Multi-Instance Explainable Learning

for decoding stereo-dependent biological effects

Vasilii Fastovskii

PhD Candidate,
Strasbourg University,
Laboratory of Chemoinformatics

Sanofi

Dr Marc Bianciotto

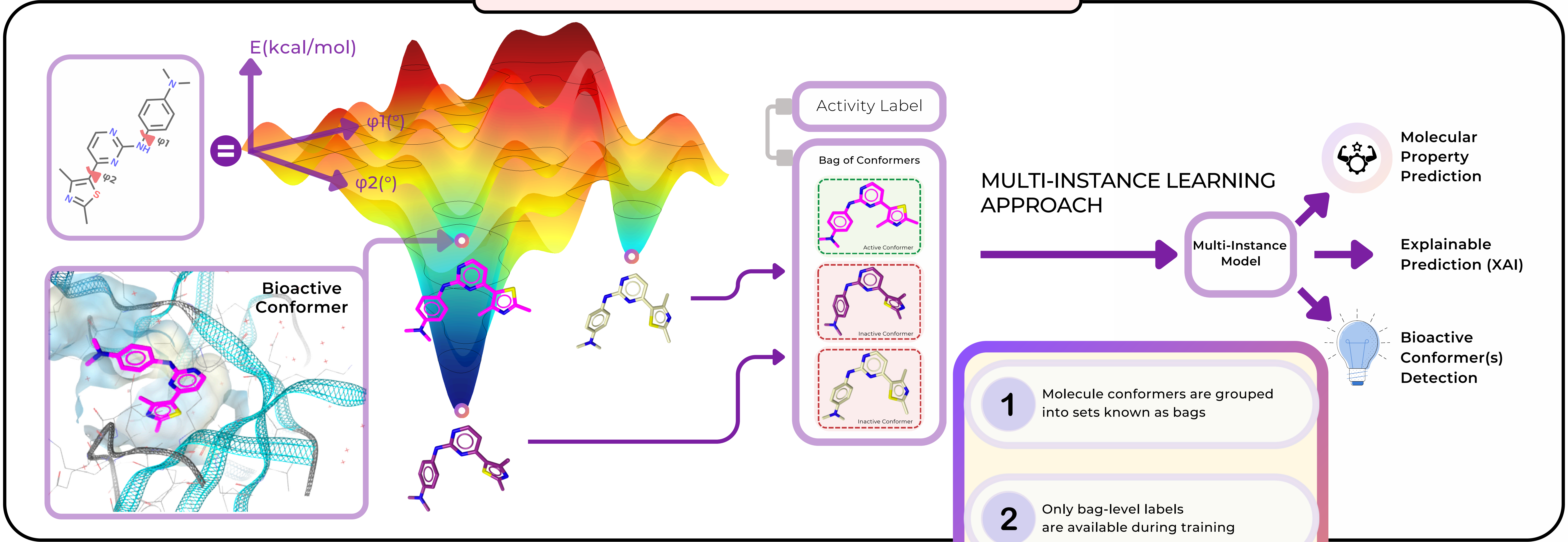
Dr Christoph Grebner

Strasbourg University,
Laboratory of Chemoinformatics

Dr Gilles Marcou

Dr Prof Alexandre Varnek

1. INTRODUCTION



Bioactive Conformer

Activity Label

Bag of Conformers

MULTI-INSTANCE LEARNING APPROACH

Multi-Instance Model

Molecular Property Prediction

Explainable Prediction (XAI)

Bioactive Conformer(s) Detection

1 Molecule conformers are grouped into sets known as bags

2 Only bag-level labels are available during training

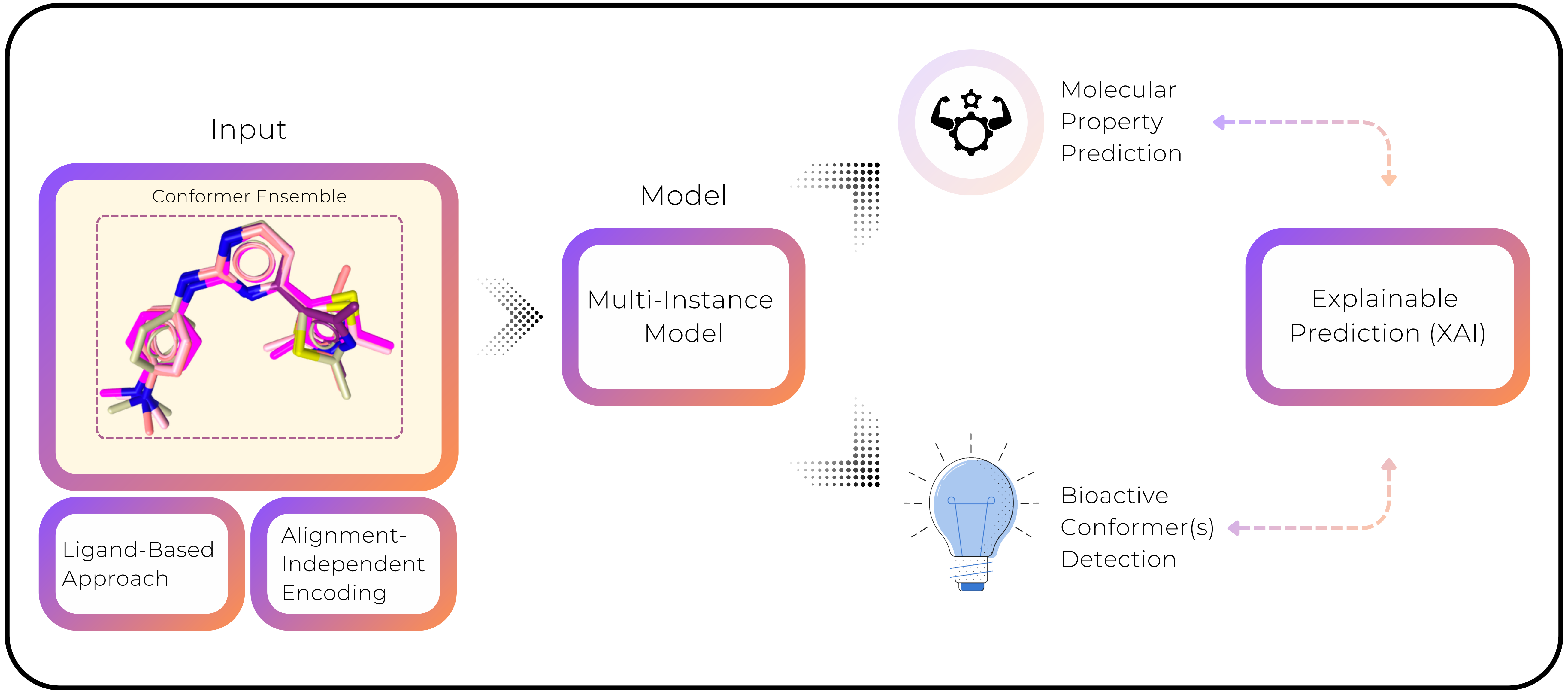
3 Standard MIL Task - Train a classifier/regressor that labels new bags

4 Key Instance Detection (KID) Task - Identify the instances (conformers) most responsible for the bag's label

2. PROJECT GOALS

We aim to identify the exact molecular form(s) responsible for the observed/predicted properties to provide a data-driven three-dimensional shape hypothesis supporting a ligand-based model

Molecules can adopt multiple forms (conformers, tautomers, protonation states)



Input

Conformer Ensemble

Ligand-Based Approach

Alignment-Independent Encoding

Model

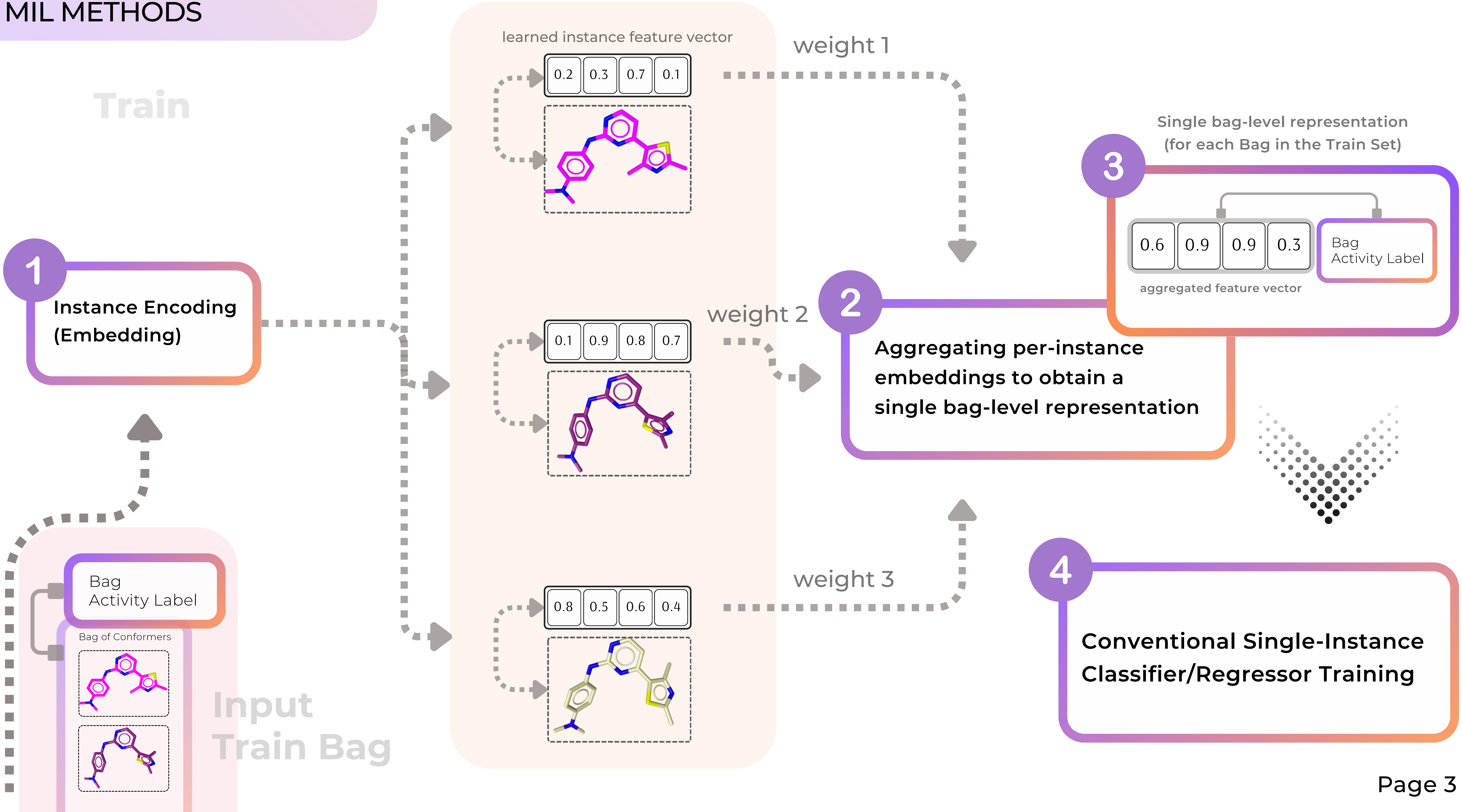
Multi-Instance Model

Molecular Property Prediction

Bioactive Conformer(s) Detection

Explainable Prediction (XAI)

EMBEDDED-SPACE (ES) MIL METHODS



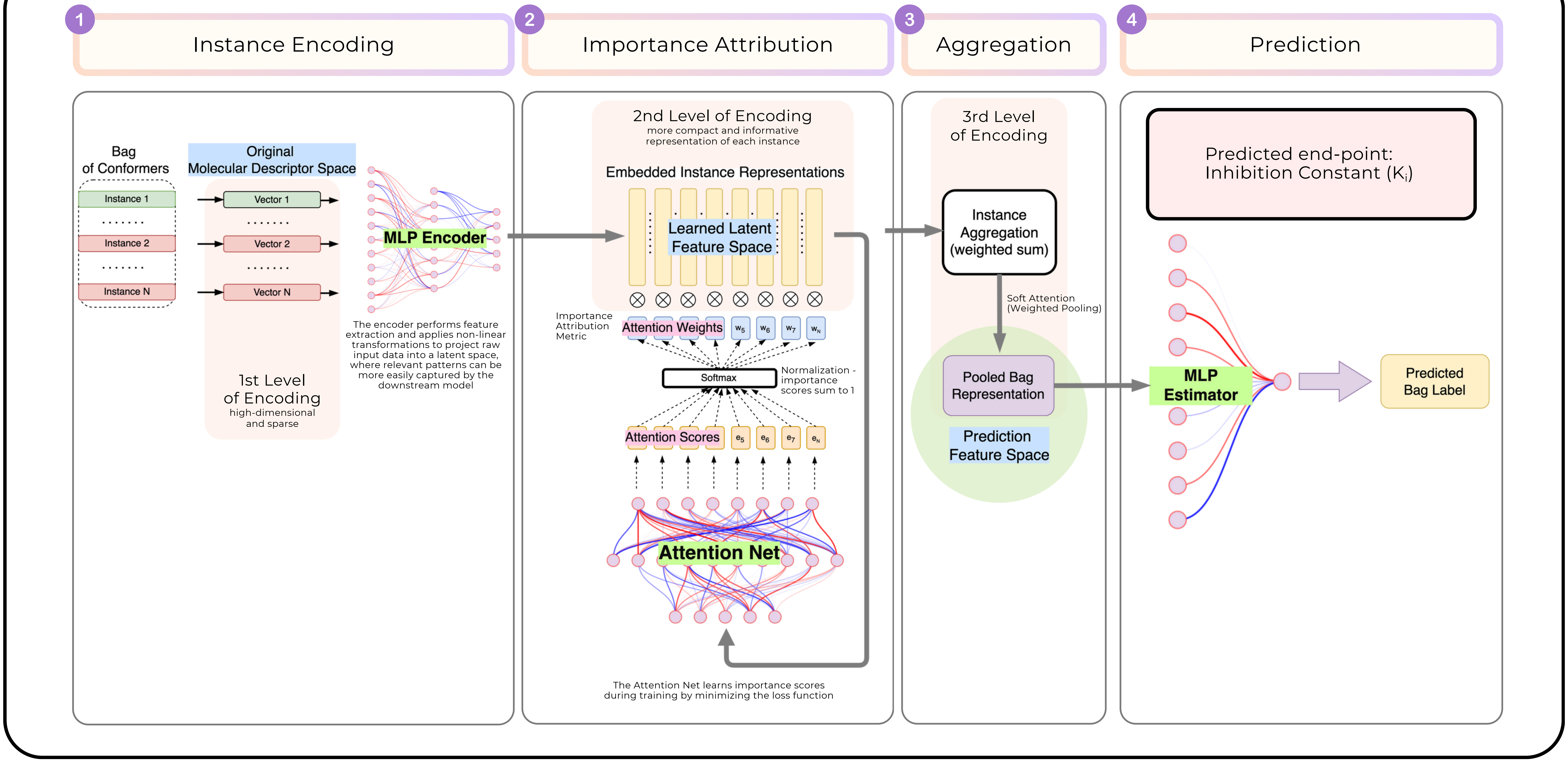
1 Instance Encoding (Embedding)

2 Aggregating per-instance embeddings to obtain a single bag-level representation

3 Single bag-level representation (for each Bag in the Train Set)

4 Conventional Single-Instance Classifier/Regressor Training

3. PIPELINE



1 Instance Encoding

2 Importance Attribution

3 Aggregation

4 Prediction

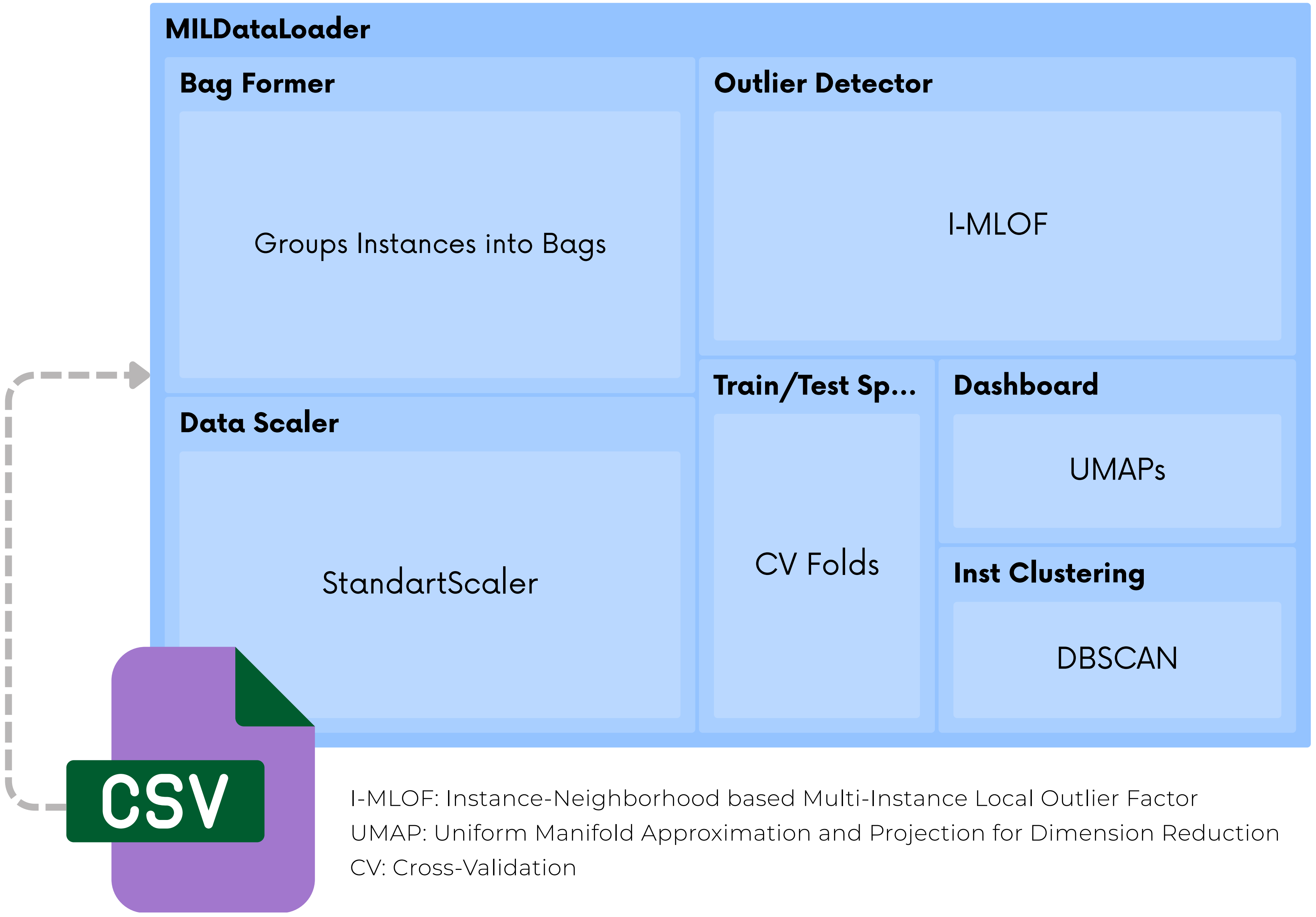
The encoder performs feature extraction and applies non-linear transformations to project raw input data into a latent space, where relevant patterns can be more easily captured by the downstream model

The Attention Net learns importance scores during training by minimizing the loss function

Zankov D. et al. *J Chem Inf Model.* **2021**, 61, 10, 4913-4923.

MIL-Based QSAR Pipeline

1. Data Loading and Preprocessing


MILDataLoader

Bag Former

Groups Instances into Bags

Outlier Detector

I-MLOF

Data Scaler

StandartScaler

Train/Test Sp...

CV Folds

Dashboard

UMAPs

Inst Clustering

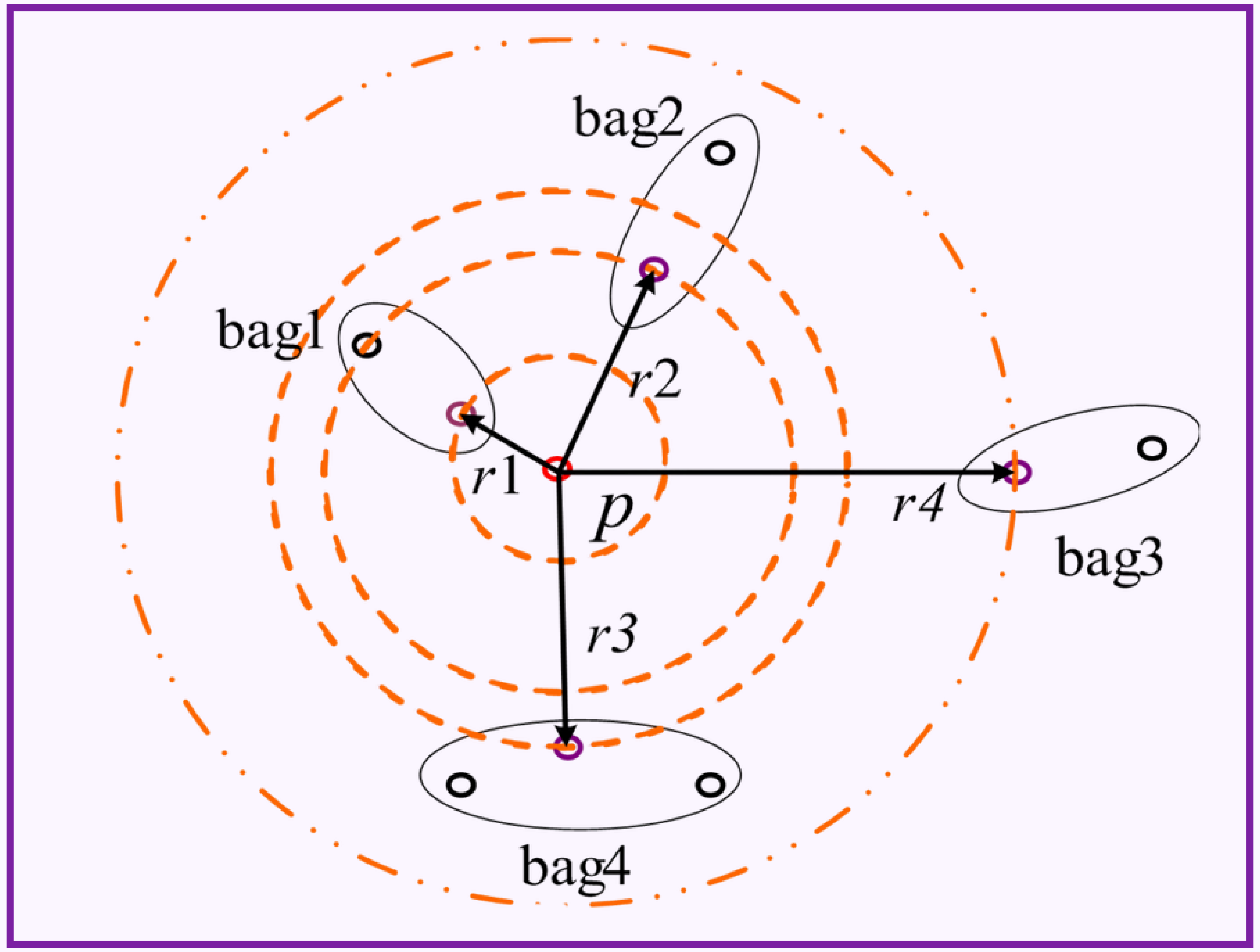
DBSCAN

CSV

Descriptors

I-MLOF: Instance-Neighborhood based Multi-Instance Local Outlier Factor
UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction
CV: Cross-Validation

Instance-Neighborhood Based Multi-Instance Local Outlier Factor (I-MLOF)

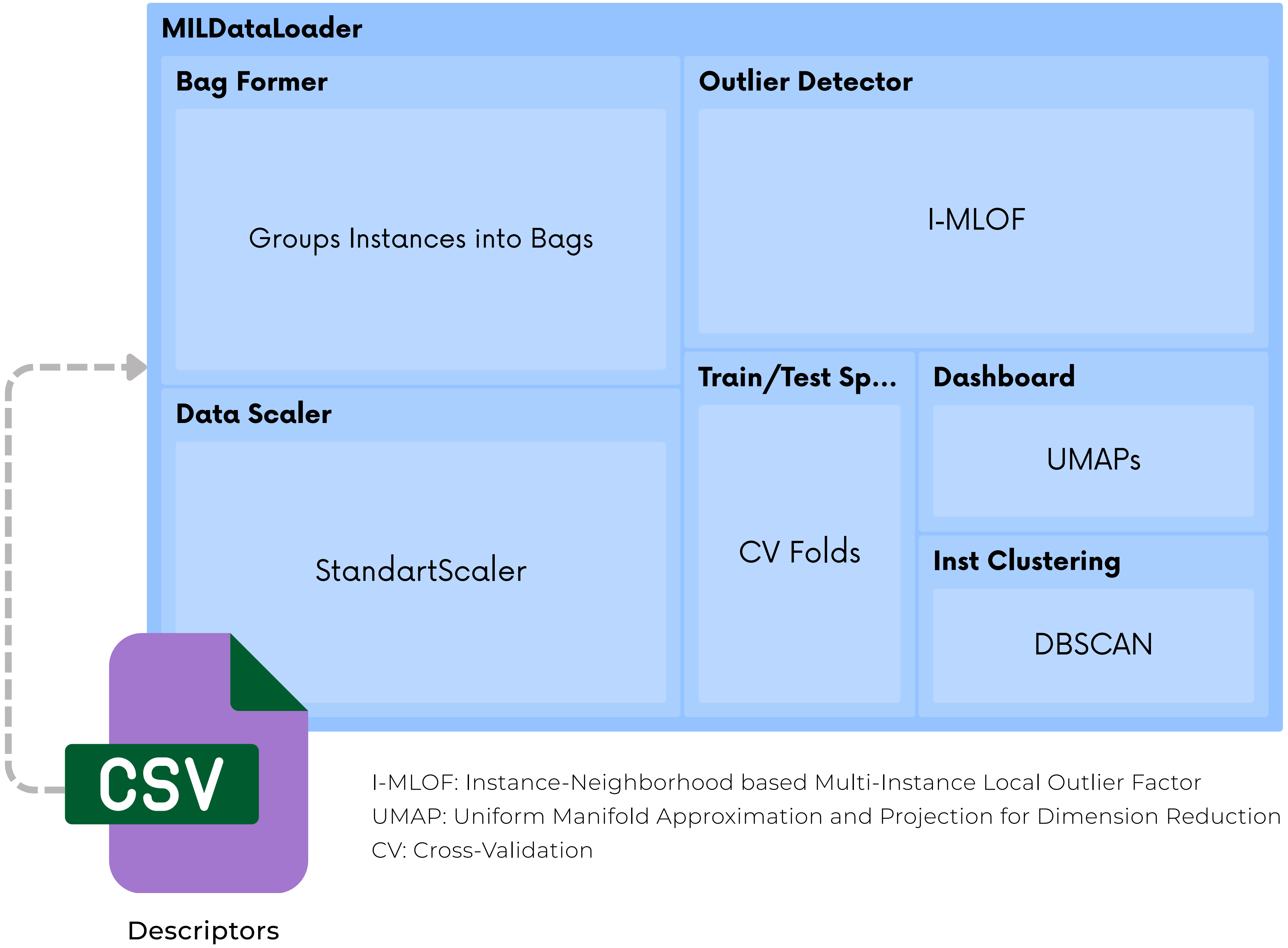


$$I{-}MLOF_{MinPts}(B) = \frac{\sum\limits_{C\in N_{MinPts}(B)} \frac{lrd_{MinPts}(C)}{lrd_{MinPts}(B)}}{|N_{MinPts}(B)|}$$

Wu, O., Li, B., Hu, W., Gao, J., & Zhu, M. (2010). Identifying Multi-instance Outliers. 430–441.
https://doi.org/10.1137/1.9781611972801.38

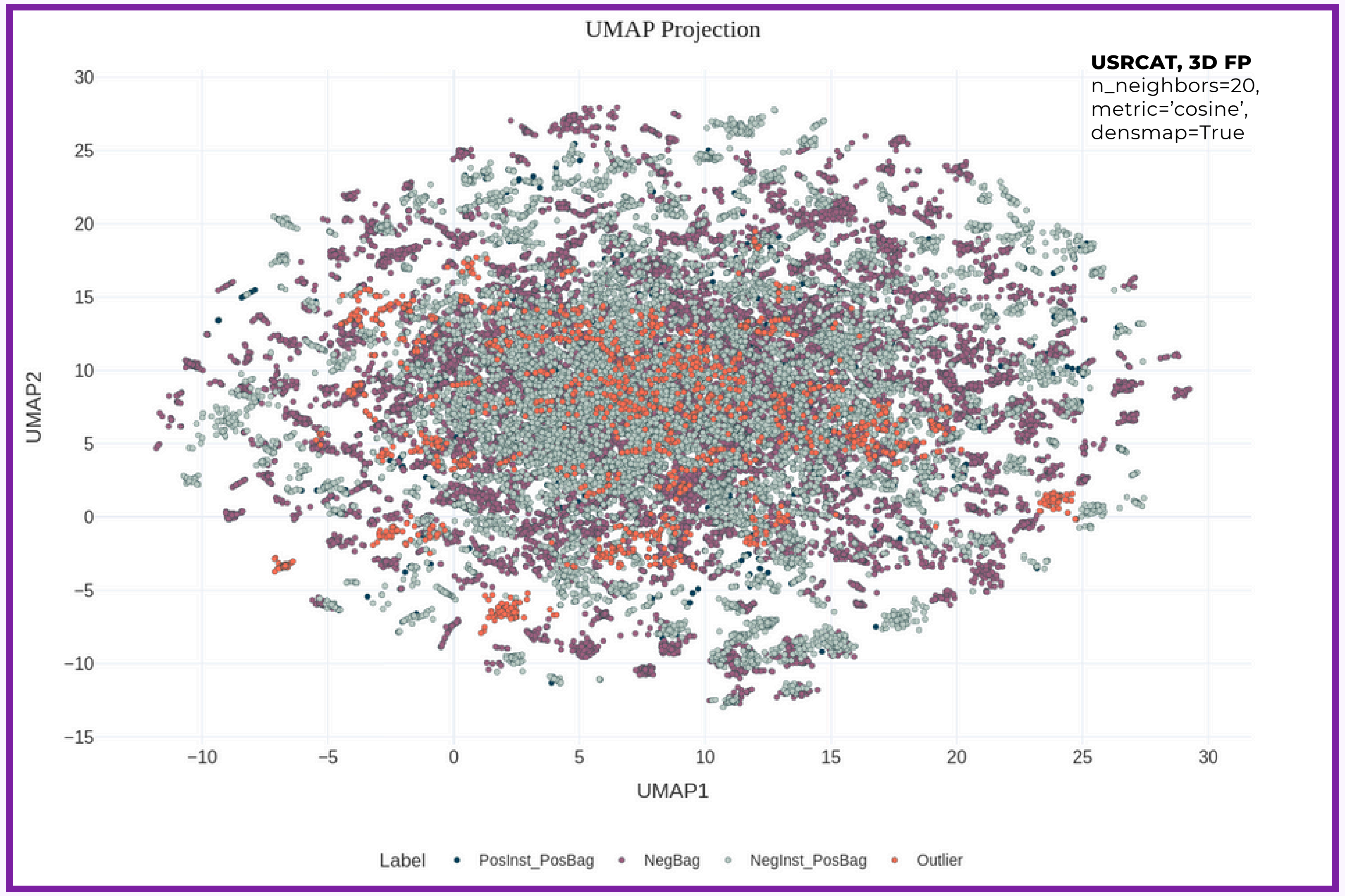
MIL-Based QSAR Pipeline

1. Data Loading and Preprocessing



I-MLOF: Instance-Neighborhood based Multi-Instance Local Outlier Factor
UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction
CV: Cross-Validation

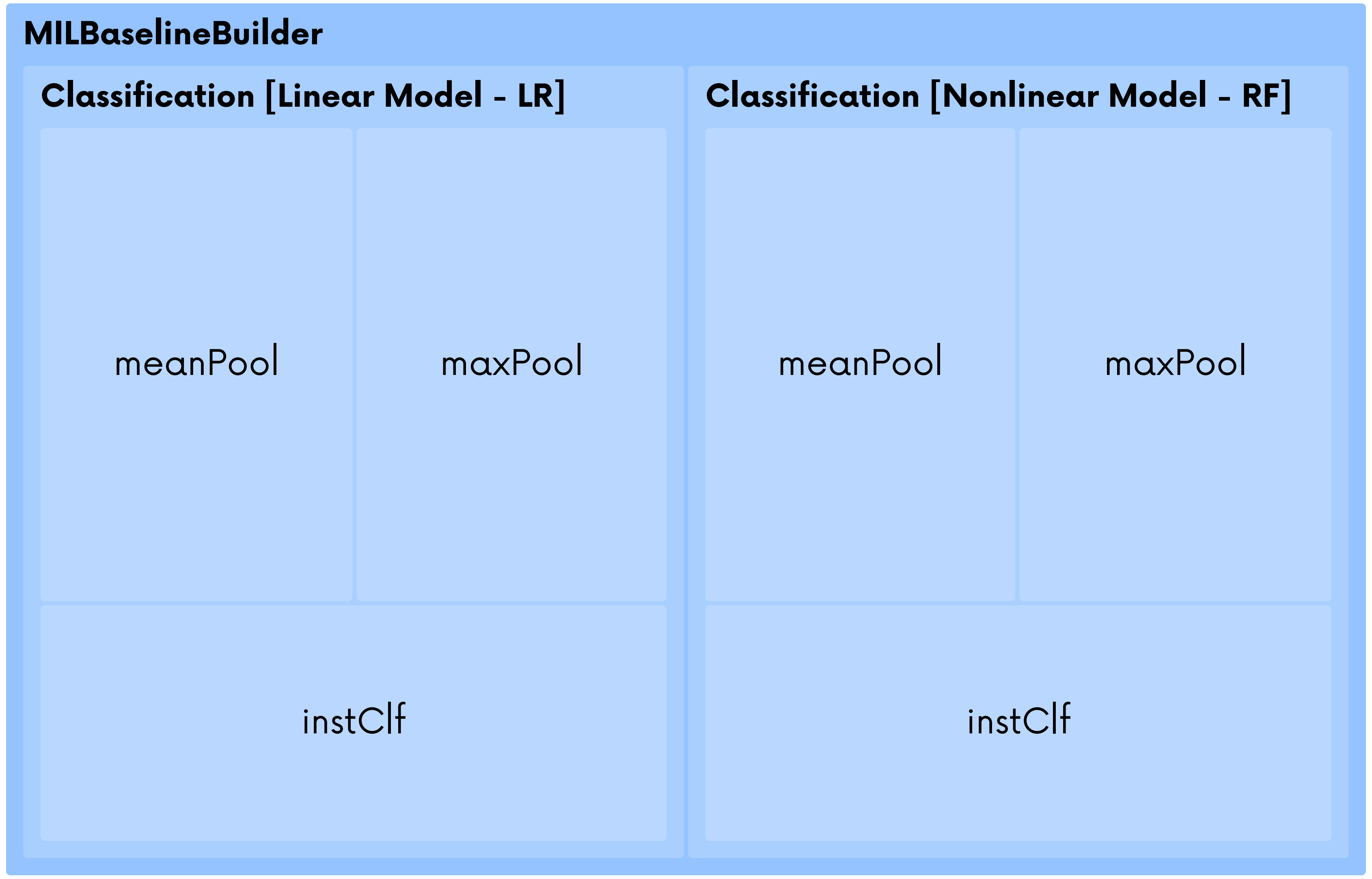
Uniform Manifold Approximation and Projection (UMAP)



An interactive Plotly Dash dashboard can be launched locally on your machine. The interface provides key dataset statistics and a visualization of the original descriptor space

MIL-Based QSAR Pipeline

2. Baseline Evaluation



MILBaselineBuilder

Classification [Linear Model - LR]

meanPool

maxPool

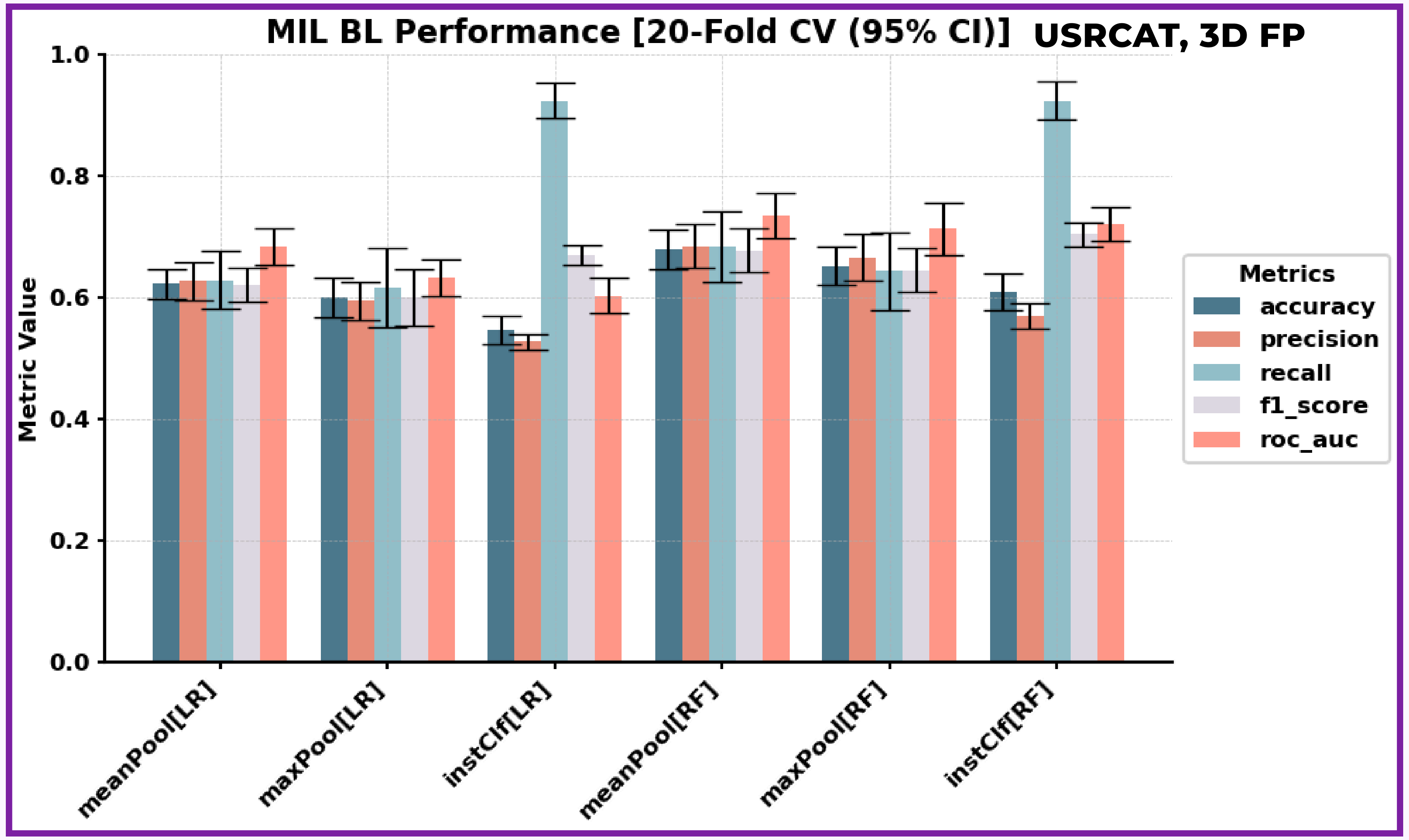
instClf

Classification [Nonlinear Model - RF]

meanPool

maxPool

instClf

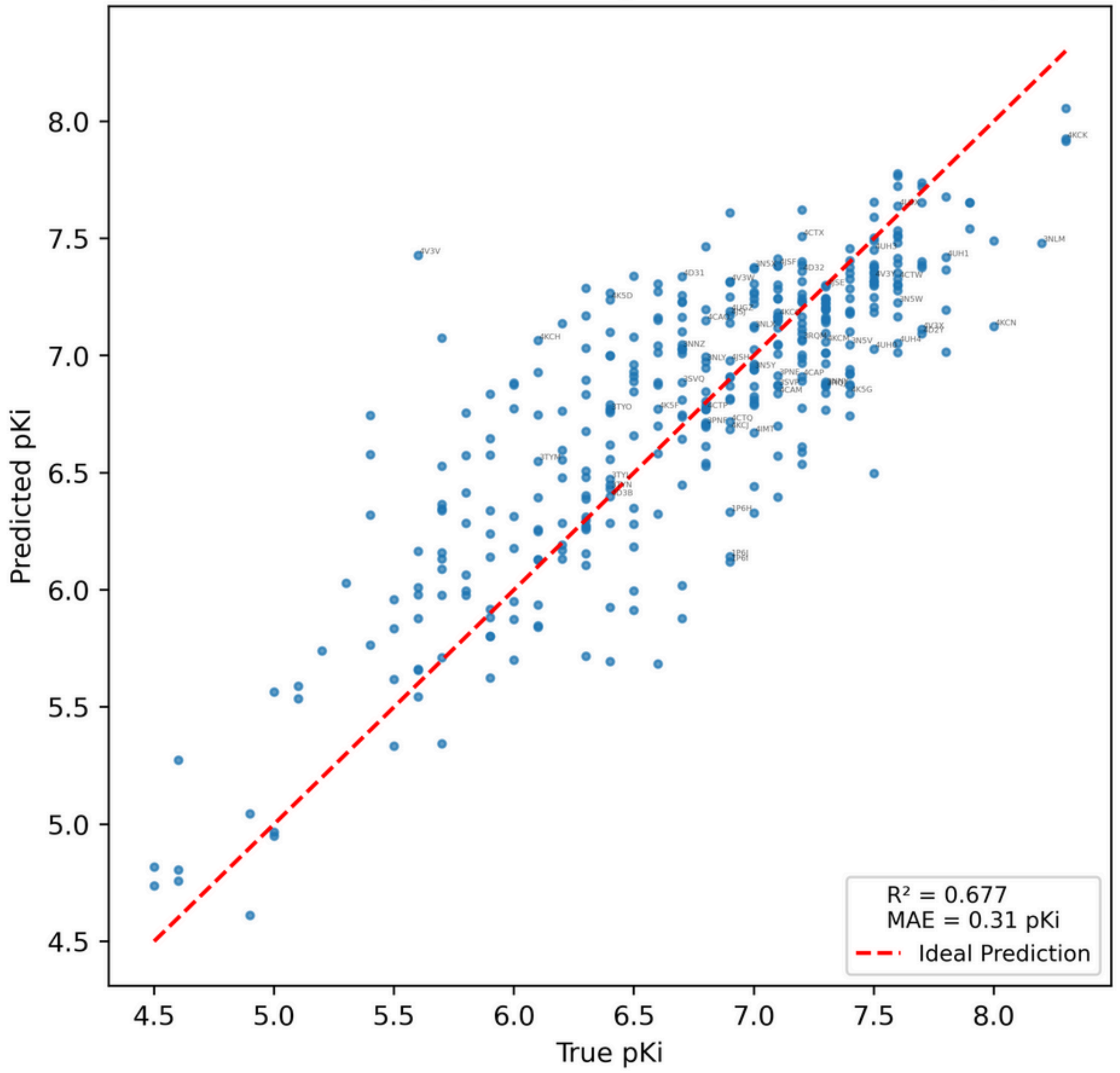


The MILBaselineBuilder Class provides several baseline MIL strategies, such as mean pooling, max pooling, or instance-level classifier with max aggregation [instClf], with Logistic Regression [LR] or Random Forest [RF]

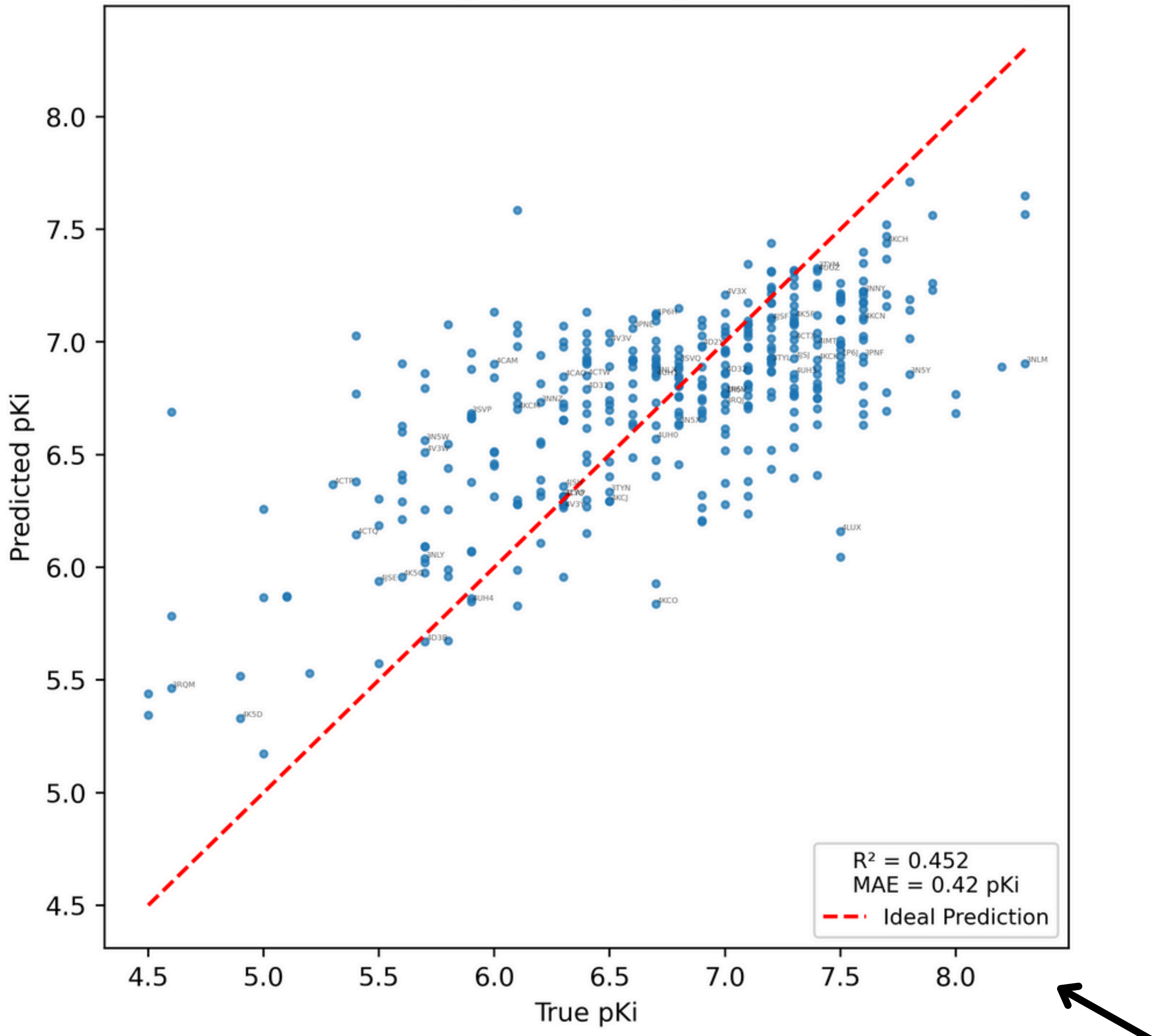
MIL-Based QSAR Pipeline: Baselines

Target: Nitric-oxide synthase, CHEMBL3048
Regression Task

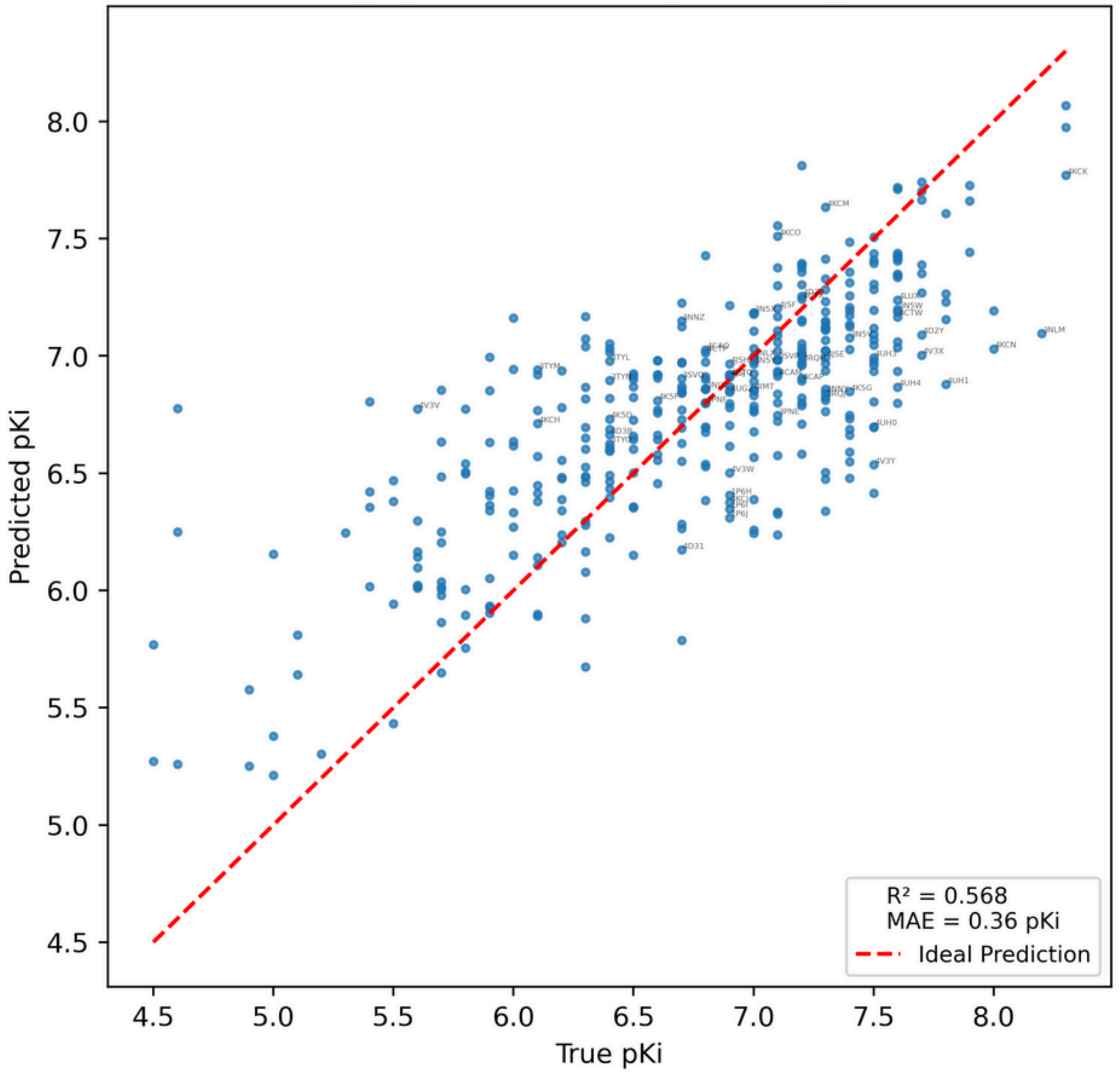
2D Model [Morgan FP, RFreg] — True vs Predicted pKi



3D SI Model [MORSEFingerprint, RFreg] — True vs Predicted pKi



3D MI Model [MORSEFingerprint, BS+Mean-pool, RFreg] — True vs Predic



MORSE FP demonstrated the strongest performance in the MI setting, but performed poorly compared to other FPs in the SI setting.

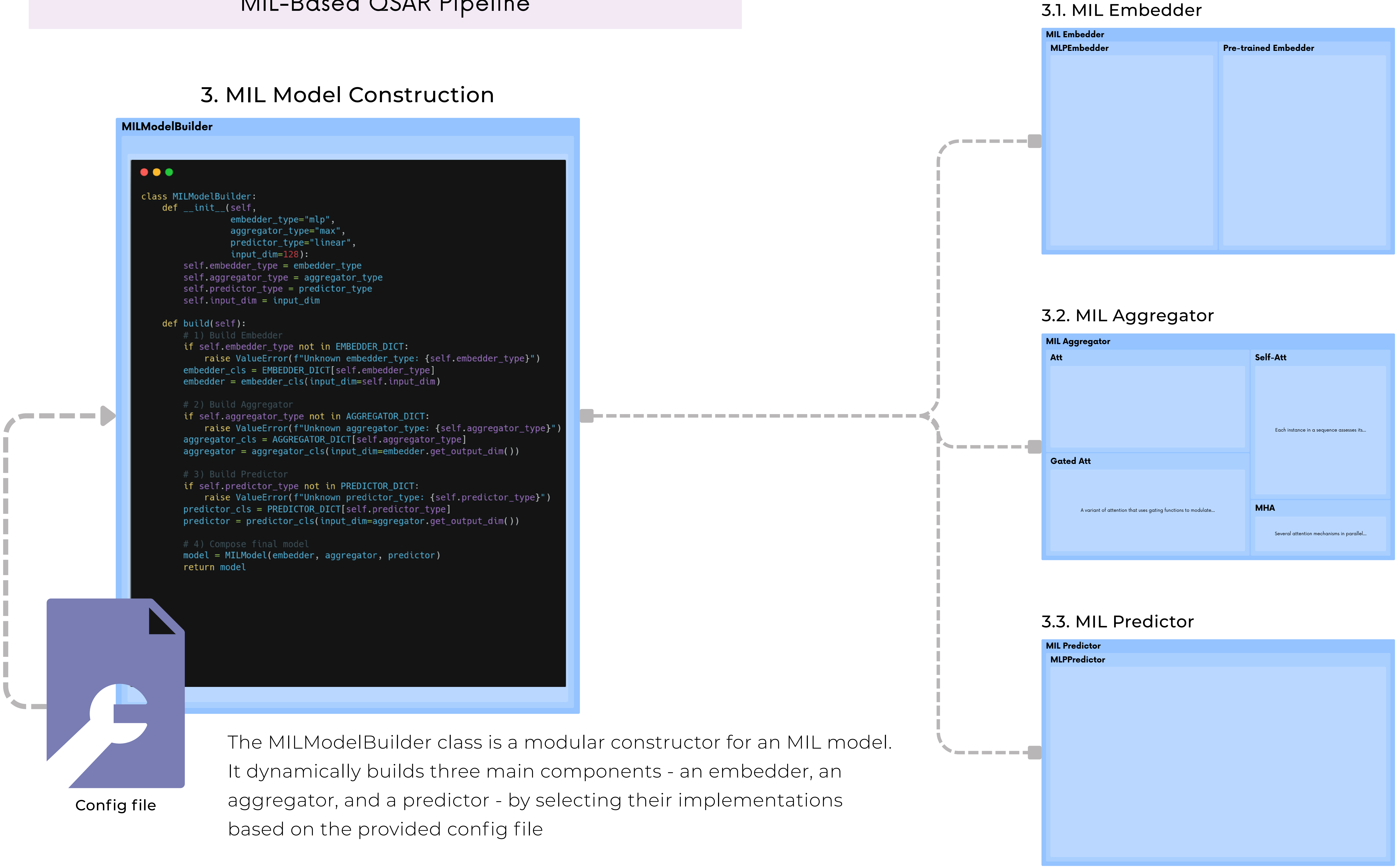
*20-Fold Cross-Validation

**3D Single-Instance (SI) Model was built on the minimum energy conformers (ETKDGv3+MMFF94)

***3D Multi-Instance (MI) Model was built on molecular conformer ensembles (50 conformers/mol, RMSD threshold 1Å, ETKDGv3+MMFF94)

MIL-Based QSAR Pipeline

3. MIL Model Construction

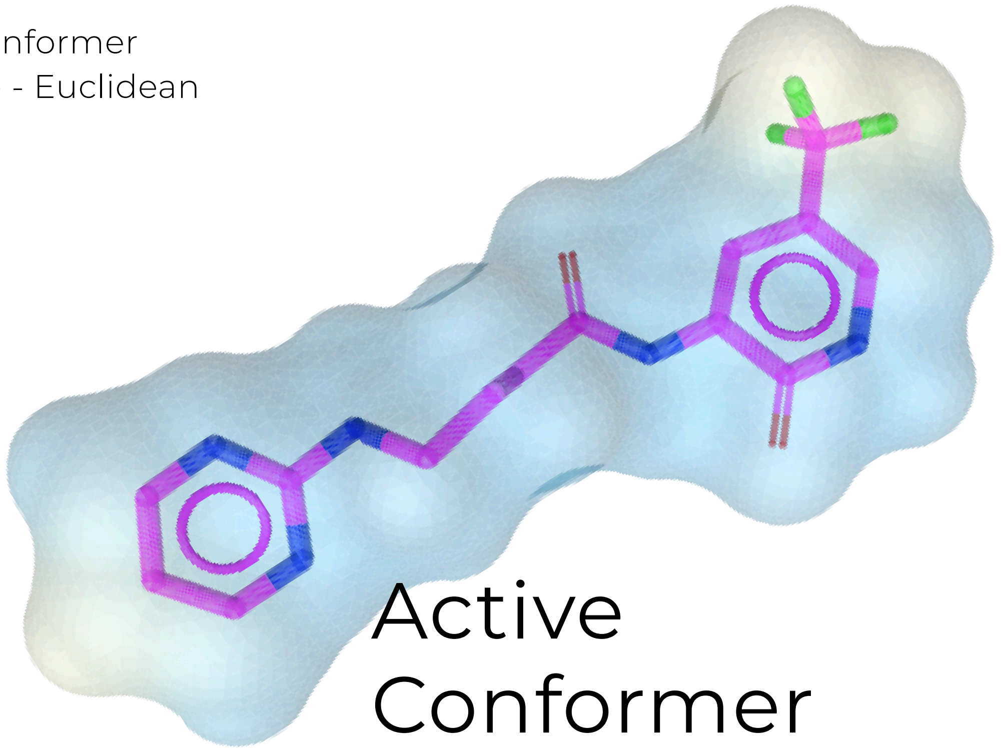


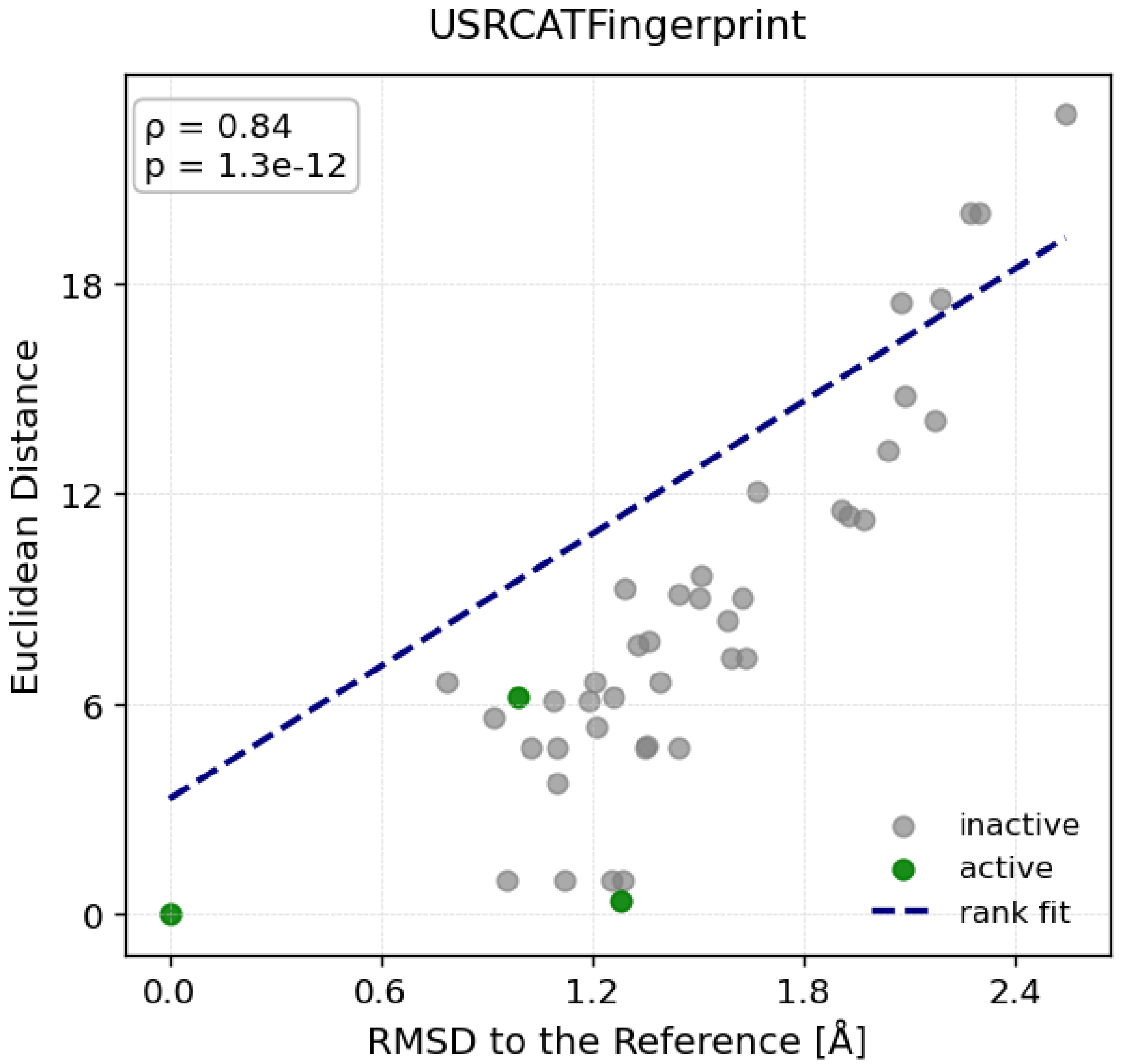
```python
class MILModelBuilder:
    def __init__(self,
                 embedder_type="mlp",
                 aggregator_type="max",
                 predictor_type="linear",
                 input_dim=128):
        self.embedder_type = embedder_type
        self.aggregator_type = aggregator_type
        self.predictor_type = predictor_type
        self.input_dim = input_dim

    def build(self):
        # 1) Build Embedder
        if self.embedder_type not in EMBEDDER_DICT:
            raise ValueError(f"Unknown embedder_type: {self.embedder_type}")
        embedder_cls = EMBEDDER_DICT[self.embedder_type]
        embedder = embedder_cls(input_dim=self.input_dim)

        # 2) Build Aggregator
        if self.aggregator_type not in AGGREGATOR_DICT:
            raise ValueError(f"Unknown aggregator_type: {self.aggregator_type}")
        aggregator_cls = AGGREGATOR_DICT[self.aggregator_type]
        aggregator = aggregator_cls(input_dim=embedder.get_output_dim())

        # 3) Build Predictor
        if self.predictor_type not in PREDICTOR_DICT:
            raise ValueError(f"Unknown predictor_type: {self.predictor_type}")
        predictor_cls = PREDICTOR_DICT[self.predictor_type]
        predictor = predictor_cls(input_dim=aggregator.get_output_dim())

        # 4) Compose final model
        model = MILModel(embedder, aggregator, predictor)
        return model
```

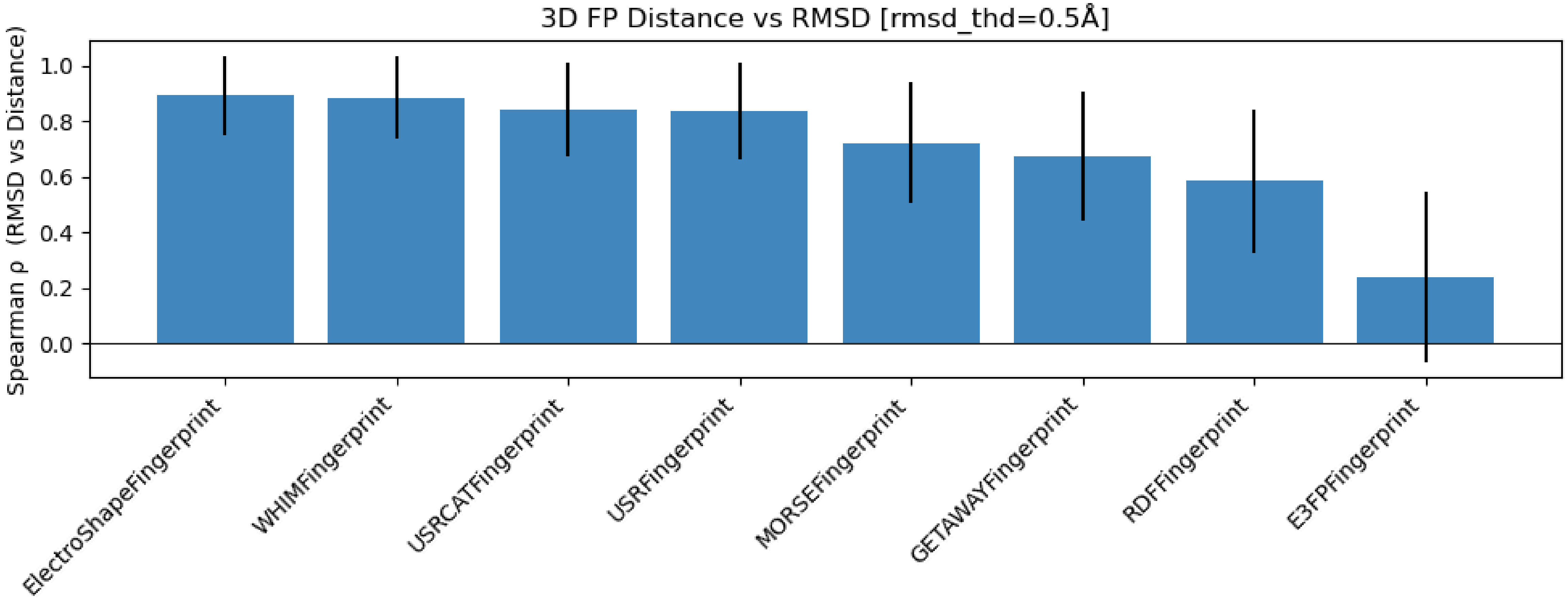
The MILModelBuilder class is a modular constructor for an MIL model. It dynamically builds three main components - an embedder, an aggregator, and a predictor - by selecting their implementations based on the provided config file

MIL-Based QSAR Pipeline: Encoding Choice

*Reference - min energy active conformer
**Distance metric in the feature space - Euclidean

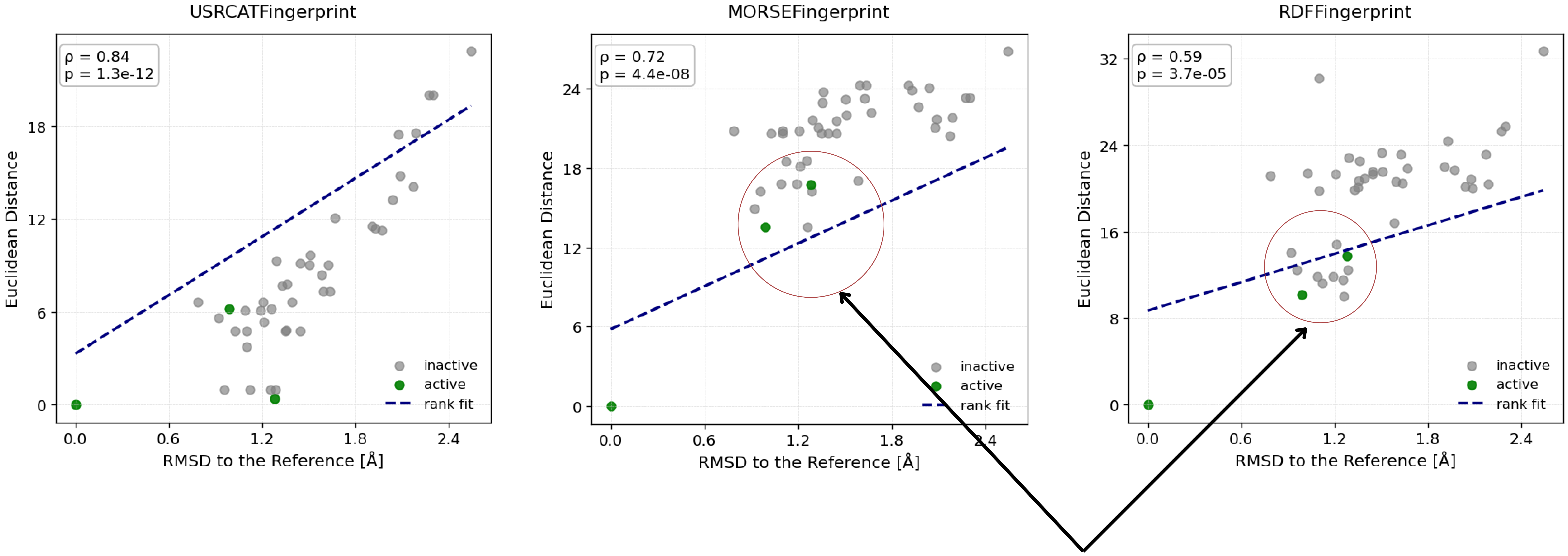






MIL-Based QSAR Pipeline: Encoding Choice

What requirements do we have for the encoding we utilize?



Clustering of bioactive conformers

Reference - min energy active conformer
*Distance metric in the feature space - Euclidean
**RMSD threshold between conformers in the Bag = 0.5 Å
***All plots were made for the same molecule

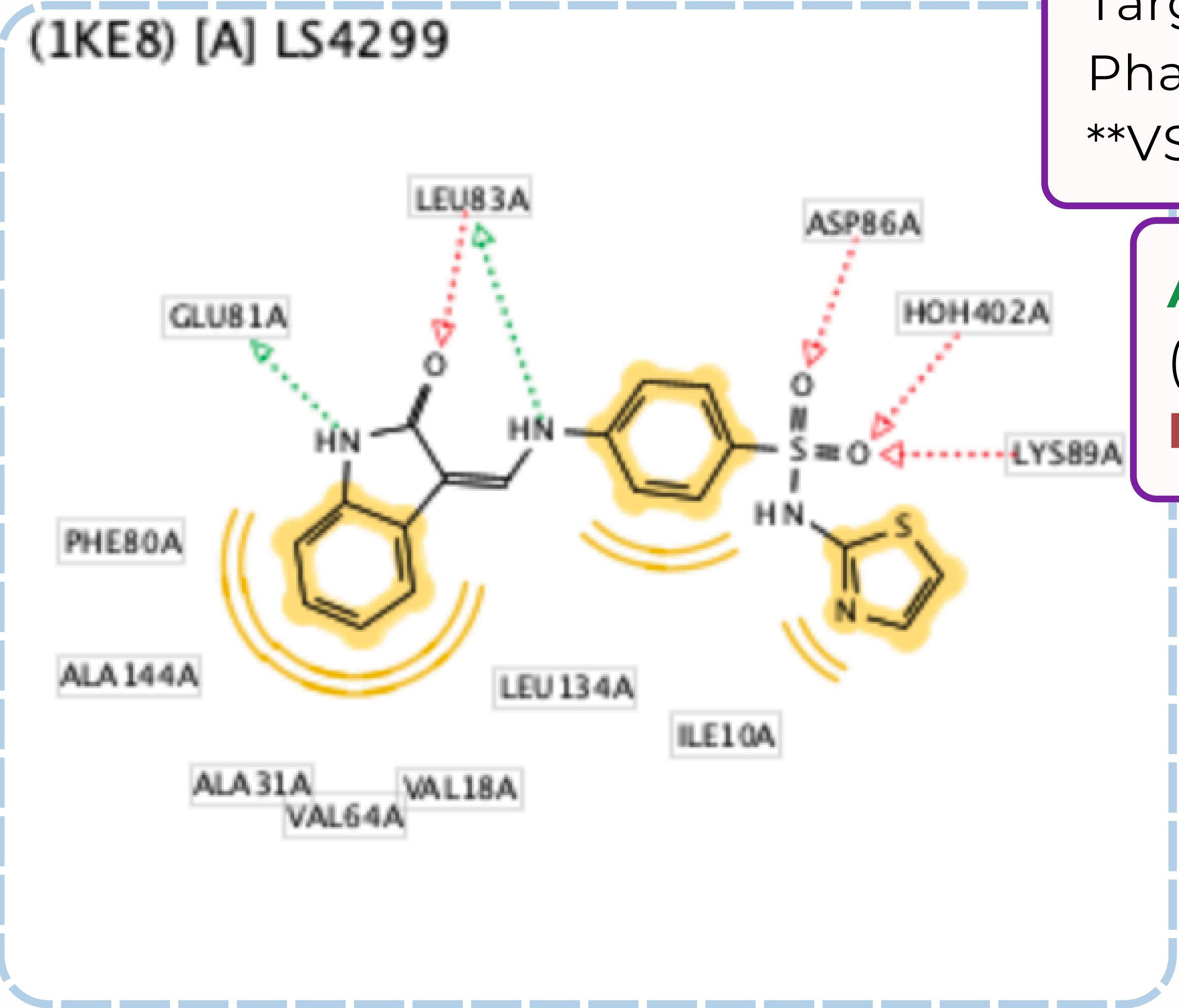
3. Results

Classification task,
target - CDK2 (1KE8)

Target: **CDK2 (1KE8)**

Pharmacophore: *7 features (3HBA, 2 HBD, 2Ar)

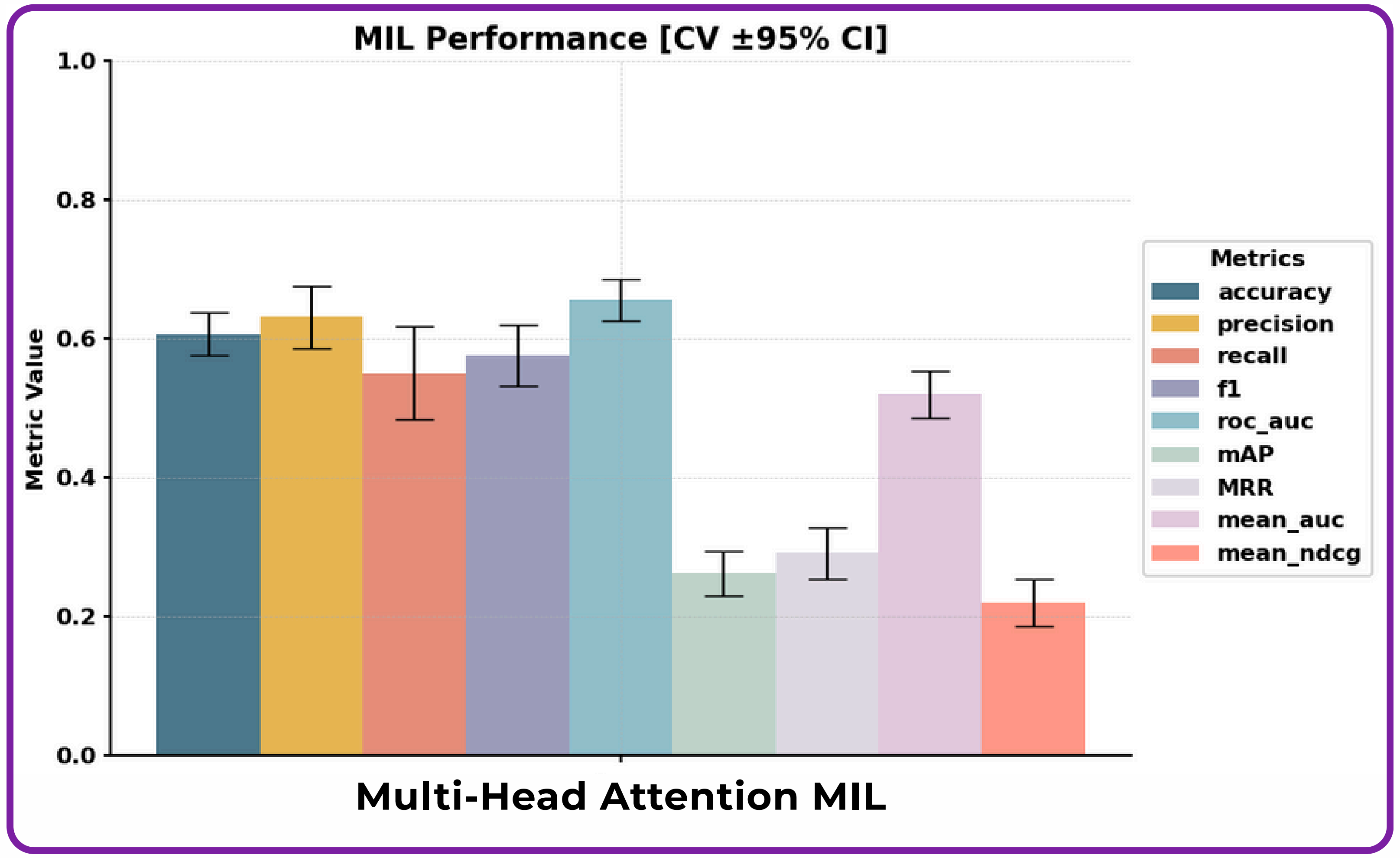
**VS: 399 actives, 399 decoys



2D Pharmacophore Model

Actives: Min. 1 Conformer matches a Pharmacophore (SMI)

Decoys: No conformers match the Pharmacophore



*7 Pharmacophore Features were considered in **Virtual Screening (VS) Procedure

3. Results

Regression task,
target - nitric oxide synthase, CHEMBL3048

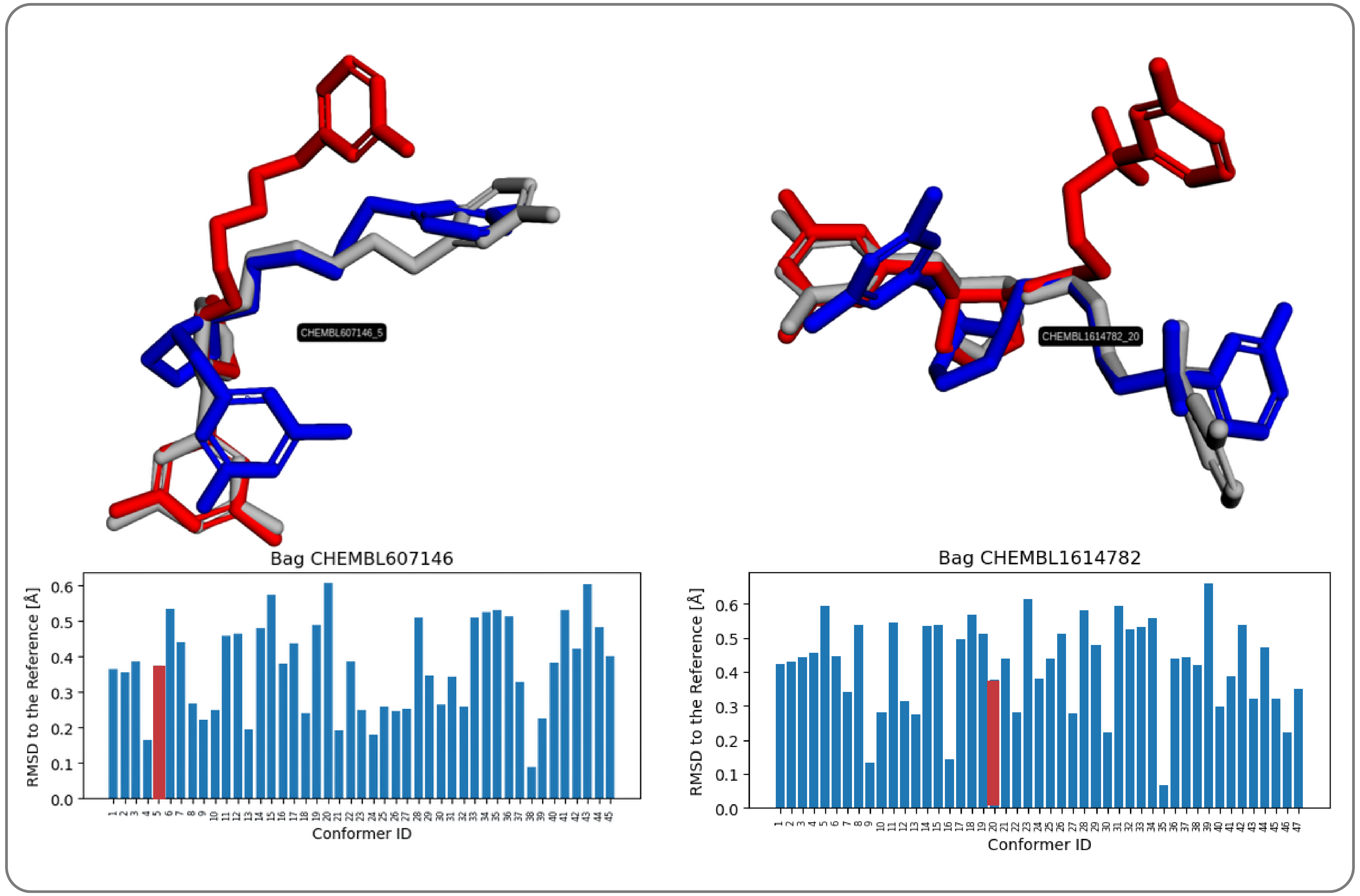


Figure 1. Key Instance Detection (KID). Top: Superposed conformers after RDKit GetO3A alignment: the lowest-energy conformer generated by RDKit ETKDGv3 (red), the experimentally observed conformer (grey), and the conformer identified as the key instance (blue). Bottom: Root-mean-square deviation (RMSD) values of all generated conformers to the experimental reference. The red bar highlights the conformer that received the highest attention weight.

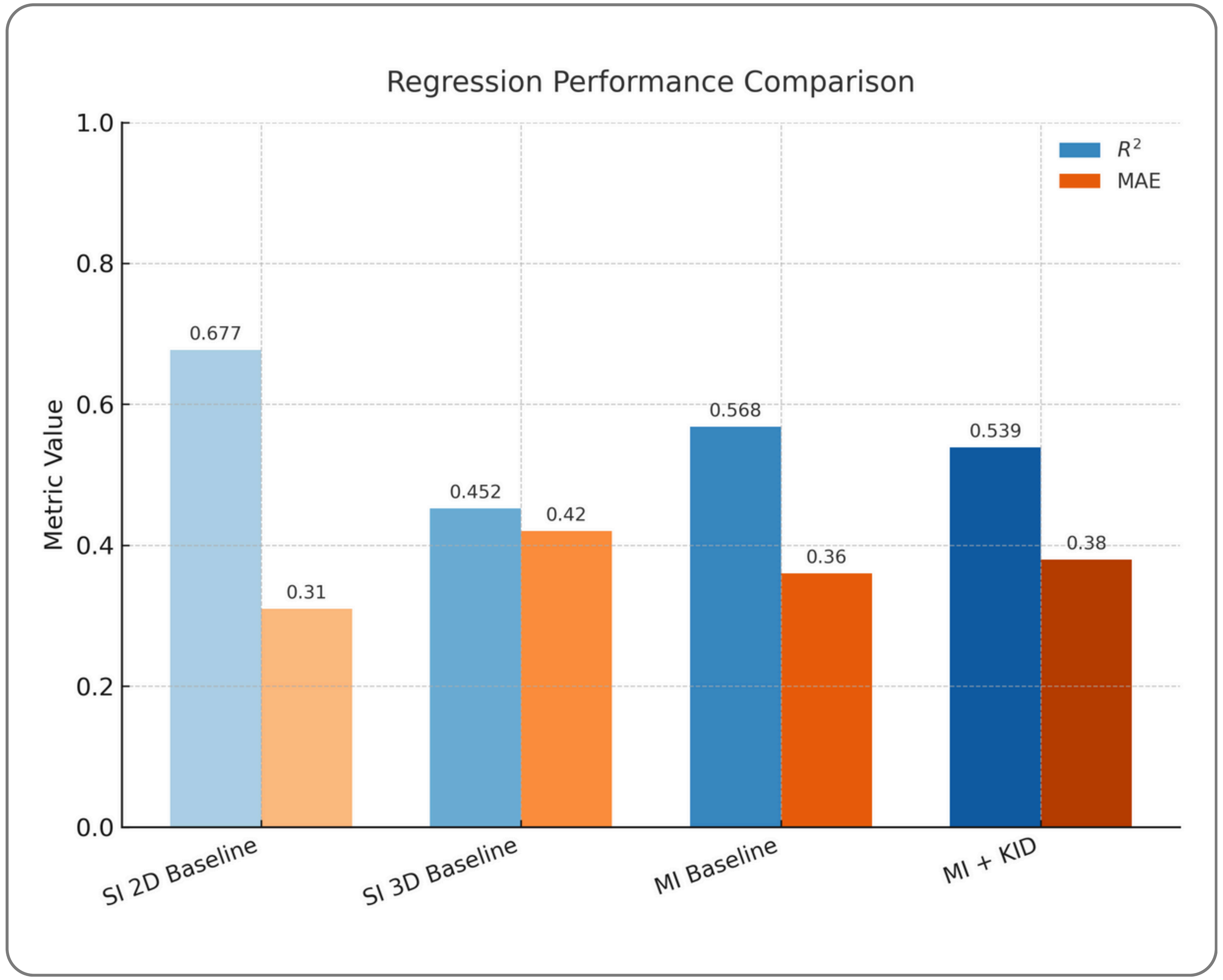
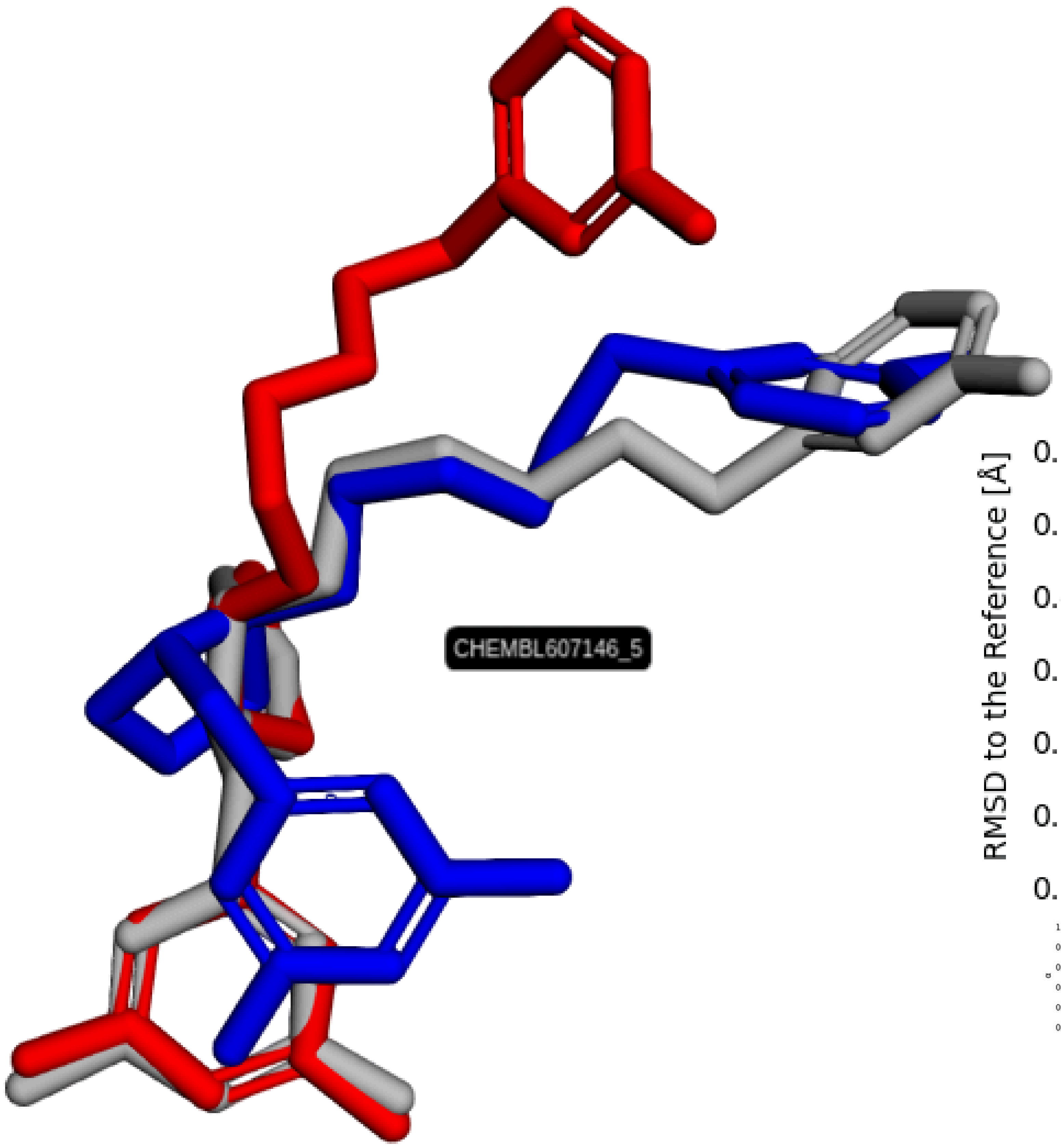
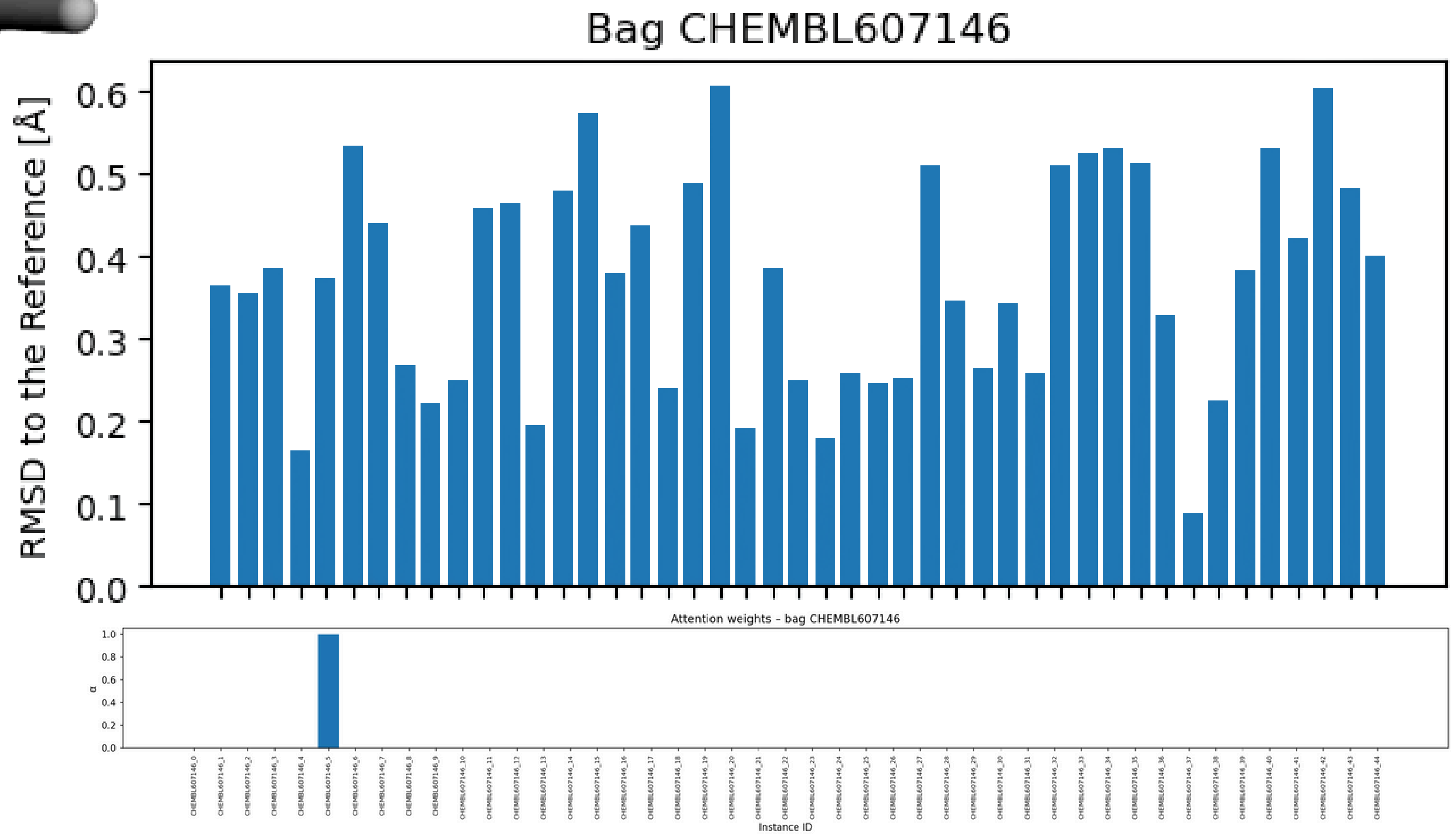


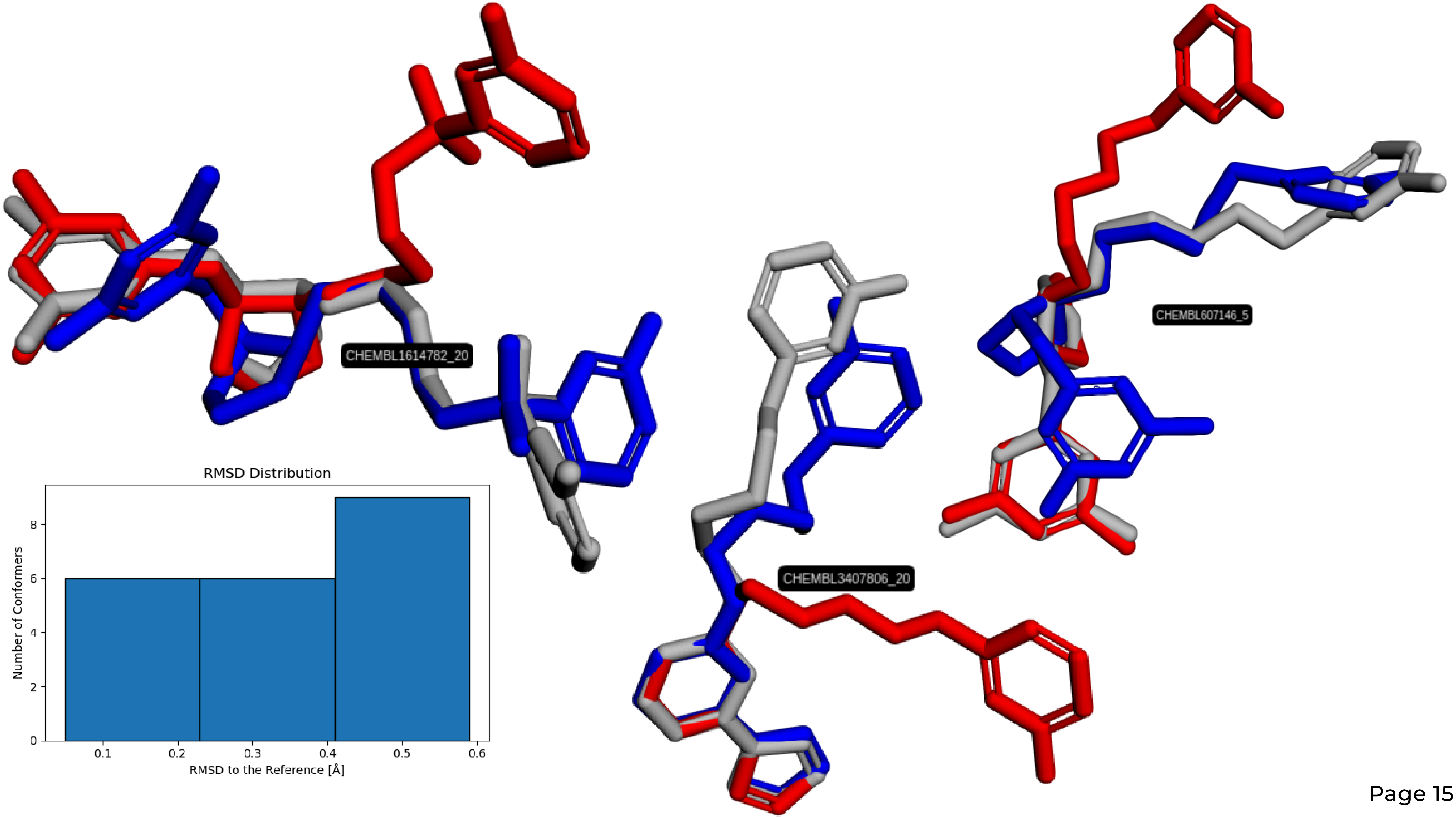
Figure 2. External validation accuracy. Regression task, target - nitric oxide synthase, CHEMBL3048: (1) SI-2D Baseline - SI RF regressor, 2D Morgan FP; (2) SI-3D Baseline - SI RF regressor trained on the minimum-energy conformers (one per molecule), 3D MoRSE FP; (3) MI Baseline - MI RF regressor; predictions obtained by mean-pooling 3D MoRSE descriptors across conformers whithin a bag; (4) MI + KID - developed method, described in the Pipeline section.

Abbreviations: RF, Random Forest; MAE, mean absolute error; MoRSE, Molecule Representation of Structures based on Electron diffraction





- Minimum-energy conformer (red)
- Experimentally observed conformer (grey)
- Predicted Key Instance (blue)


CHEMBL1614782_20
CHEMBL3407806_20
CHEMBL607146_5
RMSD Distribution
Number of Conformers
0
2
4
6
8
0.1
0.2
0.3
0.4
0.5
0.6
RMSD to the Reference [Å]

Conclusions & Outlook

- The model demonstrated the ability to prioritize conformers with lower RMSD to the reference bioactive structure, suggesting effective identification of relevant molecular shapes
- Validation of the model's attribution mechanism remains the primary goal of the project

[1] Dietterich, T. G., et al. Artificial Intelligence, 89(1–2), 31–71. (1997)

[2] Zankov, D. V., et al. Journal of Chemical Information and Modeling, 61(10), 4913–4923. (2021)

[3] Gomez, A., et al. arxiv.1706.03762. (2017)