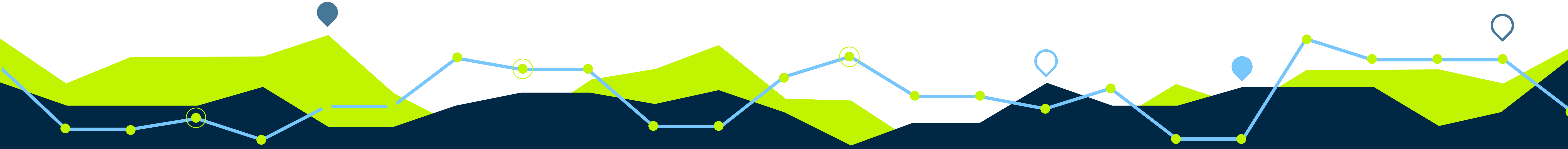
Alkermes

BREAK
THROUGH
TECH



# Forecasting Bioactivity: Predictive Models for Drug Discovery

# Meet the Team

Alkermes

## Joerg Bentzien

Director Computational Chemistry

Alkermes

## Polina Vanyukov

Assoc. Director, Enterprise Analytics

Alkermes

## Shinichiro Wachi

Principal Scientist Research

## Tiffney Aina

*Computer Science & Neuroscience*
Brown University

## Alex Lapadat

*Neuroscience*
Amherst College

## Blair Kuzniarek

*Neuroscience and Computer Science*
Northeastern University

## Ha Dong

*Neuroscience, Physics & Mathematics*
Amherst College

## Ray Qin

*Computer Science & Biology*
Smith College

# TABLE OF CONTENTS

01

## Objective

challenge summary, project tasks, intended impact

02

## Data Preprocessing & Analysis

data source, engineering, cleaning, and filtering

03

## ML Model & Results

model architecture, feature selection, optimization

04

## Conclusion

interpretations, what we learned, next steps

# 01

# Objectives

# OBJECTIVES

### CHALLENGE S MARY

Generate **predictiv** **ls** for m ioa dpoints t ting serotonin (5-HT) and dopamine (

### MAIN TECHNICAL

Use **molecular pr** **e** for drug discovery ar

HOW?

### REAL-LIFE IMPACT

**Streamline** the drug discovery pro ss by reducing costs, and improving decision-making in early-stages.

## Structural Analysis

Use different features (**SMILES & molecular fingerprints**) for similarity and property predictions.

## Machine Learning

**Random Forest** Regression with Python & Scikit-Learn to **predict bioactivity.**

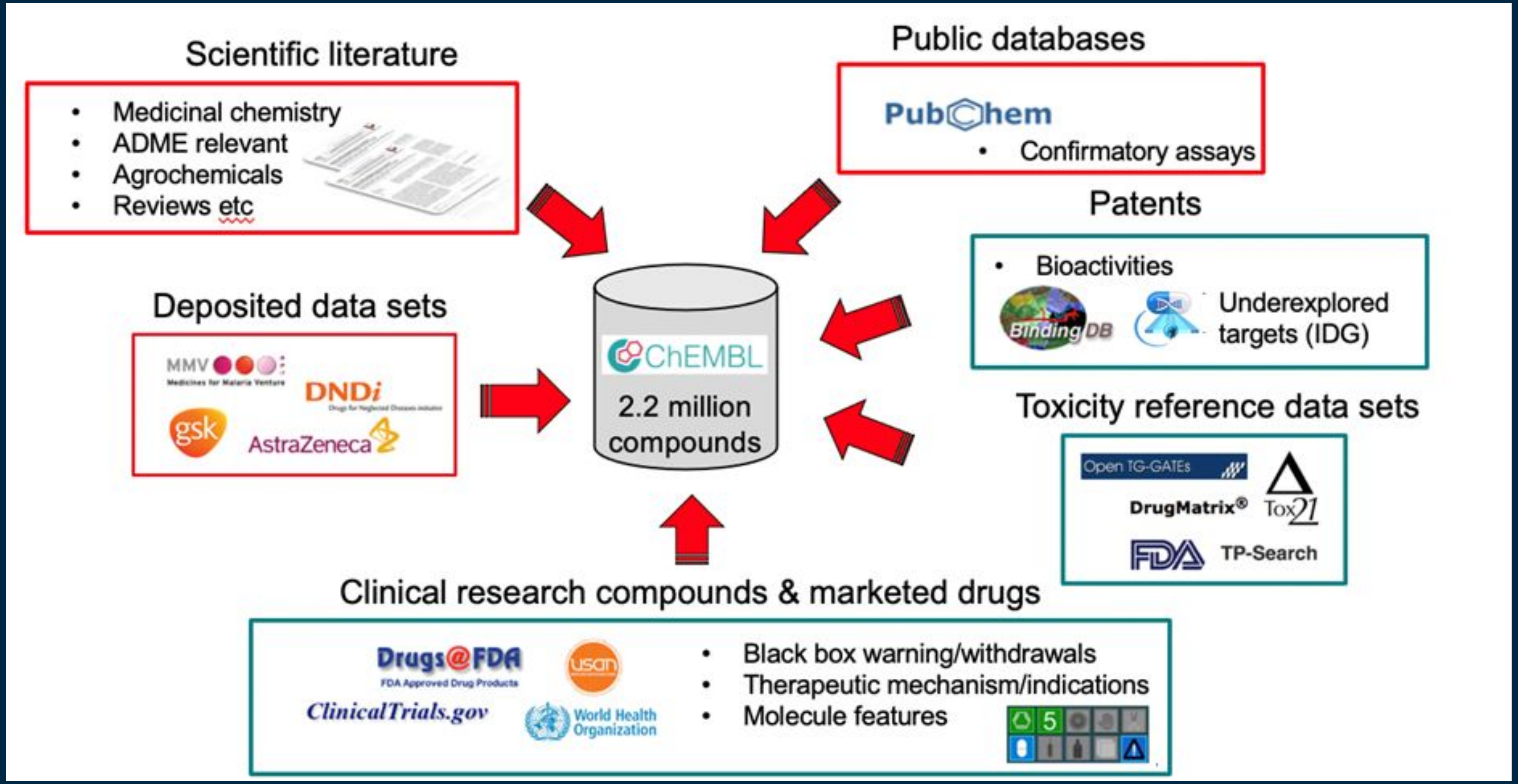
## Automation

Automate the code to make the tool **easier to deploy** for application to other datasets.

# 02

# Data Preprocessing & Analysis

# Data Source & Features



## SETTLED ON A SAMPLE DATABASE

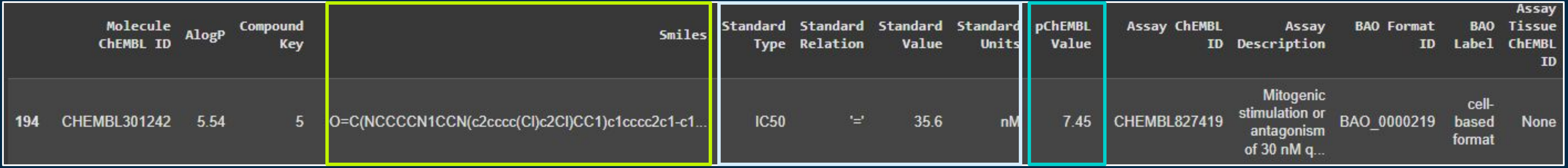
- 5 assays and 5 bioactivity tables
- **5-HT1a, 5-HT2a, 5-HT2b, 5-HT2c and D2**
- approx. 100.000 compounds, extracted from ChEMBL 34 Dataset.

**encodes the arrangement of atoms, bonds, and connectivity in a linear string**

**effectiveness of a compound in interacting with a target**

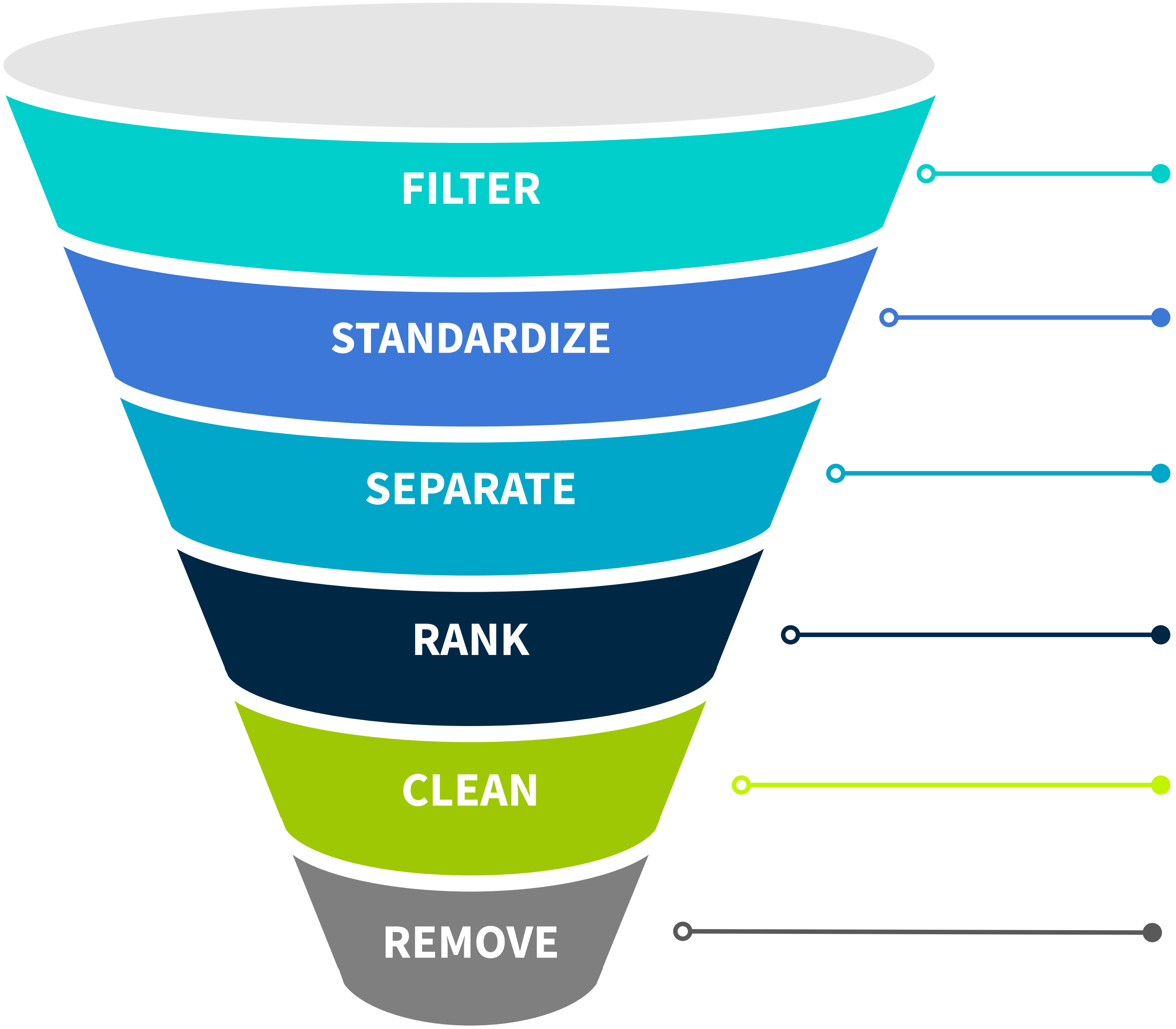
**log-transformed bioactivity (e.g., IC50, EC50), where higher values indicate stronger activity**

## RELATED COLUMNS IN TABLE



| | Molecule ChEMBL ID | AlogP | Compound Key | Smiles | Standard Type | Standard Relation | Standard Value | Standard Units | pChEMBL Value | Assay ChEMBL ID | Assay Description | BAO Format ID | BAO Label | Assay Tissue ChEMBL ID |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 194 | CHEMBL301242 | 5.54 | 5 | O=C(NCCCCN1CCN(c2cccc(Cl)c2Cl)CC1)c1cccc2c1-c1... | IC50 | '=' | 35.6 | nM | 7.45 | CHEMBL827419 | Mitogenic stimulation or antagonism of 30 nM q... | BAO_0000219 | cell-based format | None |

# DATA PREPROCESSING



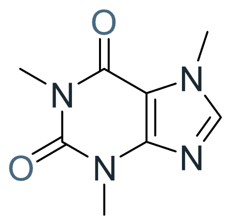
**FILTER**

Target Organism: '**Homo sapiens**'

Assay type: **‘Functional’ (“F”)**

**STANDARDIZE**

**Standardize** molecular structure into **SMILES**



O=C(N(C1=O)C)N(C2=C1N(C=N2)C)C

**SEPARATE**

**Separate** different **bioactivity** measurements (**IC50** -antagonists vs **EC50** -agonists)

**RANK**

Select the **Top 10 Assay** ChEMBL IDs by frequency

**CLEAN**

**Remove NaN** and **duplicate** values, keeping rows with the highest **pChEMBL** values

**REMOVE**

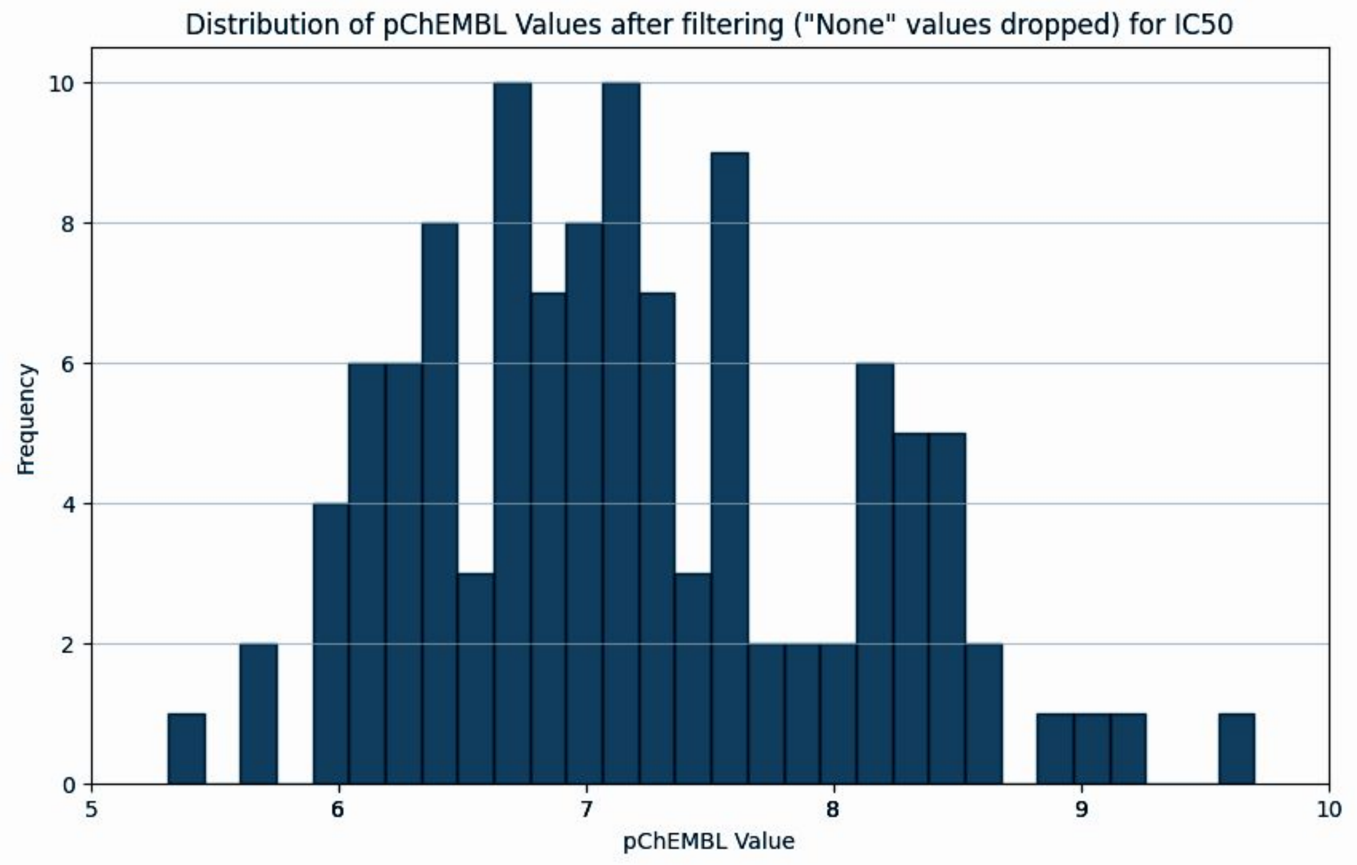
**Remove** outliers using **IQR Outlier Detection**

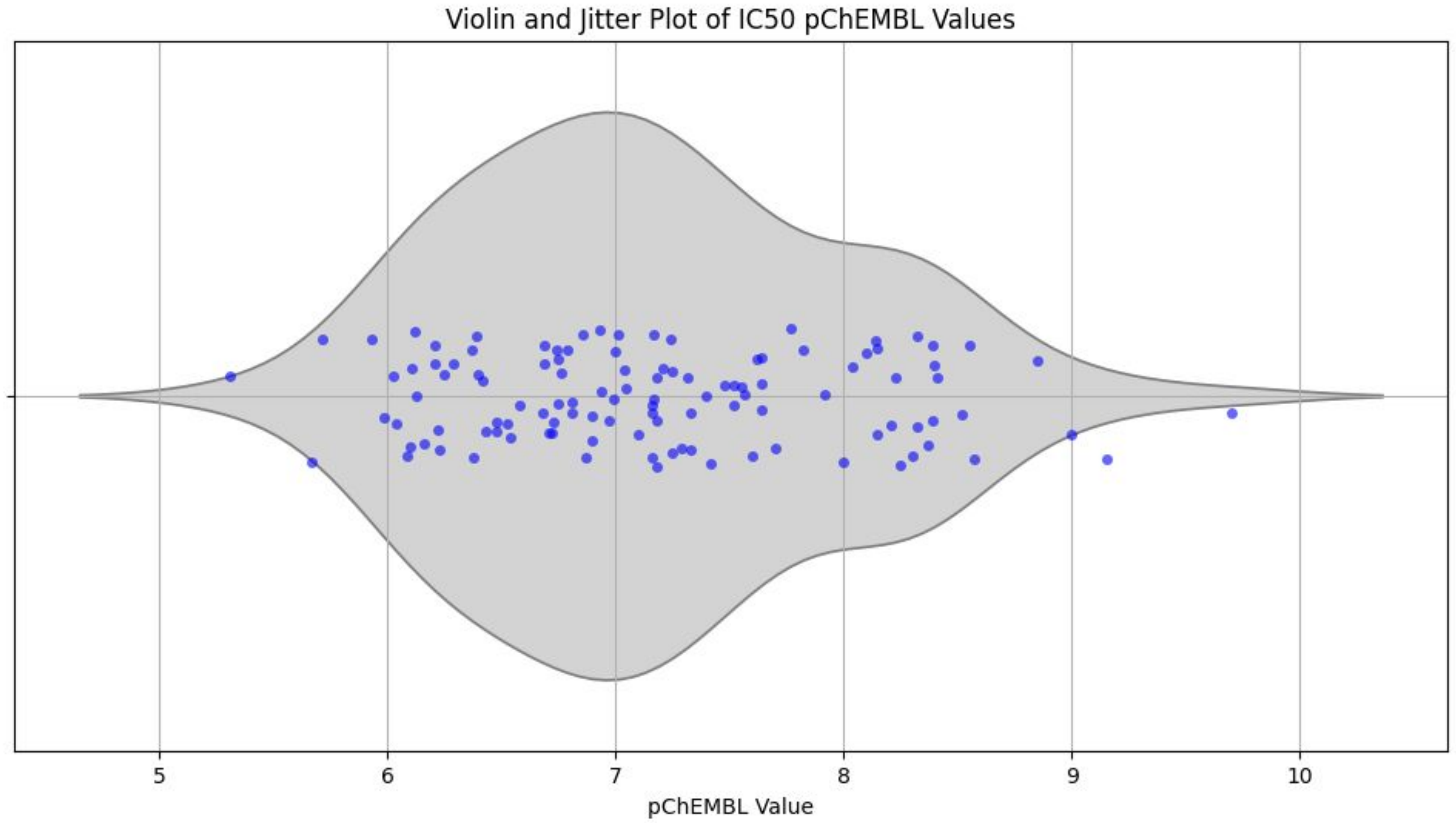
# DATA ANALYSIS

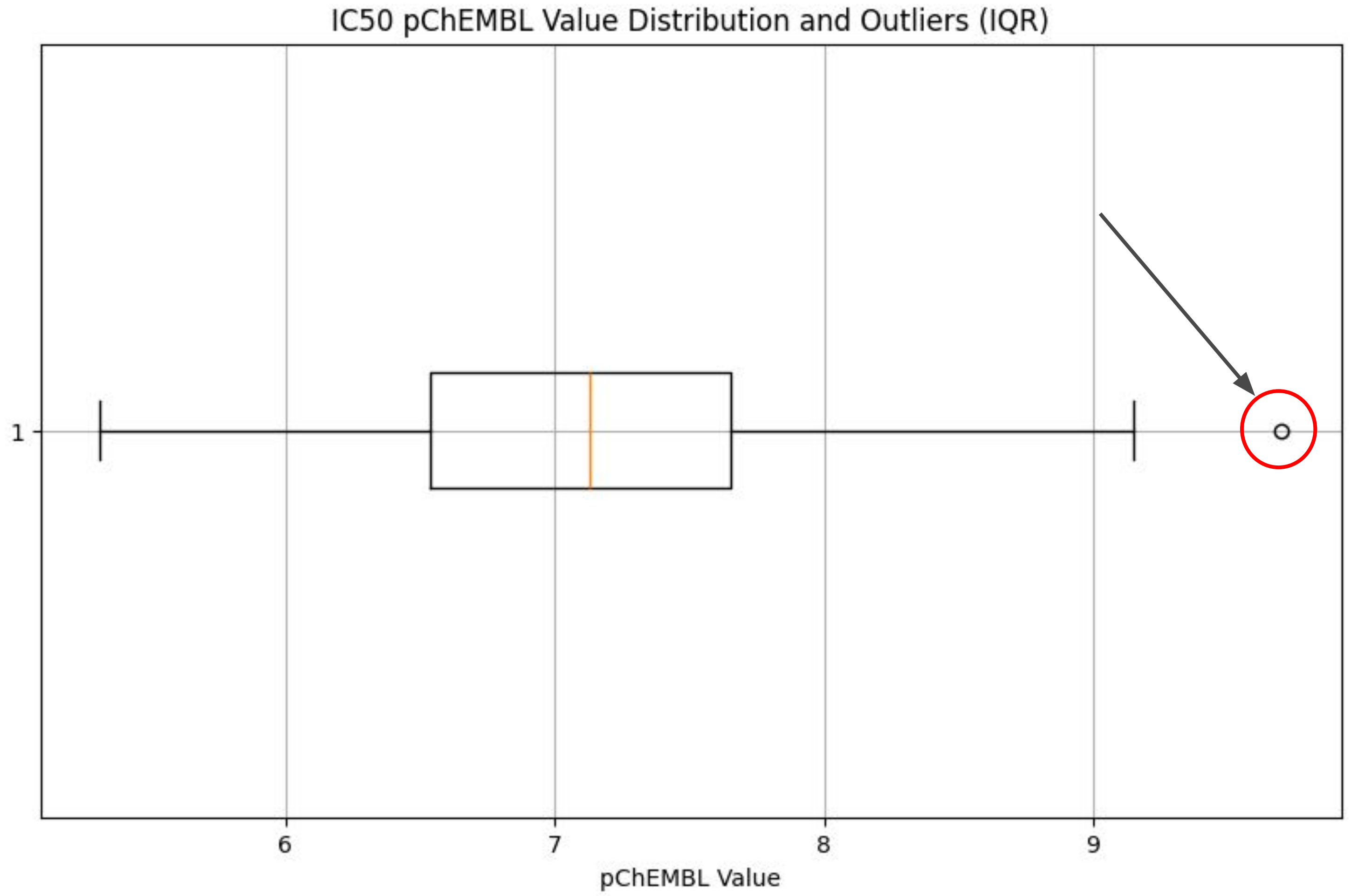
01 Visualize the distribution of pChEMBL values

02 Outlier Detection using IQR and Z-score

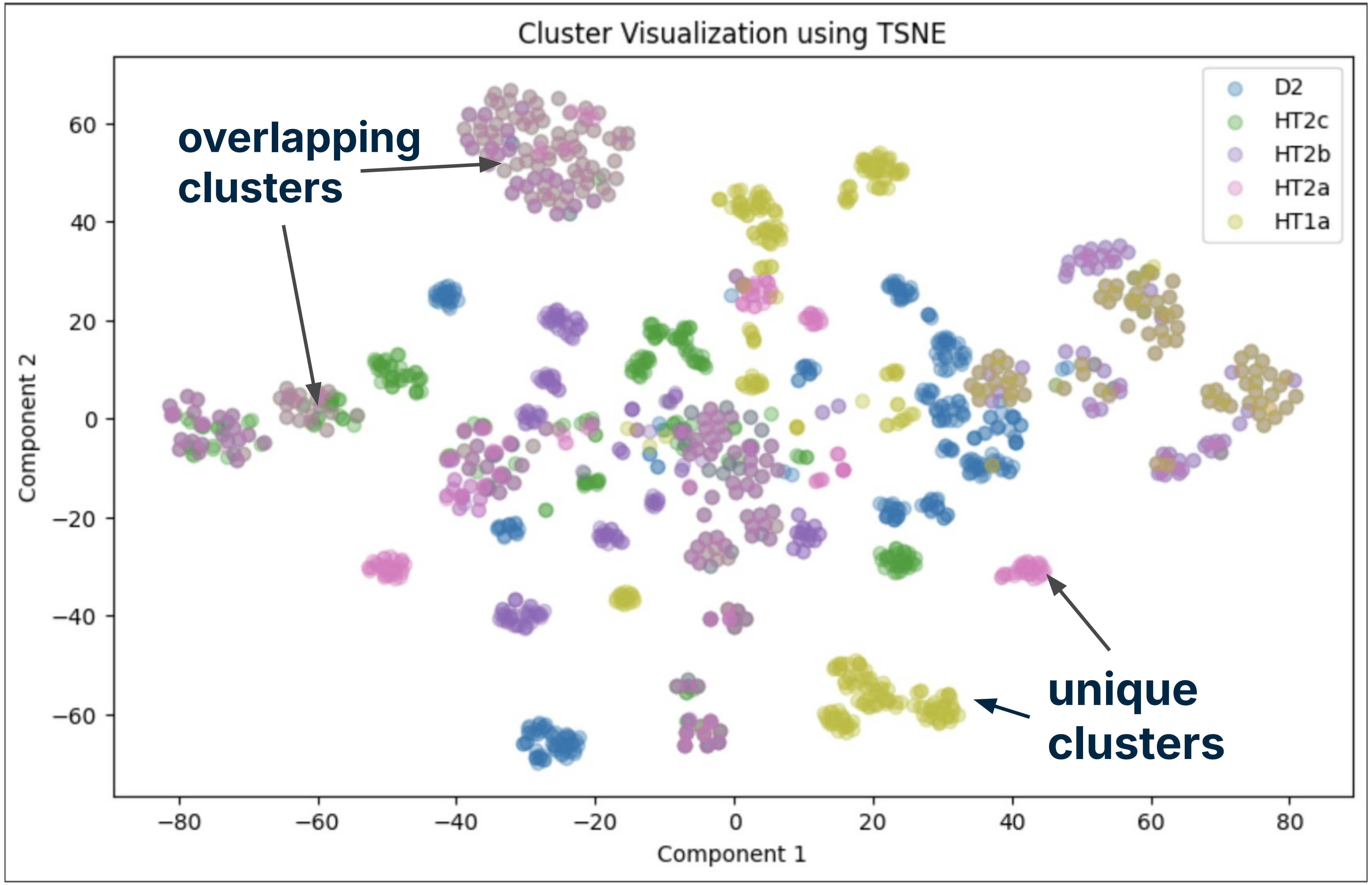
03 Key Value Identification (lowest and highest pChEMBL values)







# Tanimoto Similarity Across Endpoints ECFP6



*each color indicates a different dataset

***Do our datasets cover similar chemical space?***

**T-SNE** plot shows **molecular similarity of ECFP6 fingerprint** across five datasets, using Tanimoto Similarity.

The datasets have ***shared*** chemical space, with some ***unique*** clusters.

# 03

# Random Forest Model & Results

# Training – Baseline model

We trained a `RandomForestRegressor` on our five datasets using two sets of features:

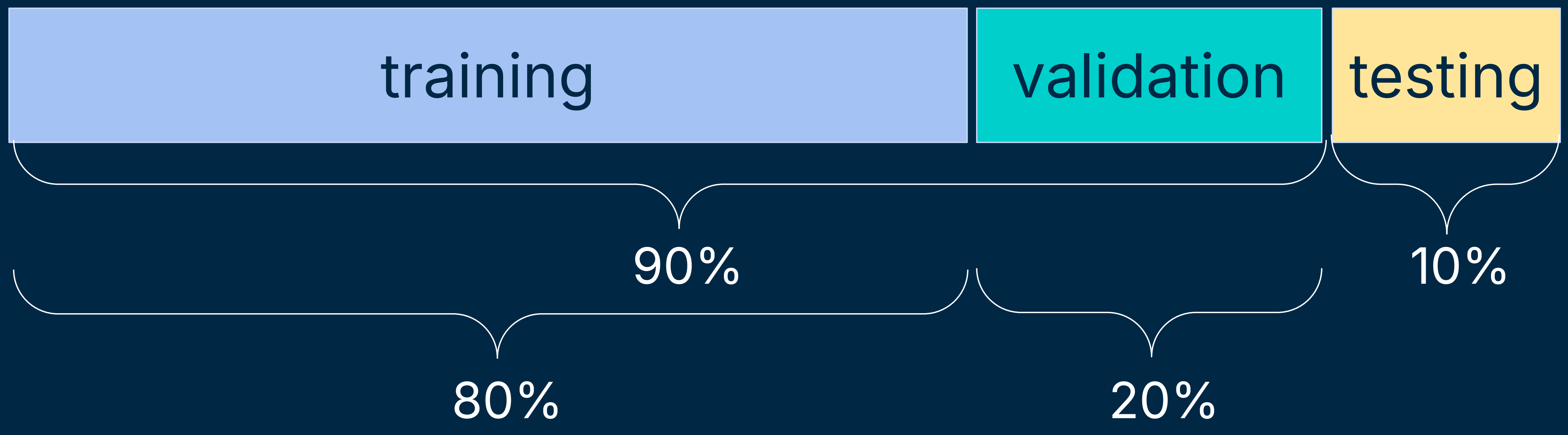
**Input**:

- ECFP6 Fingerprint
- 1613 2D Mordred Descriptors

**Output**: pChEMBL Value - Normalized Potency

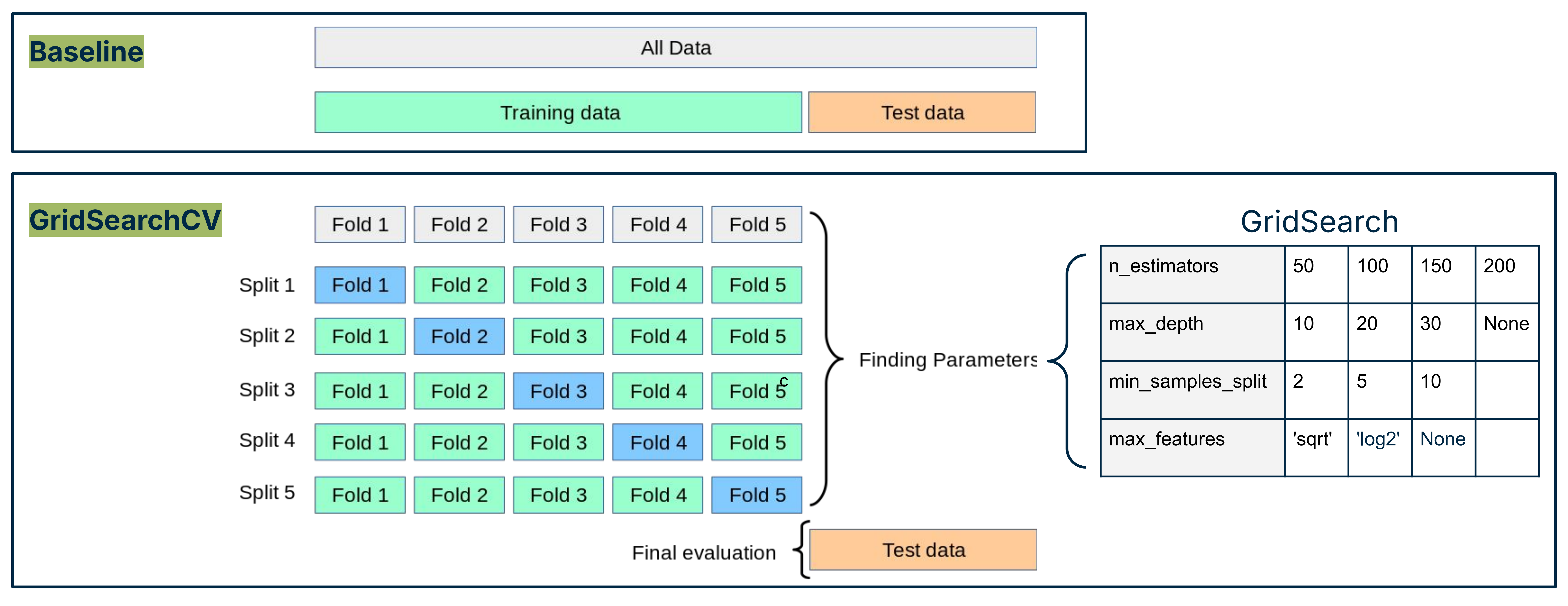
**Training parameters:**

```
- n_estimators=100
- train:test = 9:1
```



# Experiment 1: Model optimization

In additional to our baseline model (`n_estimators=100, train:test = 8:2`), we trained RF with a **five-fold cross-validation** and **hyperparameter-optimization.**



**Baseline**

All Data

Training data | Test data

**GridSearchCV**

| | | | | | |
|---|---|---|---|---|---|
| | Fold 1 | Fold 2 | Fold 3 | Fold 4 | Fold 5 |
| Split 1 | Fold 1 | Fold 2 | Fold 3 | Fold 4 | Fold 5 |
| Split 2 | Fold 1 | Fold 2 | Fold 3 | Fold 4 | Fold 5 |
| Split 3 | Fold 1 | Fold 2 | Fold 3 | Fold 4 | Fold 5 |
| Split 4 | Fold 1 | Fold 2 | Fold 3 | Fold 4 | Fold 5 |
| Split 5 | Fold 1 | Fold 2 | Fold 3 | Fold 4 | Fold 5 |

Finding Parameters

GridSearch

| | | | | |
|---|---|---|---|---|
| n_estimators | 50 | 100 | 150 | 200 |
| max_depth | 10 | 20 | 30 | None |
| min_samples_split | 2 | 5 | 10 | |
| max_features | 'sqrt' | 'log2' | None | |

Final evaluation | Test data

# Experiment 2: Explore another feature generation method

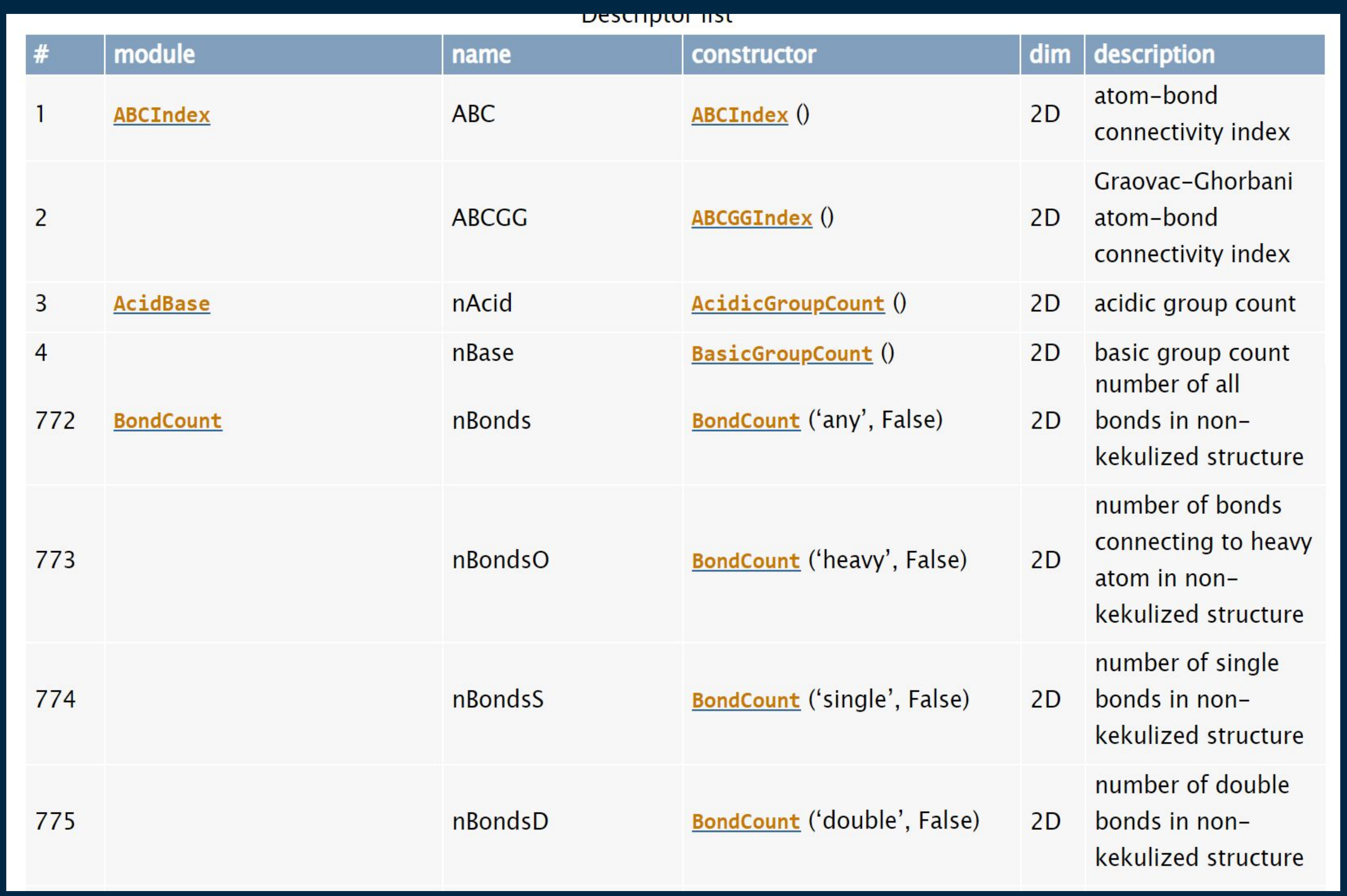
## What is Mordred?

***A novel, promising descriptor calculator library for QSAR***

- Easy installation and usage, open-source.
- Twice as fast as the well-known PaDEL-Descriptor.
- Works with other descriptor libraries (RDKit) or cheminformatics tools.
- Easy calculation for large molecules.

*We used 2D features: structural and topological properties.*

*Such as ABCIndex, EStates, BCUT, acid-base properties, bond count, aromaticity, atom count, etc.*



Descriptor list

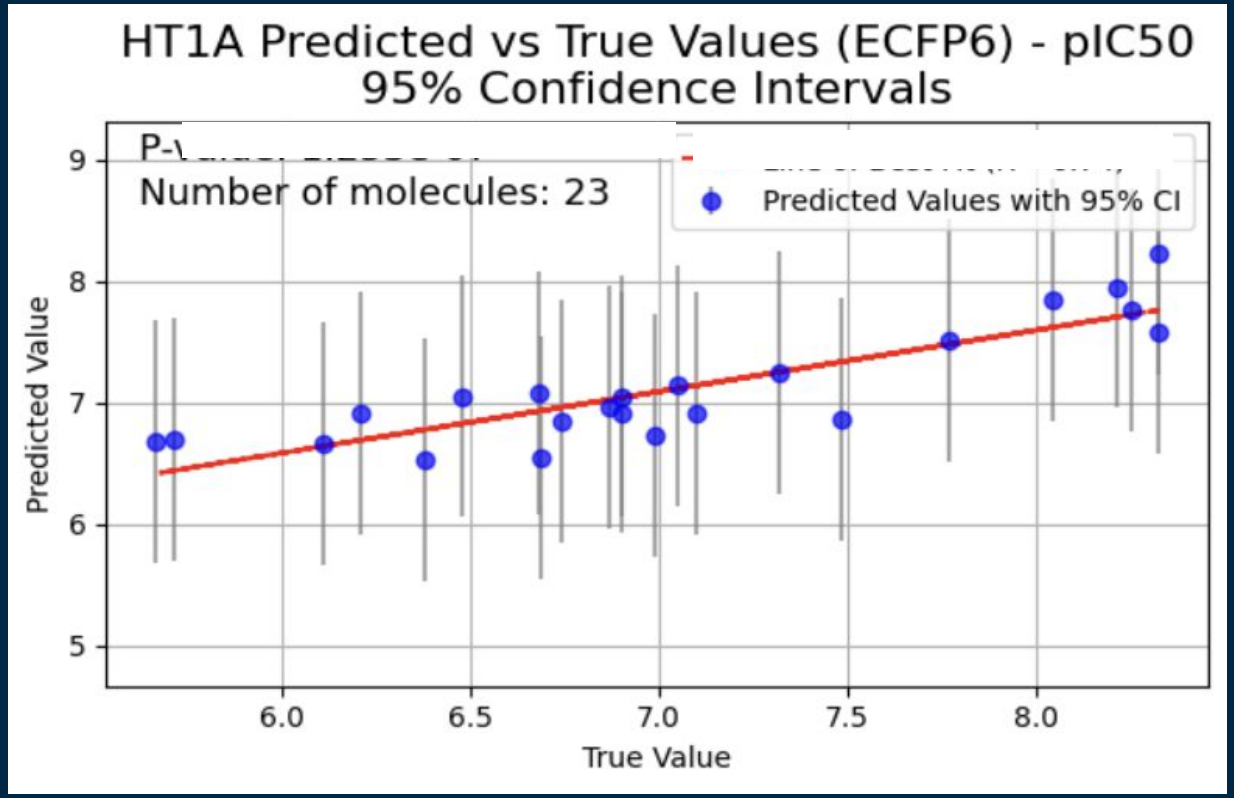
| # | module | name | constructor | dim | description |
|---|---|---|---|---|---|
| 1 | ABCIndex | ABC | ABCIndex () | 2D | atom-bond connectivity index |
| 2 | | ABCGG | ABCGGIndex () | 2D | Graovac-Ghorbani atom-bond connectivity index |
| 3 | AcidBase | nAcid | AcidicGroupCount () | 2D | acidic group count |
| 4 | | nBase | BasicGroupCount () | 2D | basic group count |
| 772 | BondCount | nBonds | BondCount ('any', False) | 2D | number of all bonds in non-kekulized structure |
| 773 | | nBondsO | BondCount ('heavy', False) | 2D | number of bonds connecting to heavy atom in non-kekulized structure |
| 774 | | nBondsS | BondCount ('single', False) | 2D | number of single bonds in non-kekulized structure |
| 775 | | nBondsD | BondCount ('double', False) | 2D | number of double bonds in non-kekulized structure |

*Figure* from Moriwaki, H., Tian, YS., Kawashita, N. *et al.* Mordred: a molecular descriptor calculator. *J Cheminform* 10, 4 (2018). https://doi.org/10.1186/s13321-018-0258-y

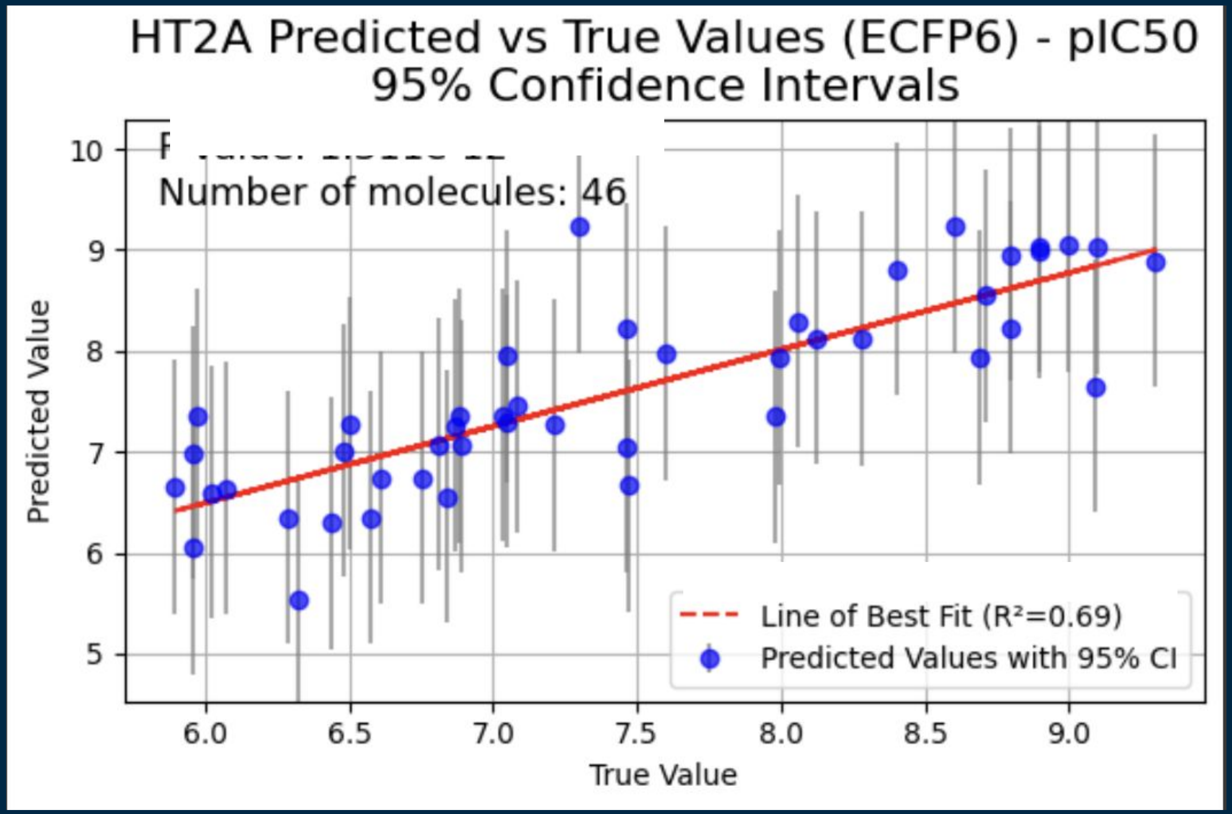
# Results: Baseline ECFP6 model

$$\text{Relative Width} = \frac{\text{CI Width}}{\text{Predicted Value}} \times 100$$

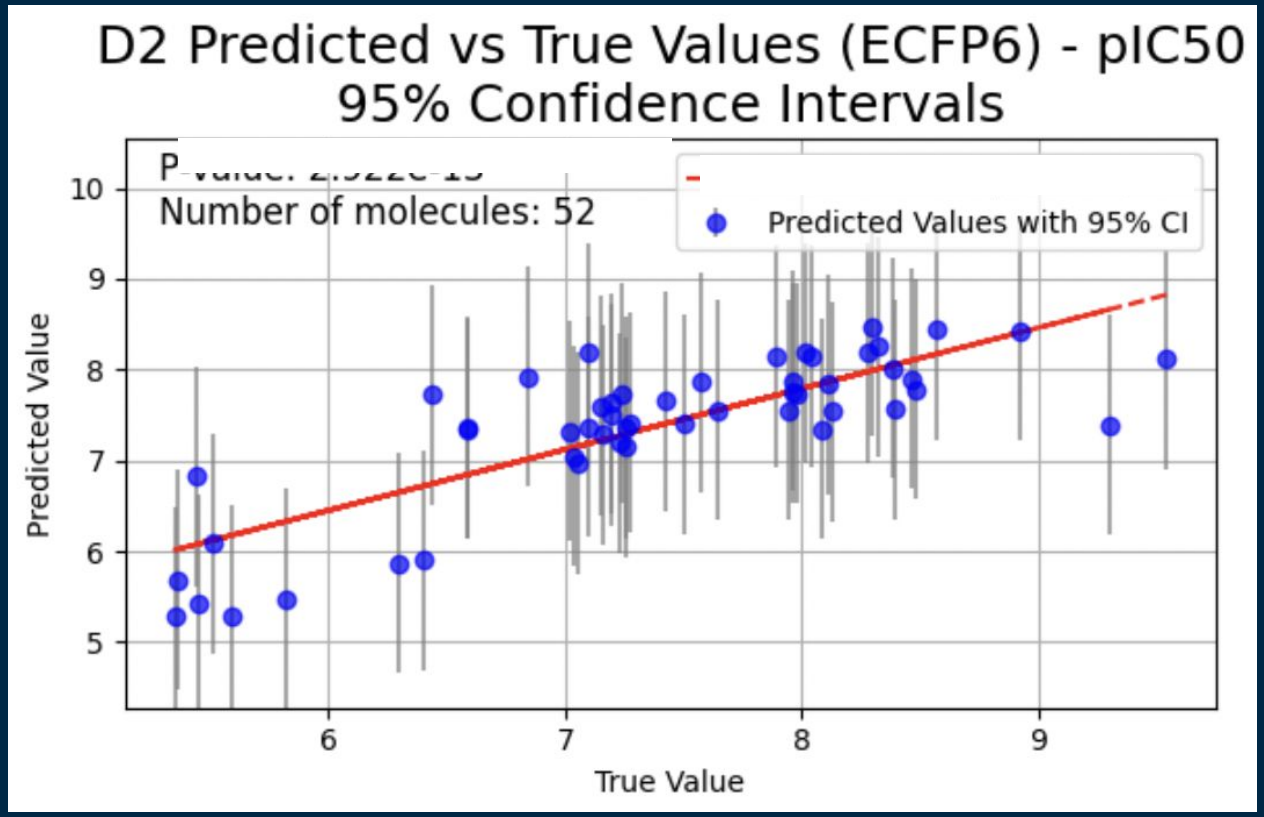
5-HT1A: $R^2$ = 0.6557


HT1A Predicted vs True Values (ECFP6) - pIC50
95% Confidence Intervals

5-HT2A: $R^2$ = 0.6557


HT2A Predicted vs True Values (ECFP6) - pIC50
95% Confidence Intervals

D2: $R^2$ = 0.6585


D2 Predicted vs True Values (ECFP6) - pIC50
95% Confidence Intervals

'Mean Relative Width': 0.28,
'Median Relative Width': 0.29,
'Standard Deviation': 0.017

'Mean Relative Width': 0.34,
'Median Relative Width': 0.34,
'Standard Deviation': 0.043

'Mean Relative Width': 0.33,
'Median Relative Width': 0.32,
'Standard Deviation': 0.045

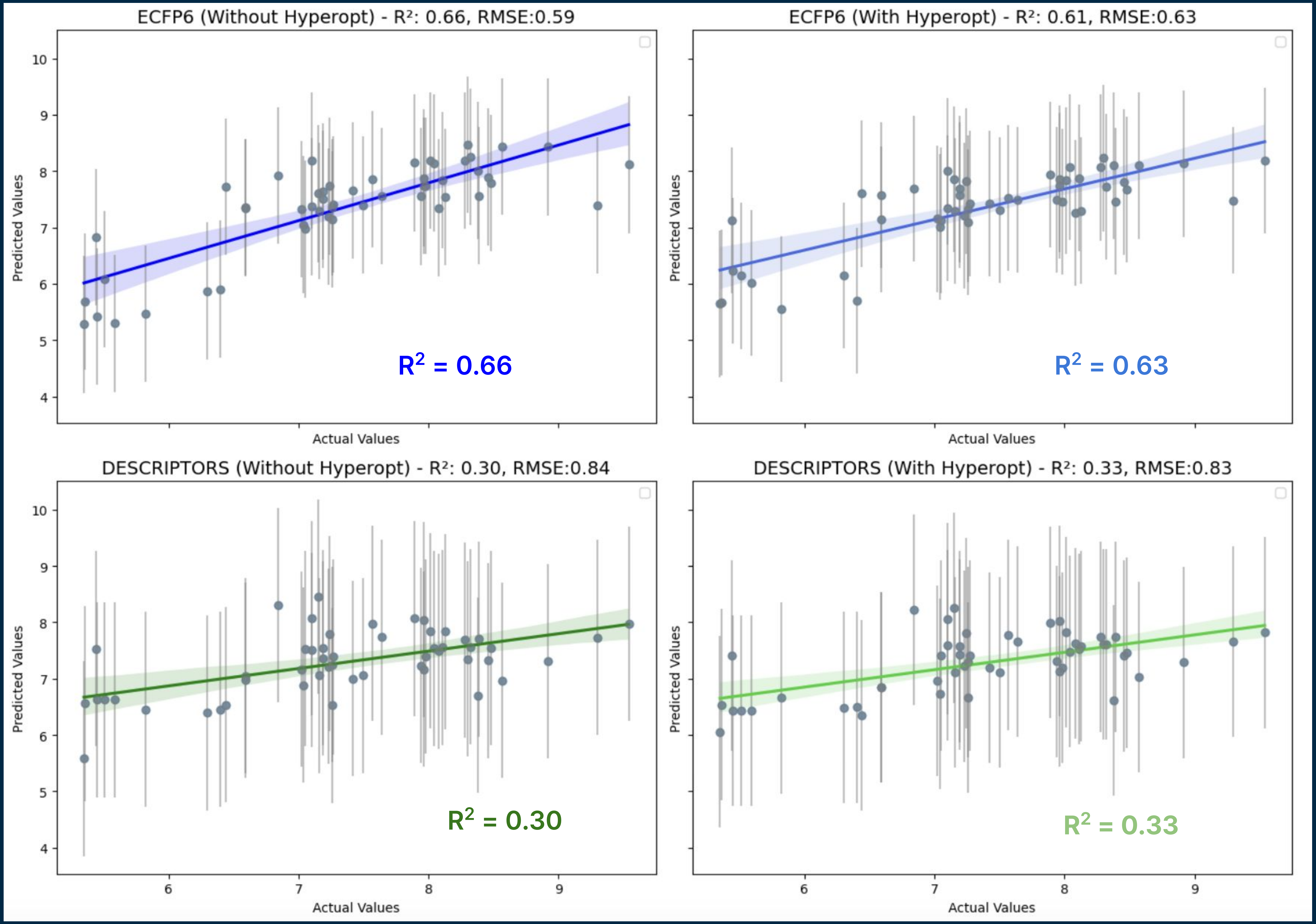
# D2

without hyper_opt

with hyper_opt

ECFP6

Morderd



● ECFP6 (Without Hyperopt)
● ECFP6 (With Hyperopt)
● DESCRIPTORS (Without Hyperopt)
● DESCRIPTORS (With Hyperopt)

**Conclusion:**
RandomForest with hyperparameter optimization and Mordred descriptors **did not improve** performance

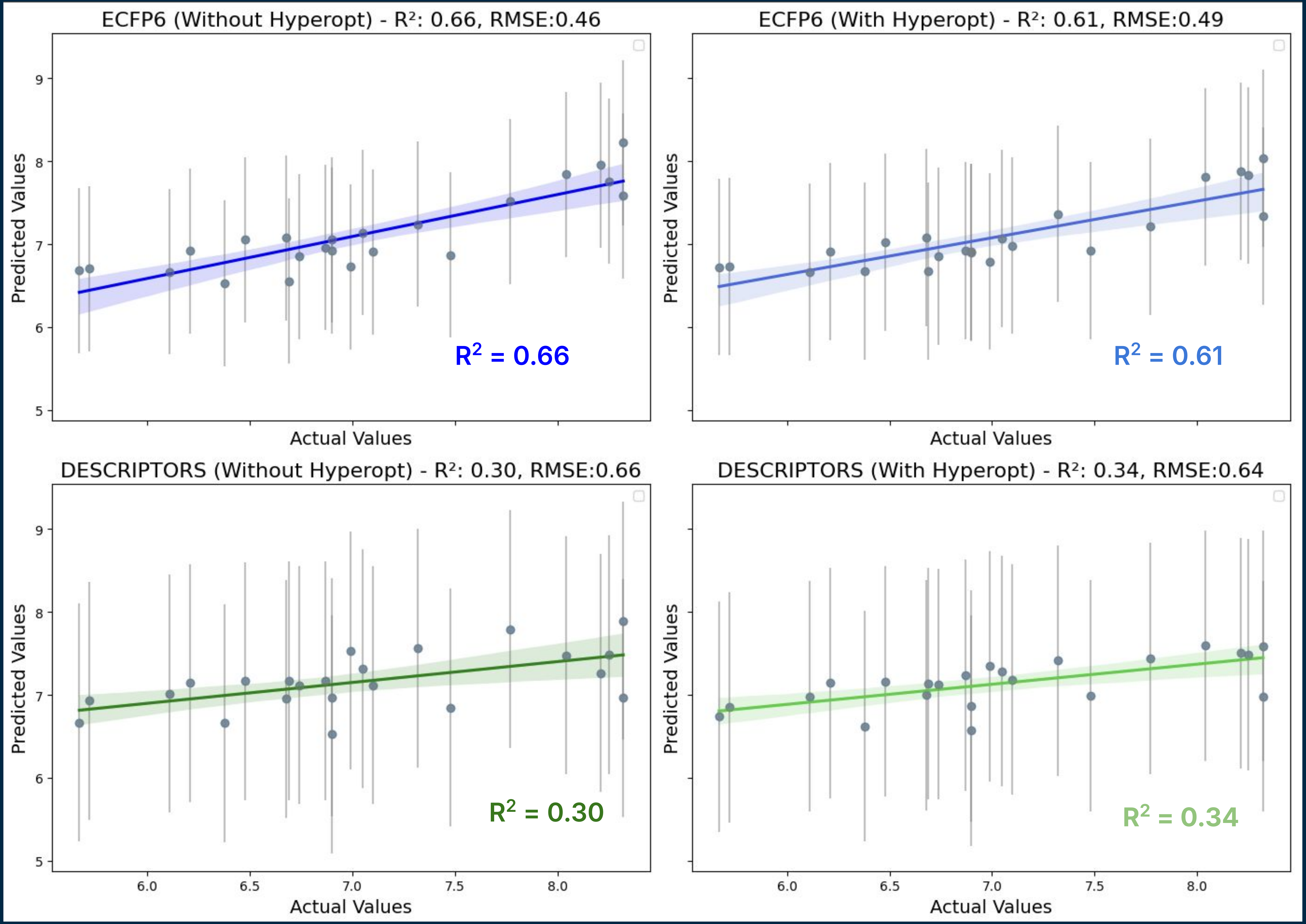
# 1A

without hyper_opt

with hyper_opt

ECFP6

Morderd



● ECFP6 (Without Hyperopt)

● ECFP6 (With Hyperopt)

● DESCRIPTORS (Without Hyperopt)

● DESCRIPTORS (With Hyperopt)

**Conclusion:**
RandomForest with hyperparameter optimization and Mordred descriptors **did not improve** performance

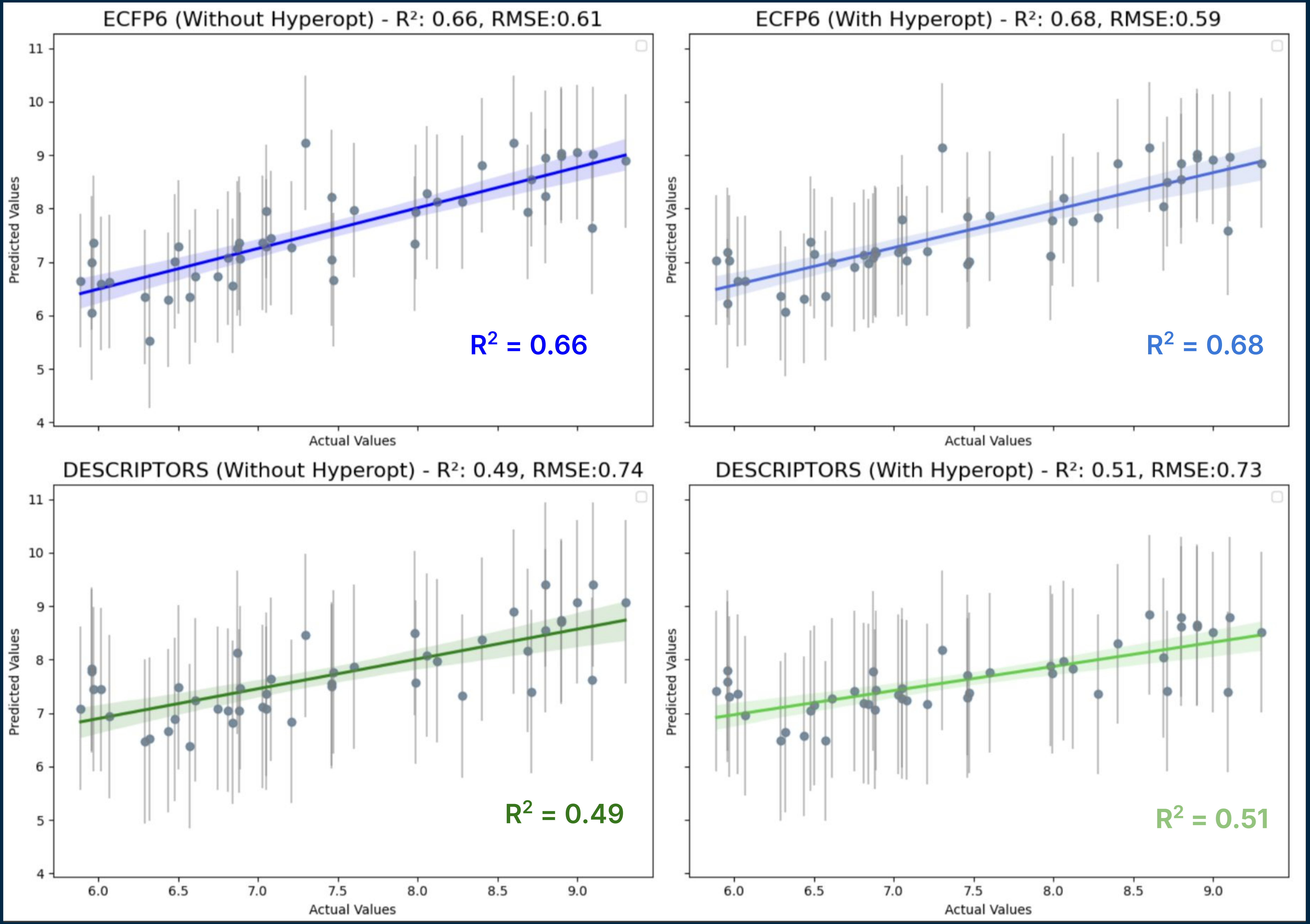
## 2A

without hyper_opt

with hyper_opt

ECFP6

Morderd



- ● ECFP6 (Without Hyperopt)
- ● ECFP6 (With Hyperopt)
- ● DESCRIPTORS (Without Hyperopt)
- ● DESCRIPTORS (With Hyperopt)

**Conclusion:**

RandomForest with hyperparameter optimization and Mordred descriptors **did not improve** performance

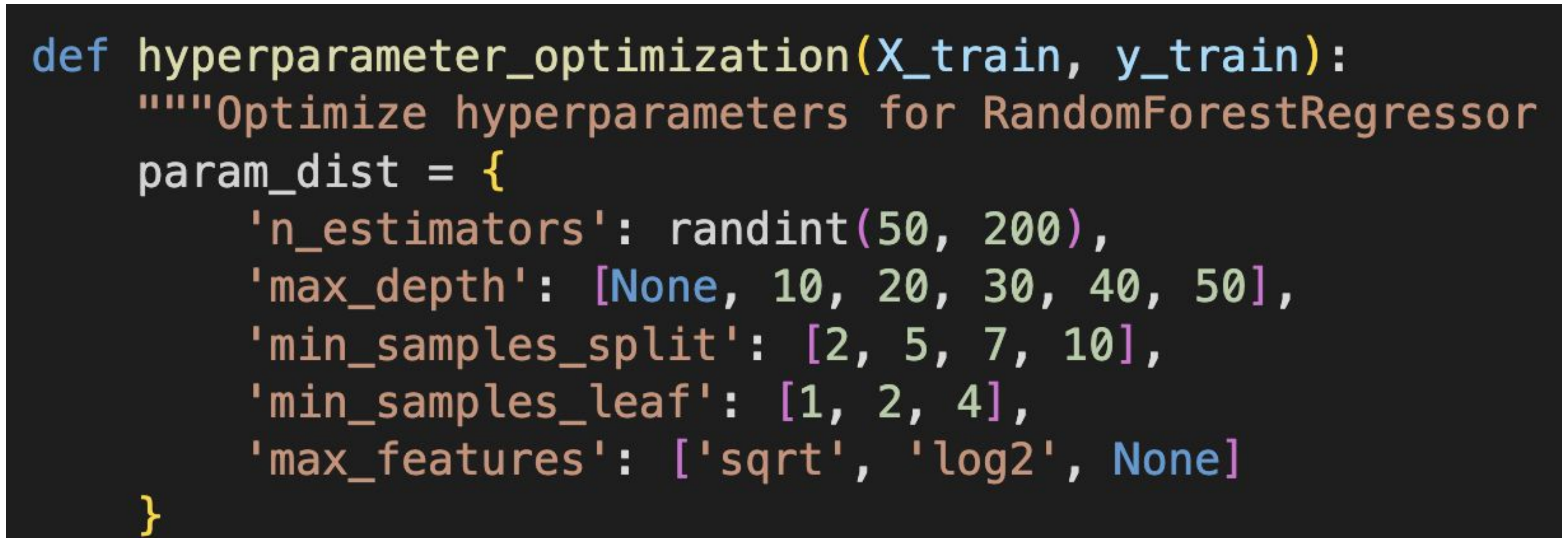
# Why?

## Hyperparamter Optimization

- our preselected values for the GridSearch were not optimal

## Mordered Descriptors

- ECFP6 is more targeted toward capturing the structural features, while mordred descriptors are **broader but less specialized**. (like weight, polarity, etc)



```python
def hyperparameter_optimization(X_train, y_train):
    """Optimize hyperparameters for RandomForestRegressor
    param_dist = {
        'n_estimators': randint(50, 200),
        'max_depth': [None, 10, 20, 30, 40, 50],
        'min_samples_split': [2, 5, 7, 10],
        'min_samples_leaf': [1, 2, 4],
        'max_features': ['sqrt', 'log2', None]
    }
```

# Baseline Results Summary

## Predictive Accuracy

- **Strong $R^2$ (0.66)** values confirm the utility of Random Forest + **ECFP6** for pChEMBL prediction.
- Endpoint 1A shows the highest reliability for serotonin receptor bioactivity.

## Prediction Confidence

- **Narrow CIs in 1A** ensure precise predictions for compound prioritization.
- Slightly **wider CIs in 2A/D2** indicate variability but remain statistically robust.

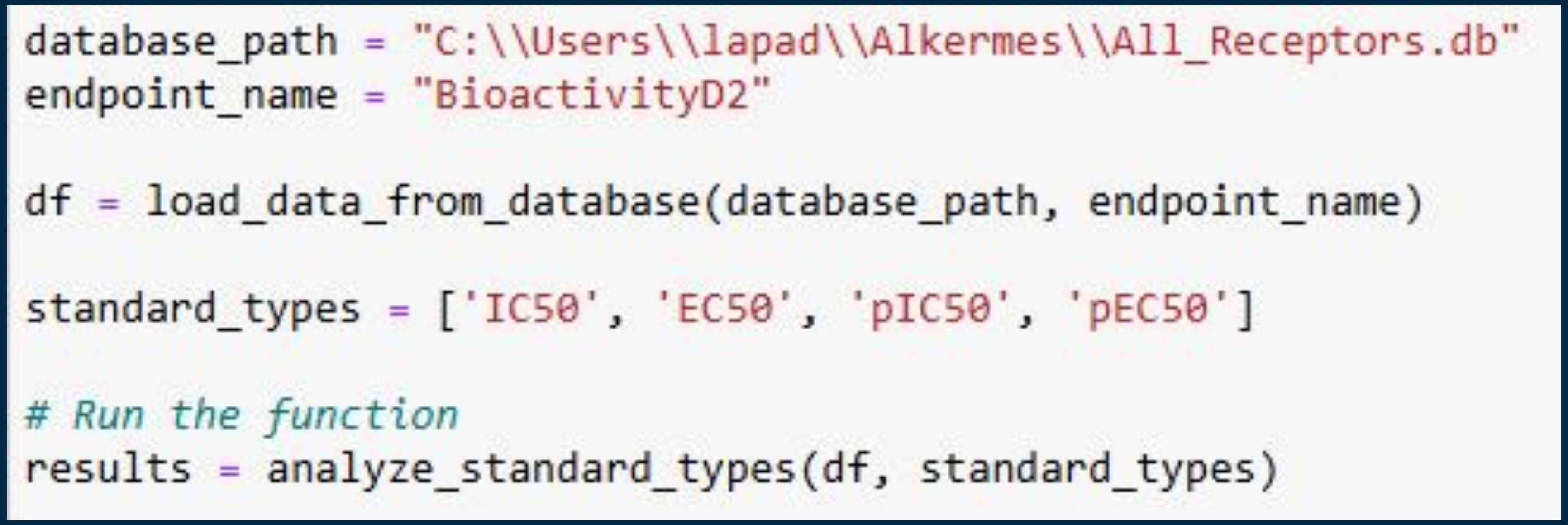
# 04

# Conclusion

# *Future Steps*

## Automation:

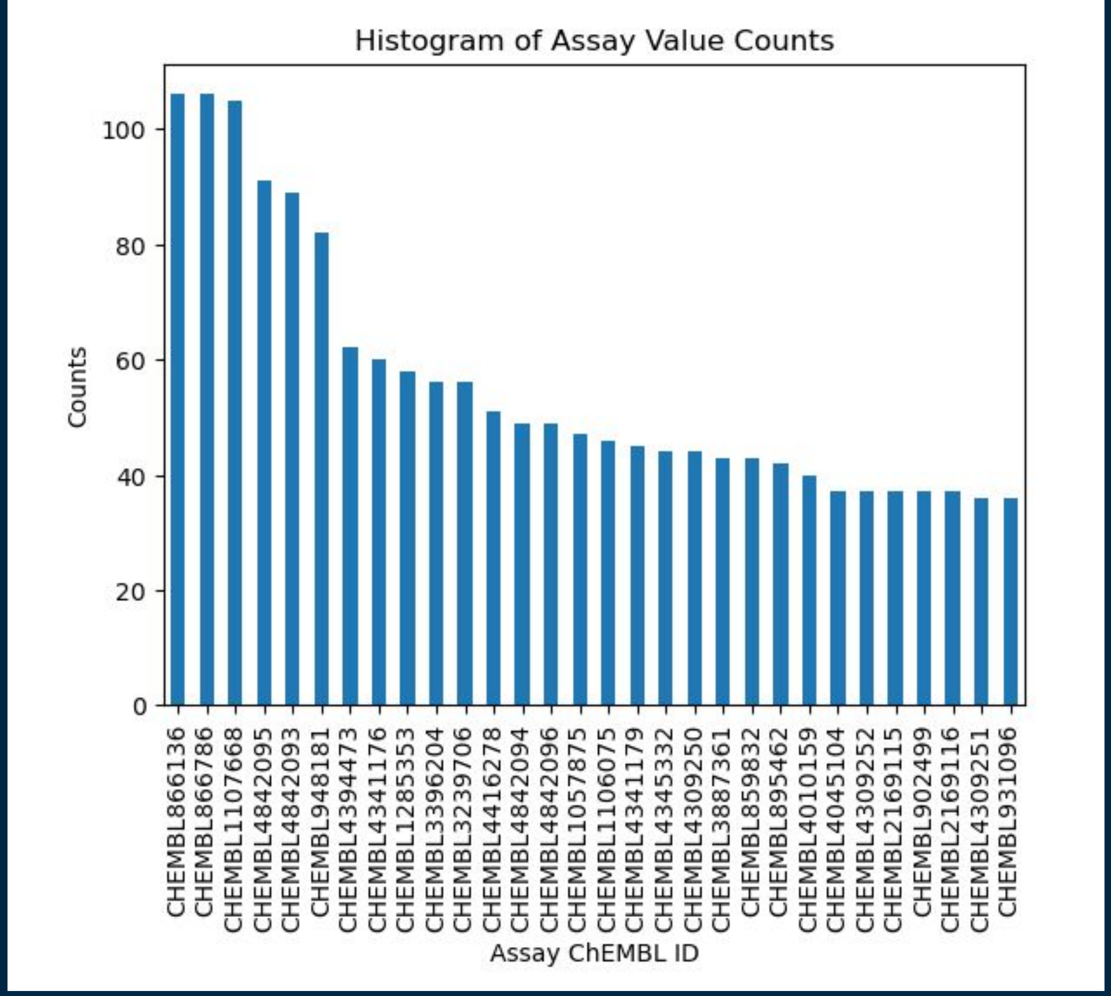
- make the process accessible and efficient
- Input data → automatic output
- No need for manual filtering or feature selection



```
database_path = "C:\\Users\\lapad\\Alkermes\\All_Receptors.db"
endpoint_name = "BioactivityD2"

df = load_data_from_database(database_path, endpoint_name)

standard_types = ['IC50', 'EC50', 'pIC50', 'pEC50']

# Run the function
results = analyze_standard_types(df, standard_types)
```

## Larger sample:

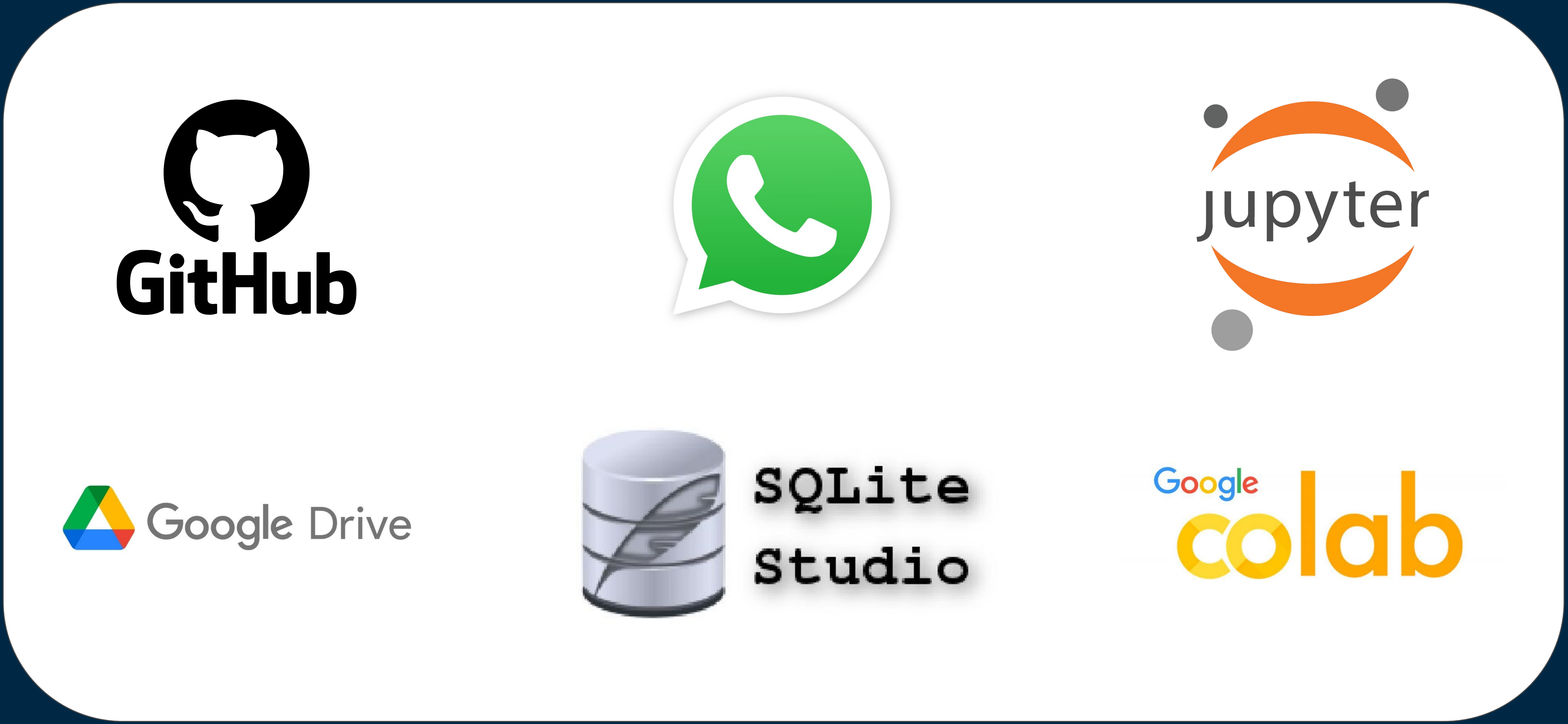
- Expand from Top 10 to Top 100 Assay ChEMBL IDs



## ChemProp:

- PyTorch-based framework for training and evaluating message-passing neural networks (MPNNs)
- User-friendly molecular property prediction

# *How we made it work?*



# *Lessons and Future Directions from our Team Members*

**Alex Lapadat**: BTTAI and our Alkeres project were incredible experiences - I gained a **strong foundation in machine learning**, applied it to **neuroscience**, all while overcoming challenges as a junior, and developing valuable skills for my future research career in biostatistics.

**Tiffney Aina**: I learned how big of a role **data engineering** plays. It became clear that building a model is not merely a matter of inputting data and obtaining results. Instead, it requires a deep understanding of the ***underlying assumptions***, the quality of the data, and the context in which the model operates.

**Blair Kuzniarek:** I learned the importance of **accurately interpreting data** to make effective decisions. I also gained an understanding of how clear communication and **proactive organization** ensure team alignment and smooth progress.

**Ray Qin:** I learned how different data cleanup/selection methods can result in very **different ML model prediction and accuracy**. It is important to understand the specific topic that the model is predicting on, for a comprehensive consideration when building and improving the model.

**Ha Dong:** It was an amazing experience where I was able to learn **how drug discovery is practiced in an industry environment**. The level of **scientific rigor**, **attention to details**, and **problem-solving strategy** Alkermes scientists taught us will definitely come in handy in my future research.

# Thank You!

BTTAI Program Organizers: For providing this incredible learning opportunity, and hosting the Maker Days where we got to learn so much!

Joerg, Polina and Shin: For your guidance and expertise for the past 5 months, your patience and optimism!

Divya, our TA: For your advice for our presentation and your understanding!

And to our audience, as well!