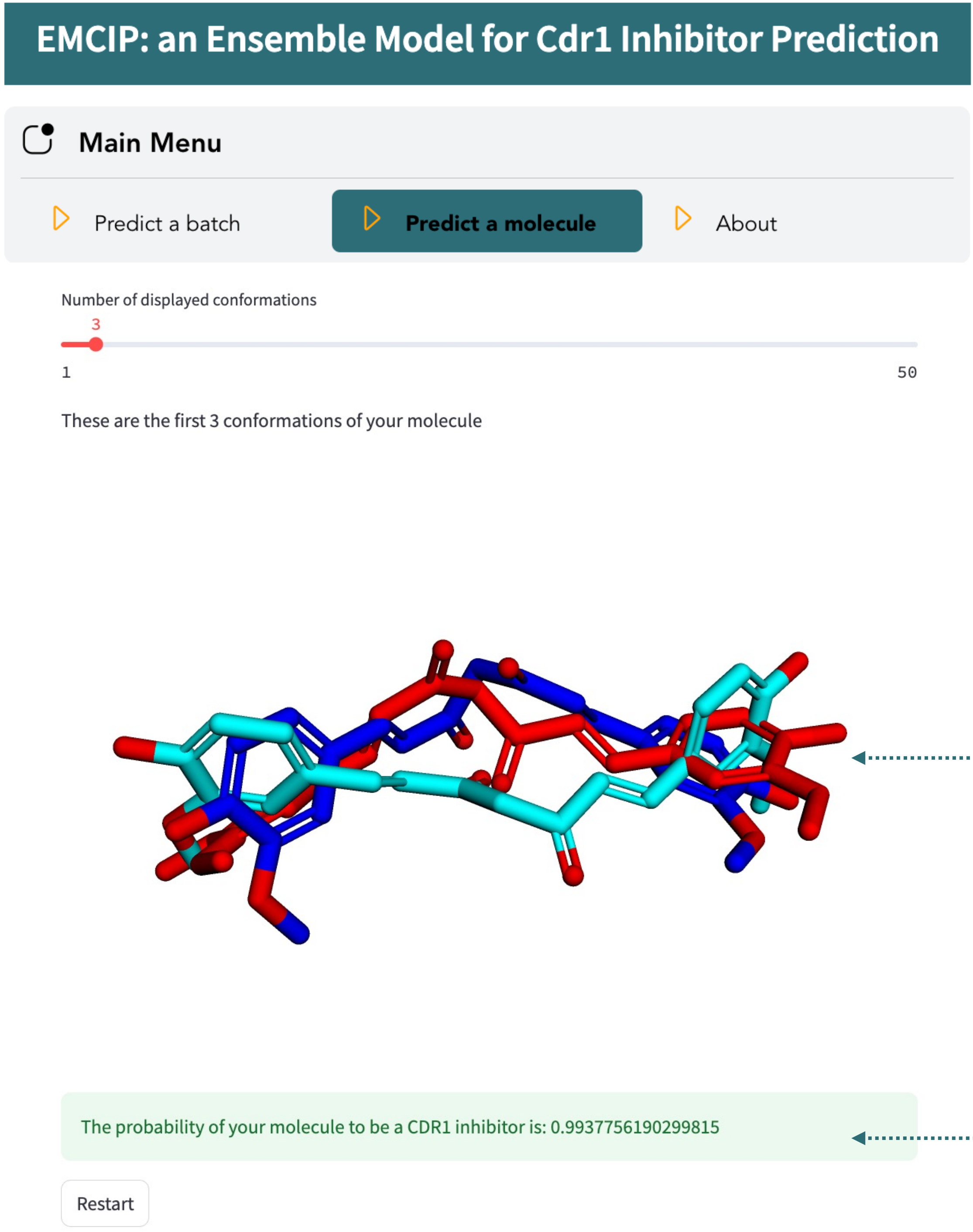
```
Standardization completed.

Graph dataset created...

Completed

Processing 3D Graph neural network: 1/1
```

Predict a molecules

Input a SMILES string

Predict

EMCIP GUI



EMCIP: an Ensemble Model for Cdr1 Inhibitor Prediction

Number of displayed conformations

3

1

50

These are the first 3 conformations of your molecule

The probability of your molecule to be a CDR1 inhibitor is: 0.9937756190299815

Generated conformations

Probability of being a Cdr1 inhibitor

4. DISCUSSION

❖ Conclusion

- **EMCIP:** The first ensemble machine learning model specific for Cdr1 inhibitor prediction.
- **MIL-3D-GNN:** A novel 3D graph neural network for multi-instance learning**.**
- **Promising results on the external test set and validation set,** demonstrating the generalizability of the EMCIP model and its effectiveness on unseen data.
- **A practical GUI for EMCIP**, making it accessible and user-friendly even to AI non-experts.
- **A practical workflow**, conducting **ligand-based predictive AI models for other targets.**

❖ Limitations

- **Data scarcity**: Test more compounds to augment the dataset.
- **Lack of experimental structure of Cdr1 protein**: Prevents integration of protein information and implementation of structure-based drug design.

THANK YOU FOR YOUR ATTENTION

Email: the-chuong.trinh@etu.univ-grenoble-alpes.fr

EMCIP Online Version