Uppsala University

June 5th, 2025

Conformal Prediction: Study case of Regression and Classification for Binding Affinity Modelling

By Dinh Long Huynh

CONFORMAL PREDICTION

Classification model

$\alpha_i = 1 - p(class|x_i)$

$\alpha_i = -\log[p(class|x_i)]$

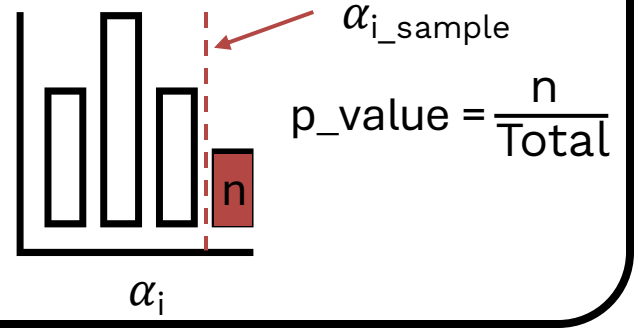
Regression model

$\alpha_i = |y\_pred_i - y\_true_i|$

$\alpha_i = \frac{|y\_predi - y\_truei|}{residual\_pred_i}$

Nonconformity score (α_i)

Calibration α_i distiribution



$p\_value = \frac{n}{Total}$

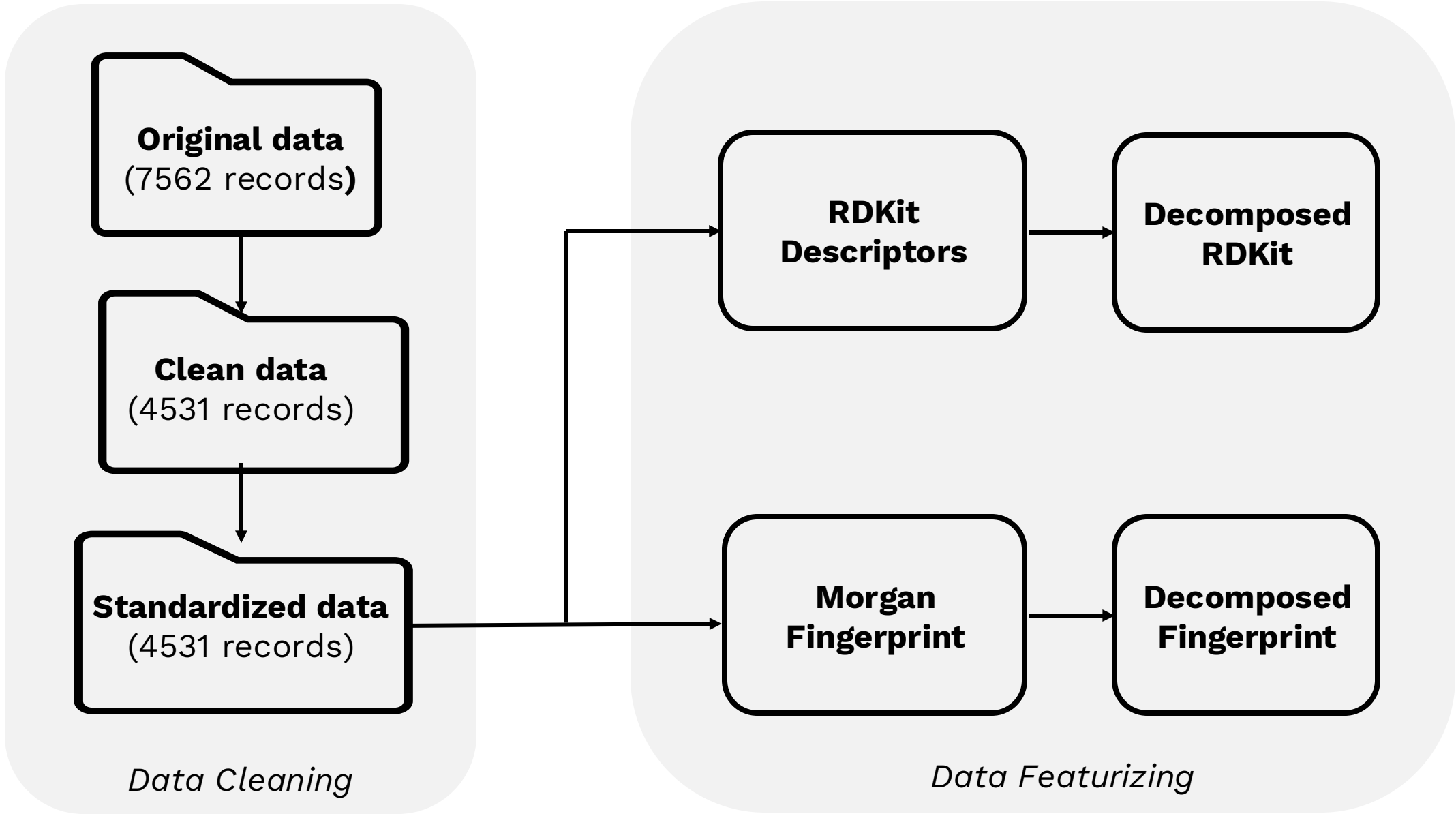
Relationship:
Small error
=> Small α_i
=> Large p_value

Classification
Given a certain class
- Calculate α_{i_sample}
- Calculate p_value
- If **p_value > ε** then **class**

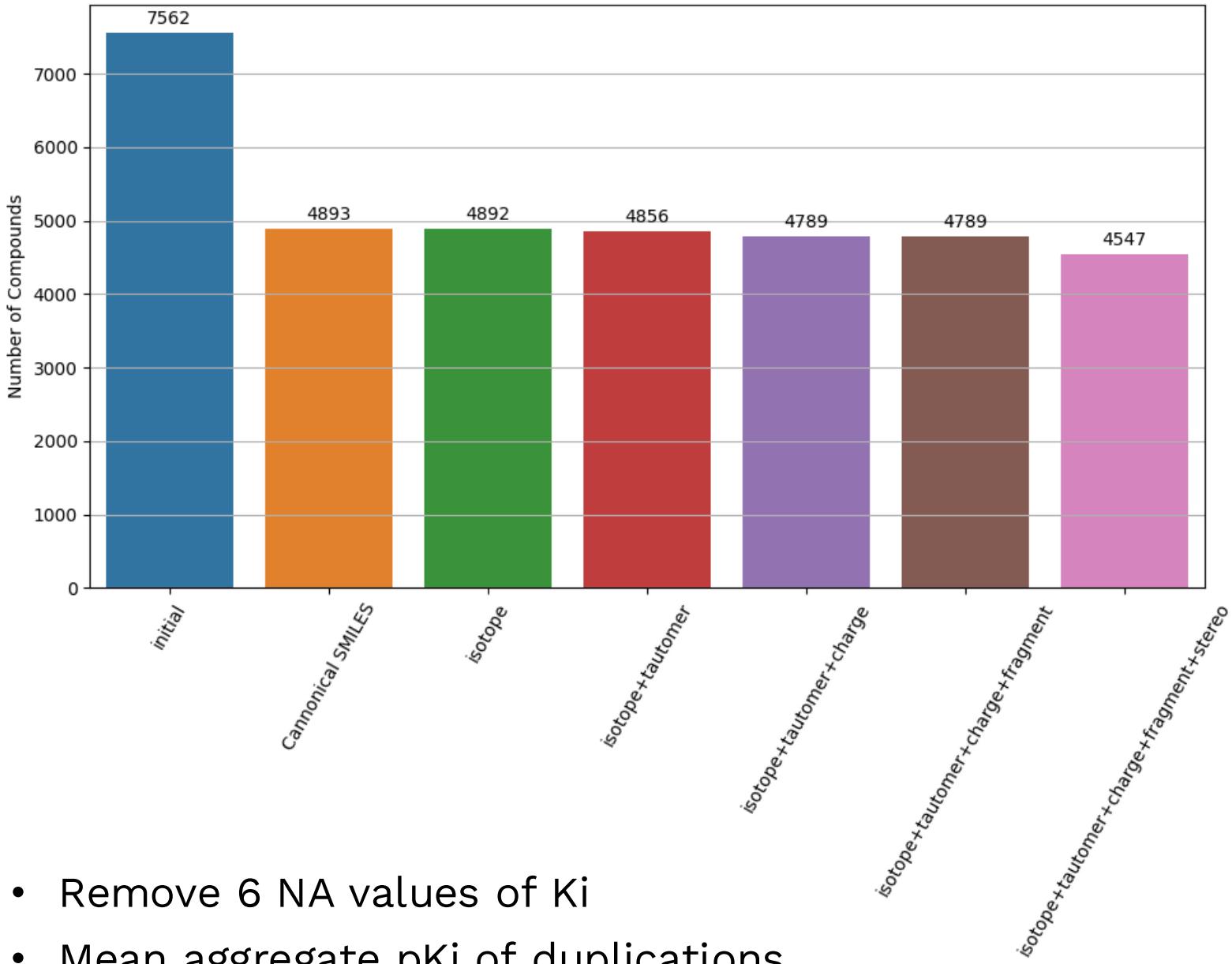
Regression
Given ε, compute confidence interval

Arvidsson et al., *Journal of Cheminformatics, 2024*

DATA HANDLING PIPELINE



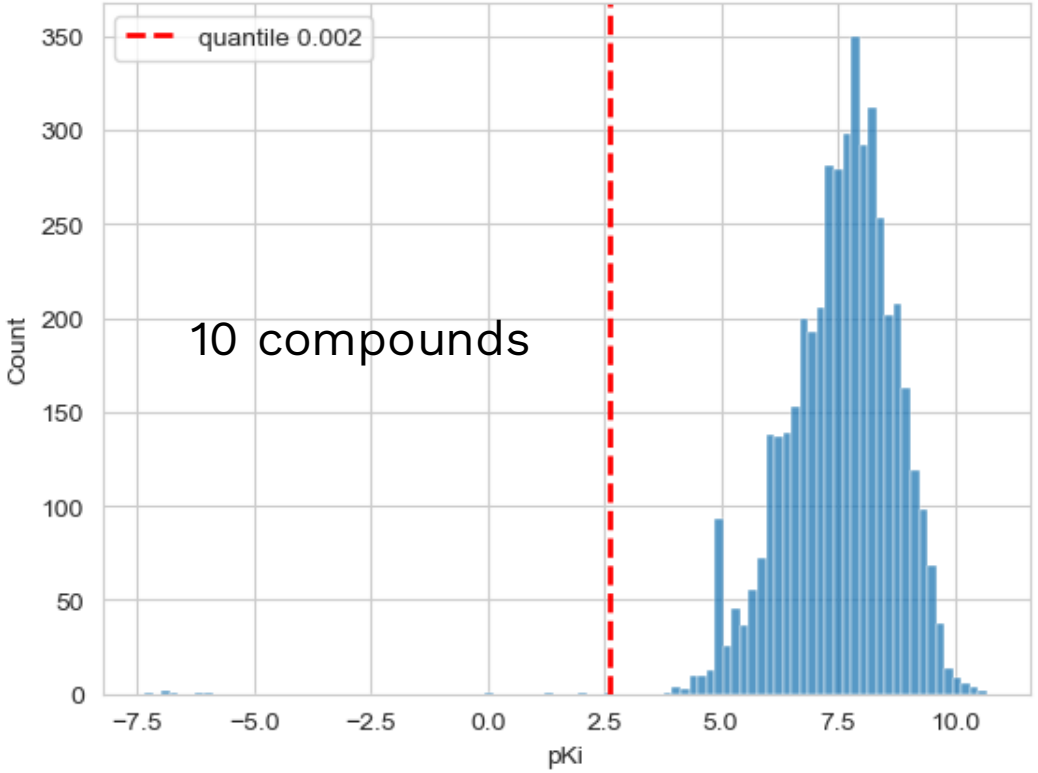
DATA CLEANING

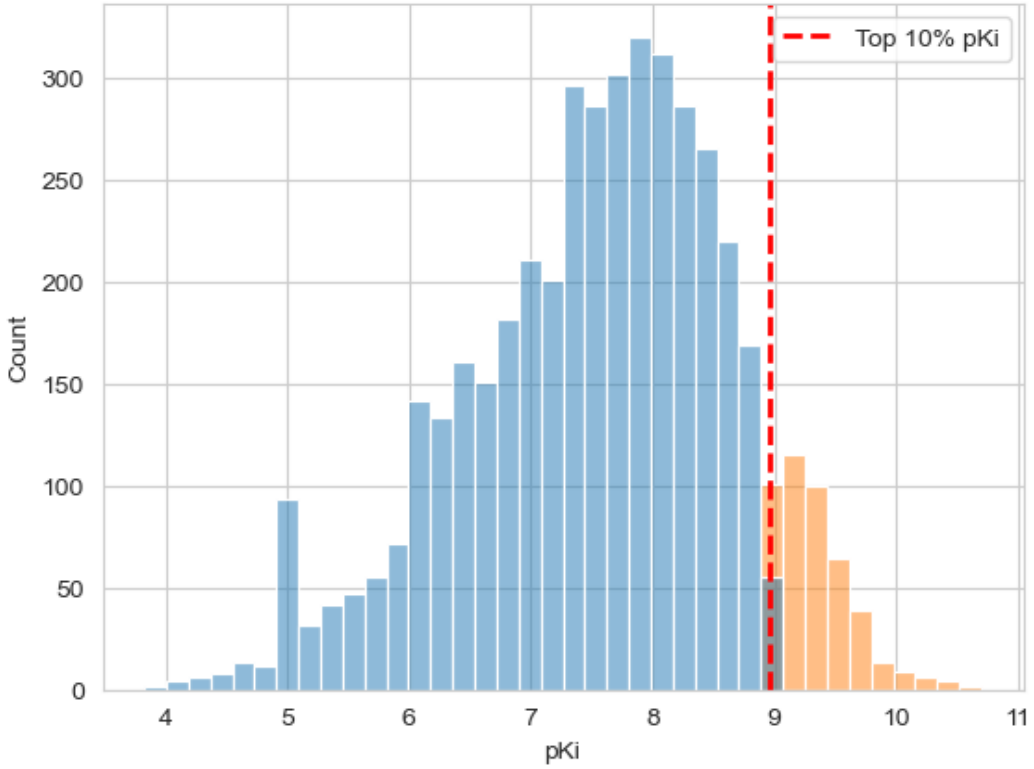


- Remove 6 NA values of Ki
- Mean aggregate pKi of duplications

DATA EXPLORATION

Threshold selection



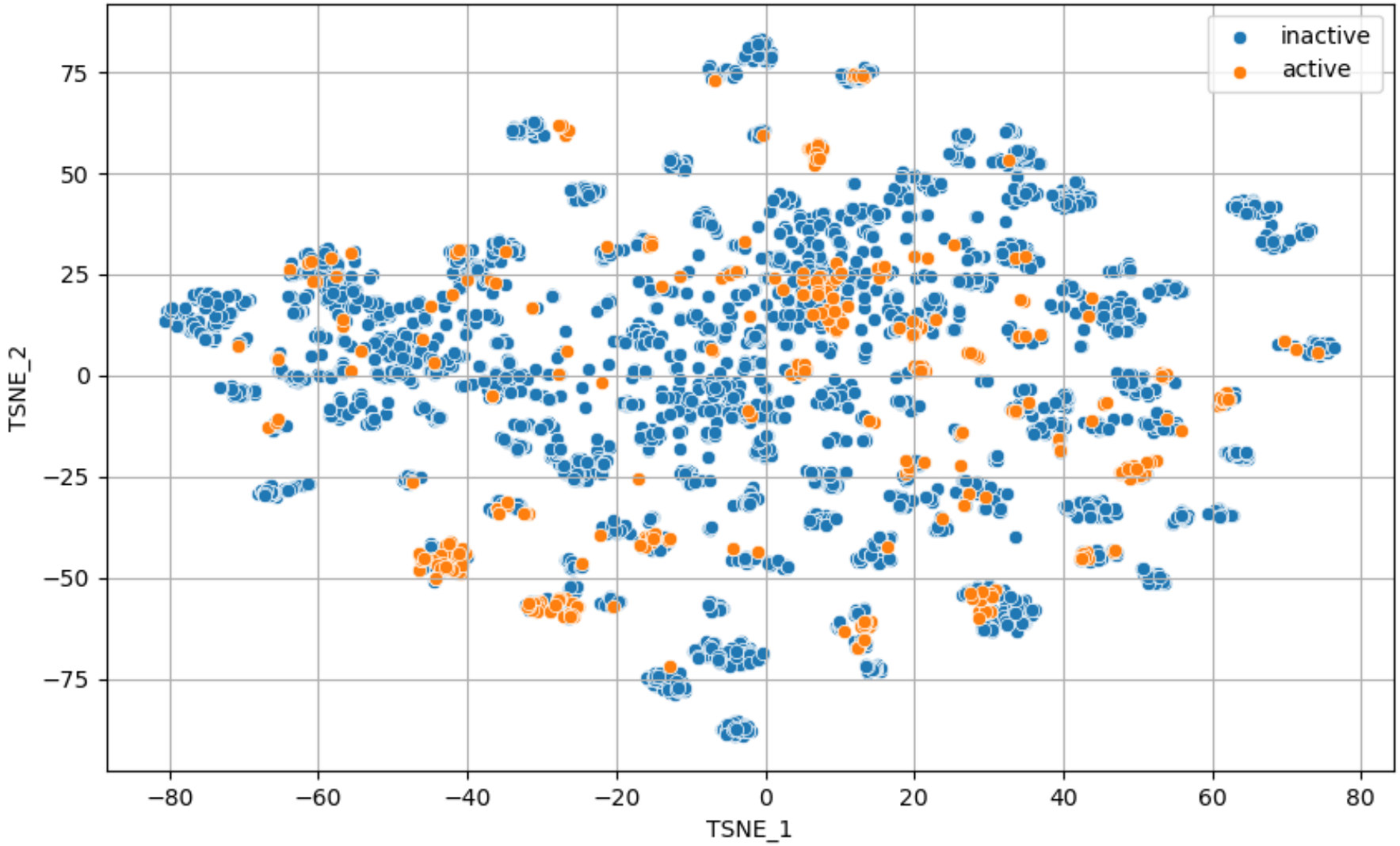


Threshold: $pK_i = 9$, $K_i = 1nM$

Total: 4531 compounds

DATA EXPLORATION

Chemical Space of whole dataset



DATA FEATURIZING

RDKit Descriptors

217 descriptors

Filter descriptors with var = 0

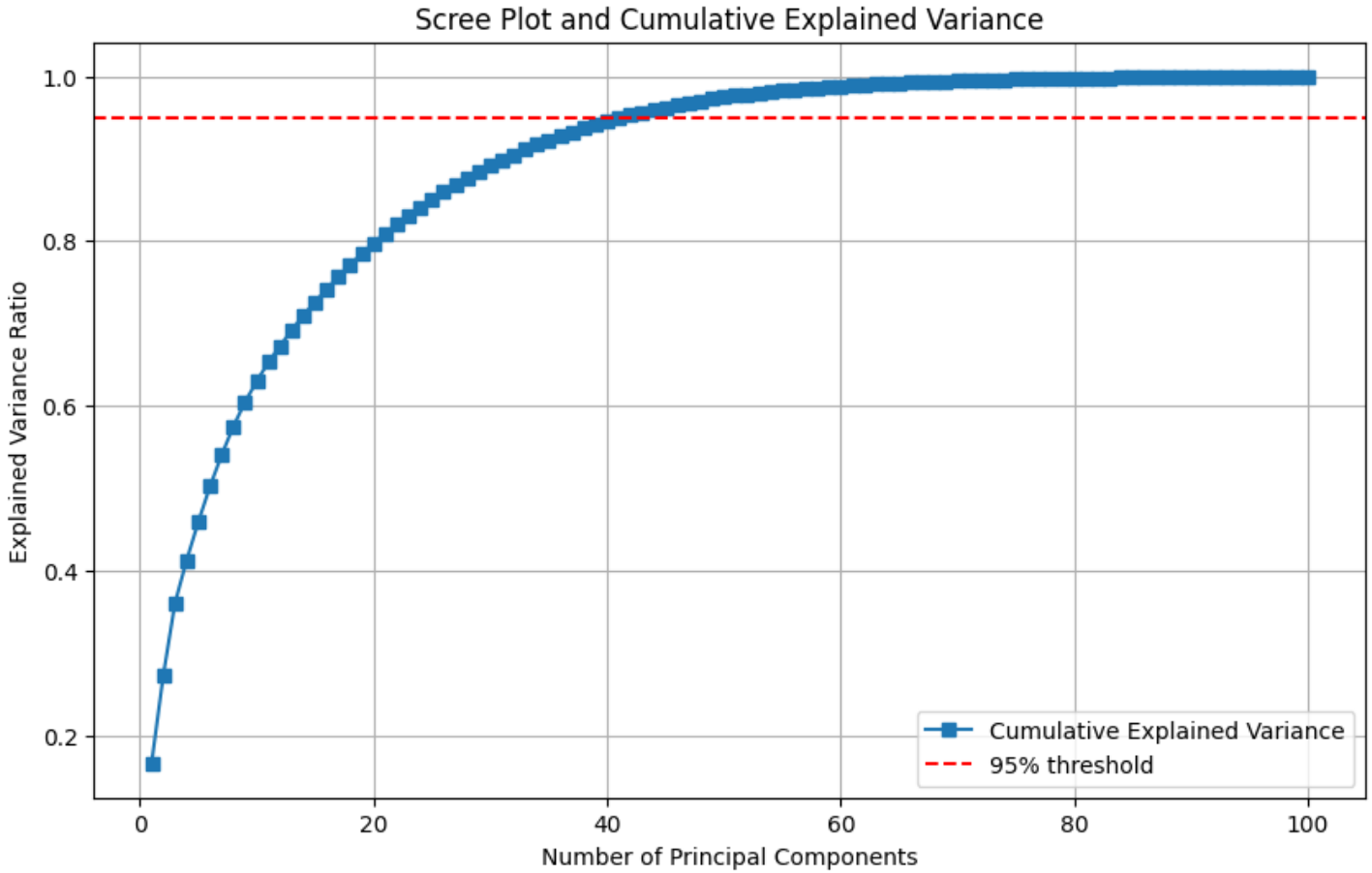
200 var!=0 descriptors

Filter descriptors with corr > 0.7

121 low-correlated descriptors

Guarantee > 95% explained variance

50 PCs



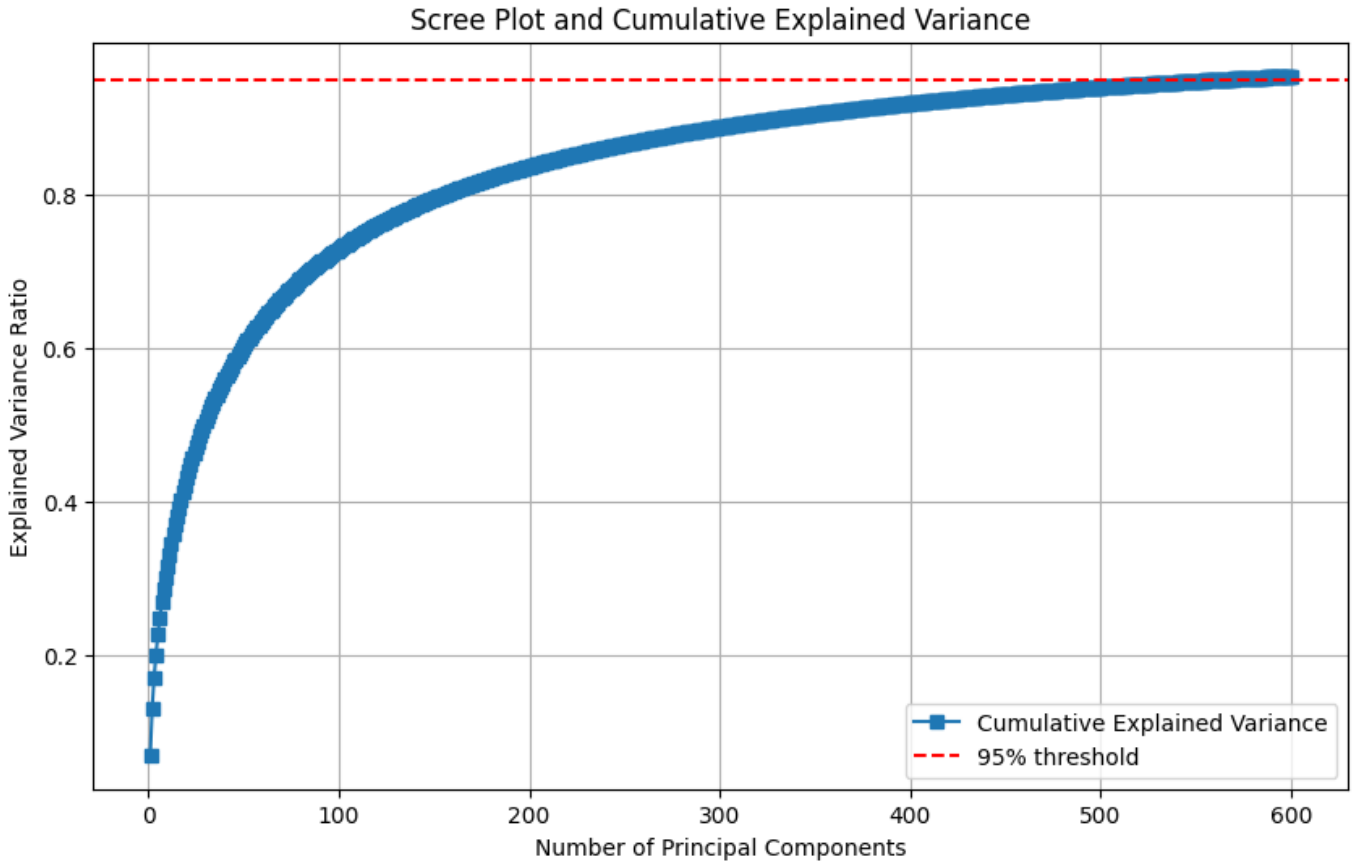
DATA FEATURIZING

Morgan Fingerprint

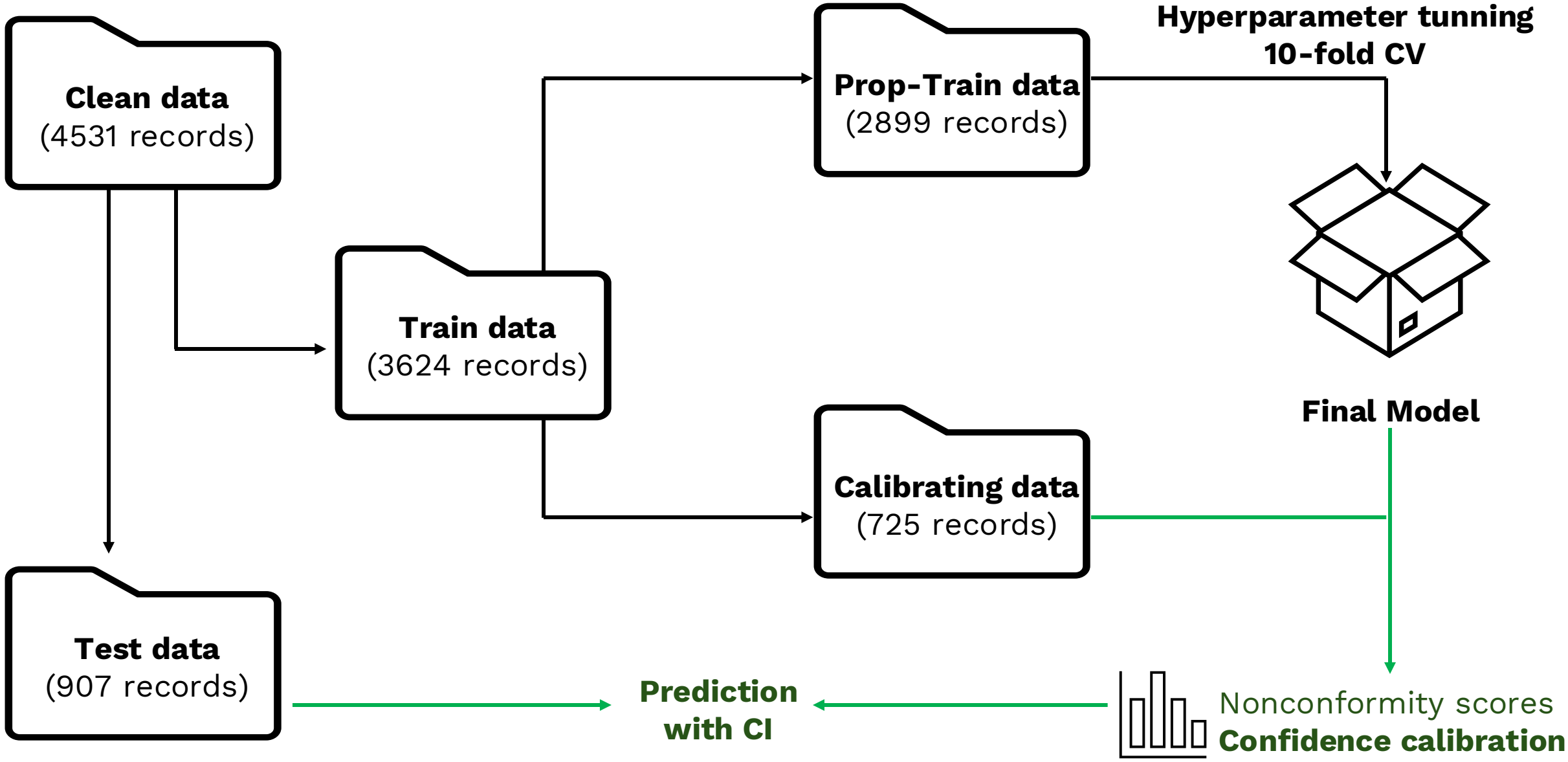
2048 bits, radius = 2

Guarantee > 95% explained variance

600 PCs

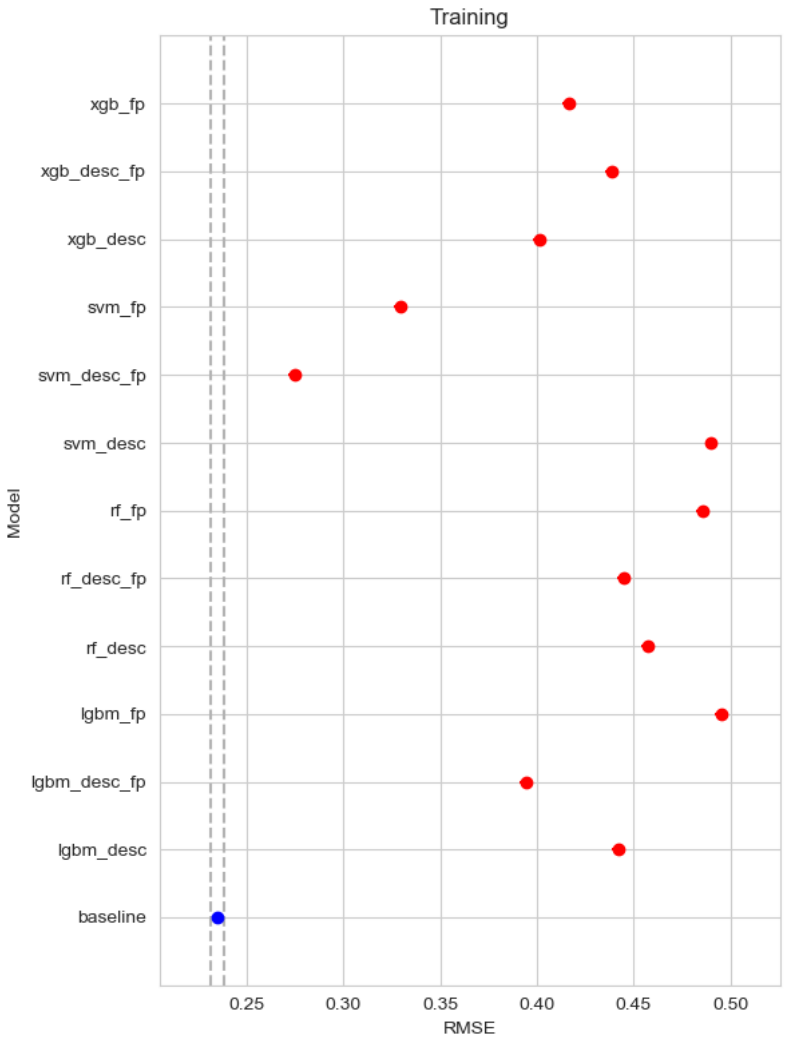


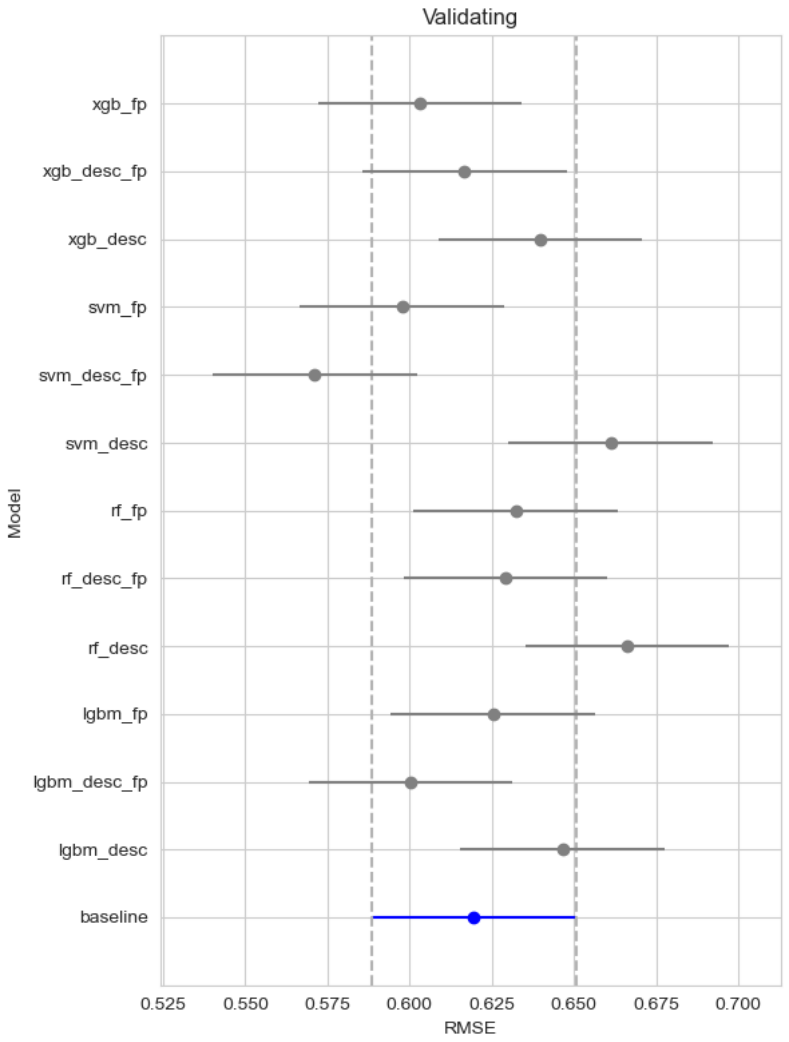
MODEL TRANING PROCESS

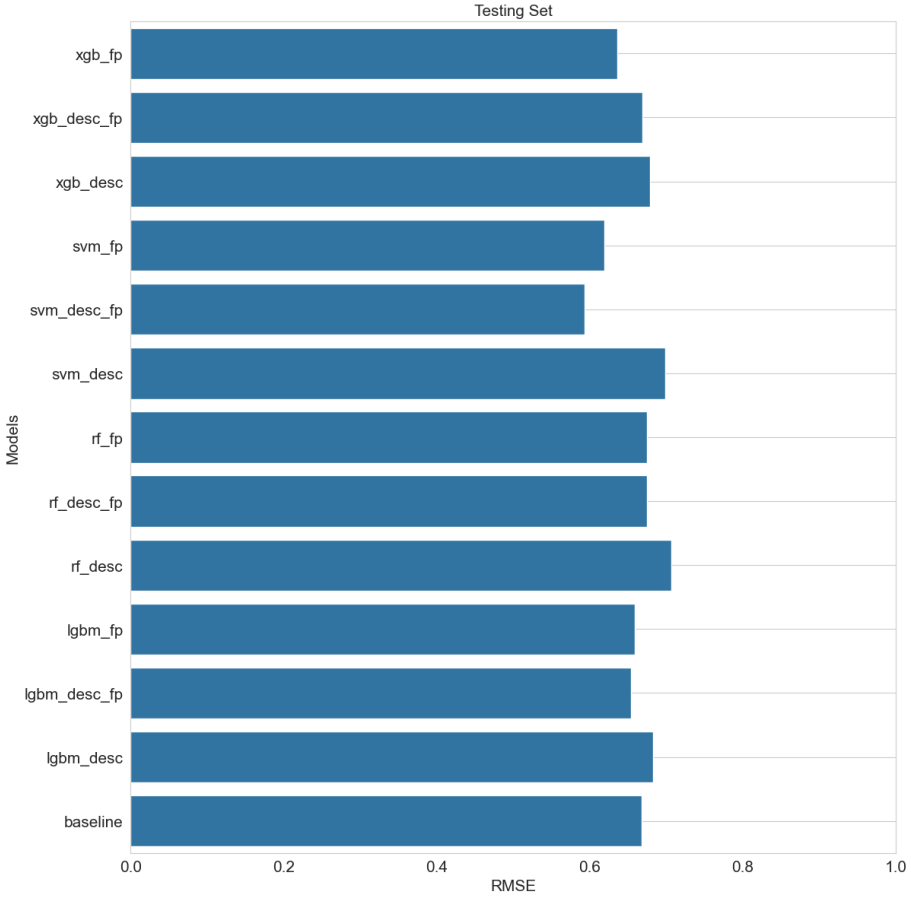


REGRESSION MODELS

Tracking metrics: RMSE



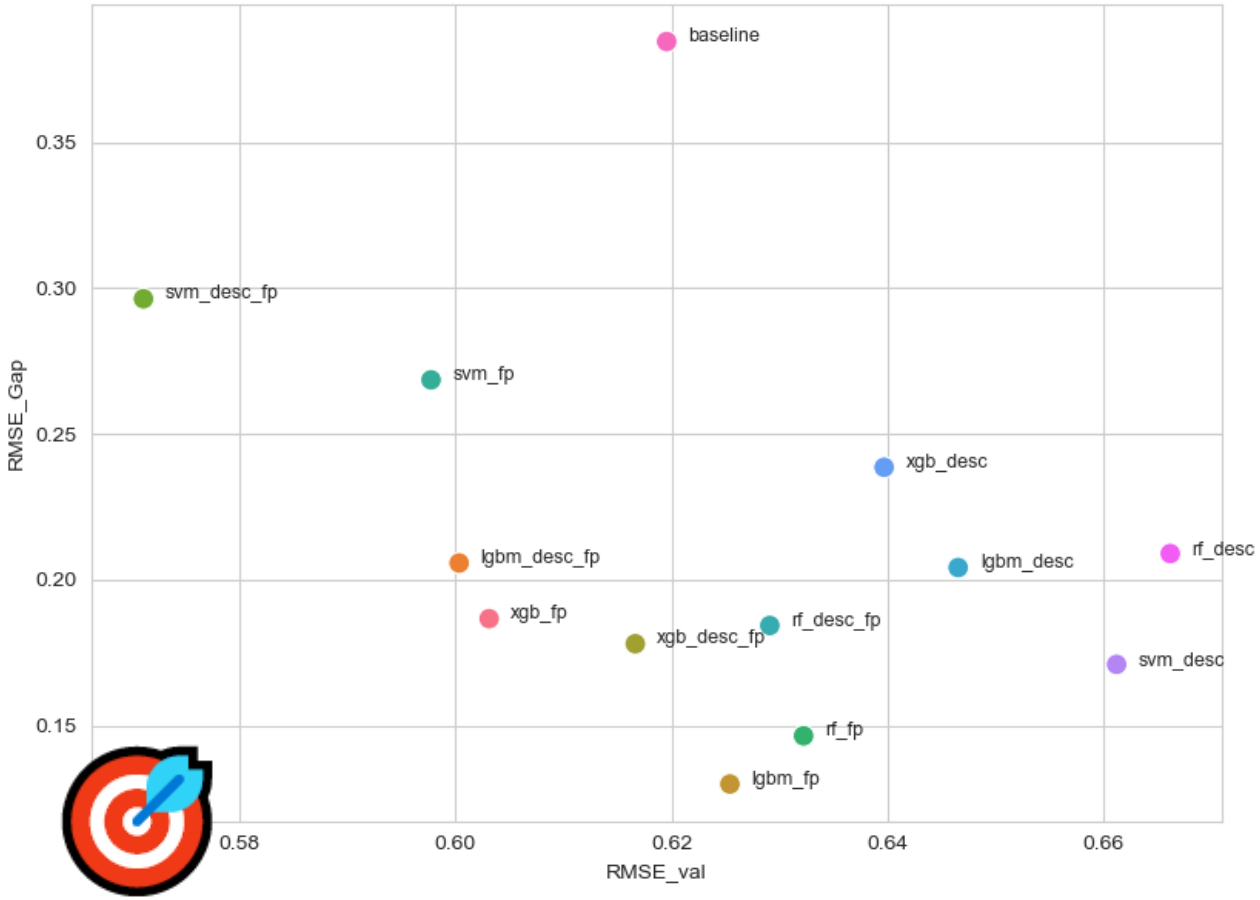


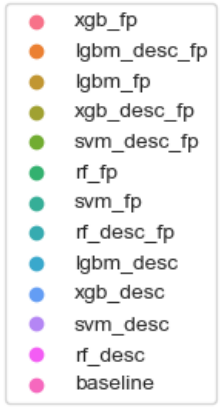


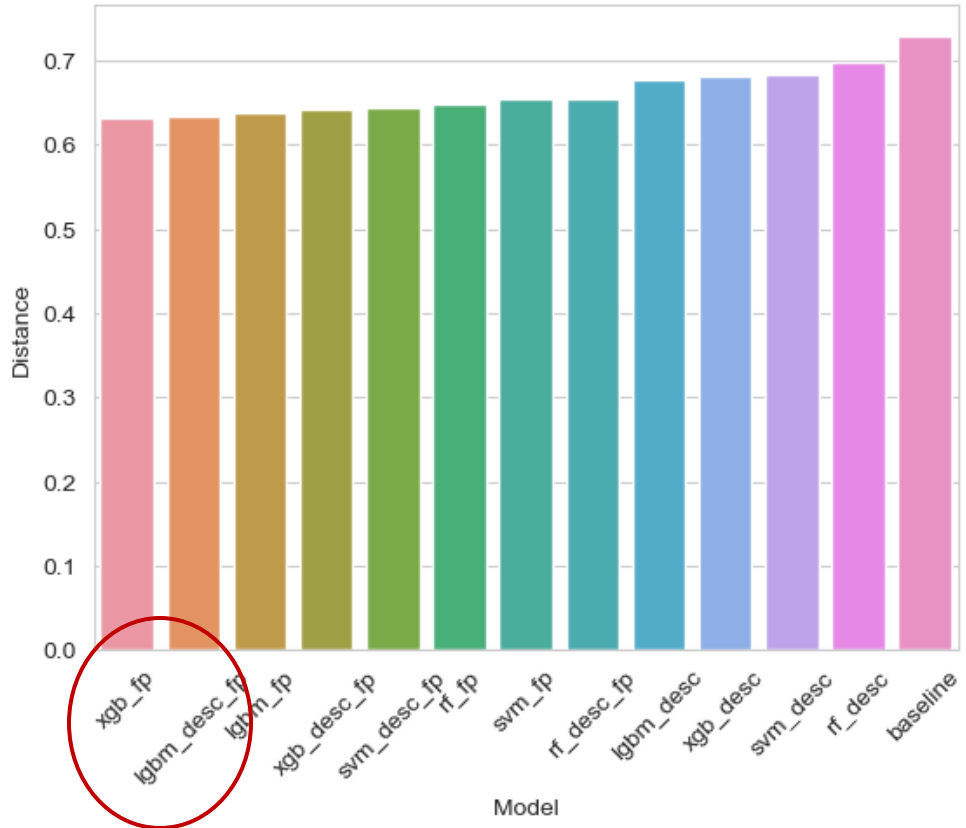
Interval from Turkey-HSD with correction

REGRESSION MODELS

Overfitting analysis

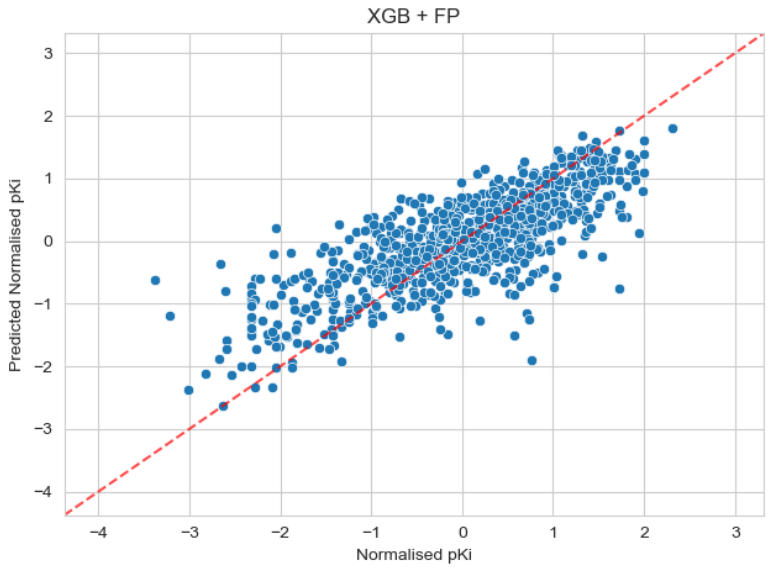


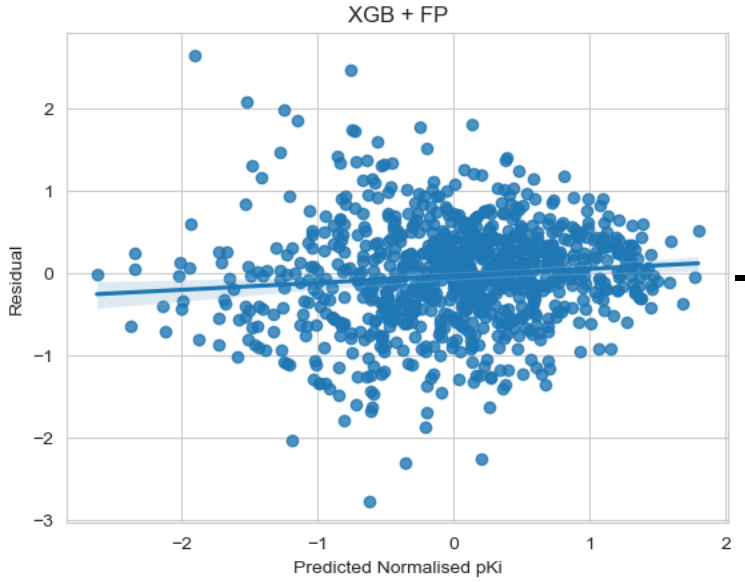




REGRESSION MODELS

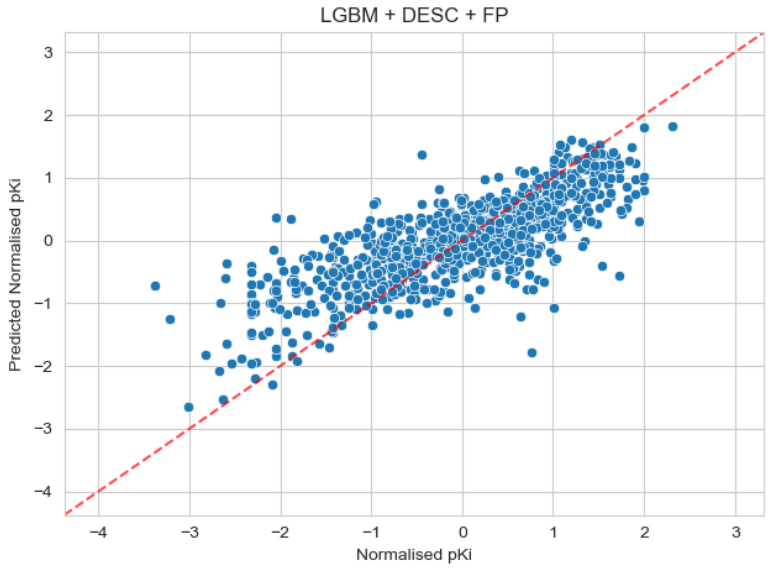
Overfitting analysis

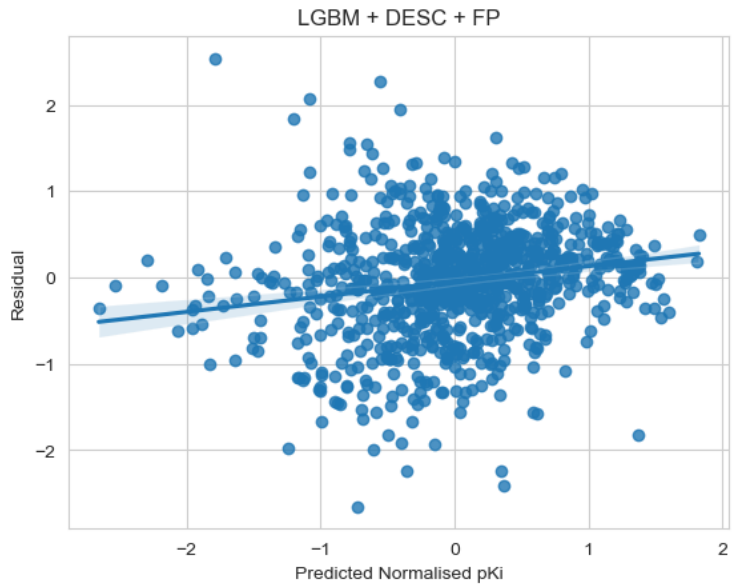




- Only use fingerprint
- More efficient

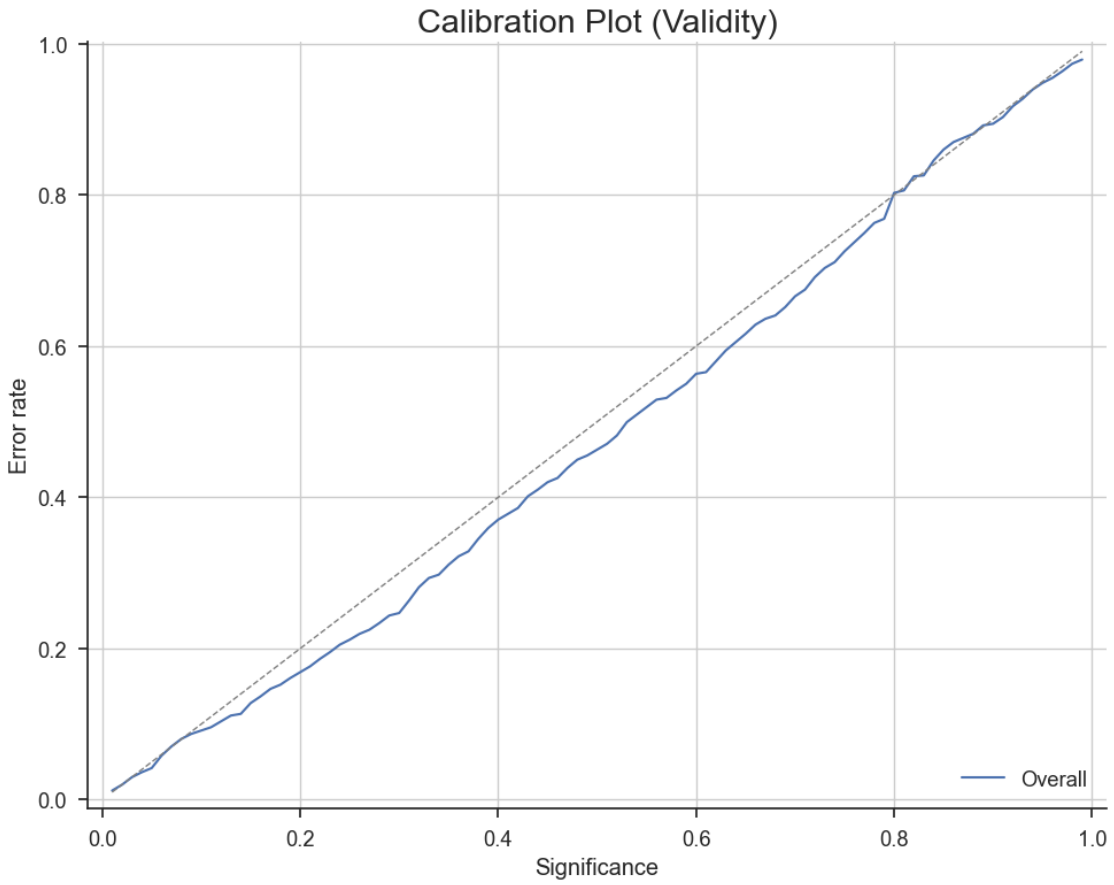
Final models: XGBoost + FP

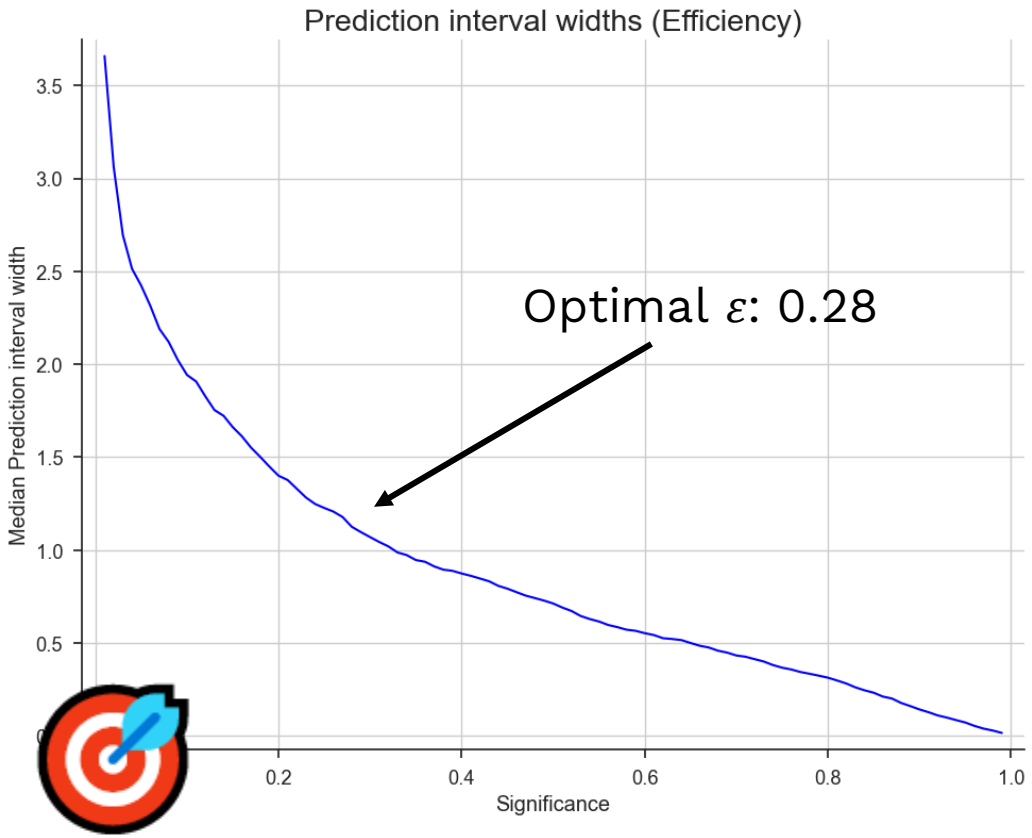




REGRESSION MODELS

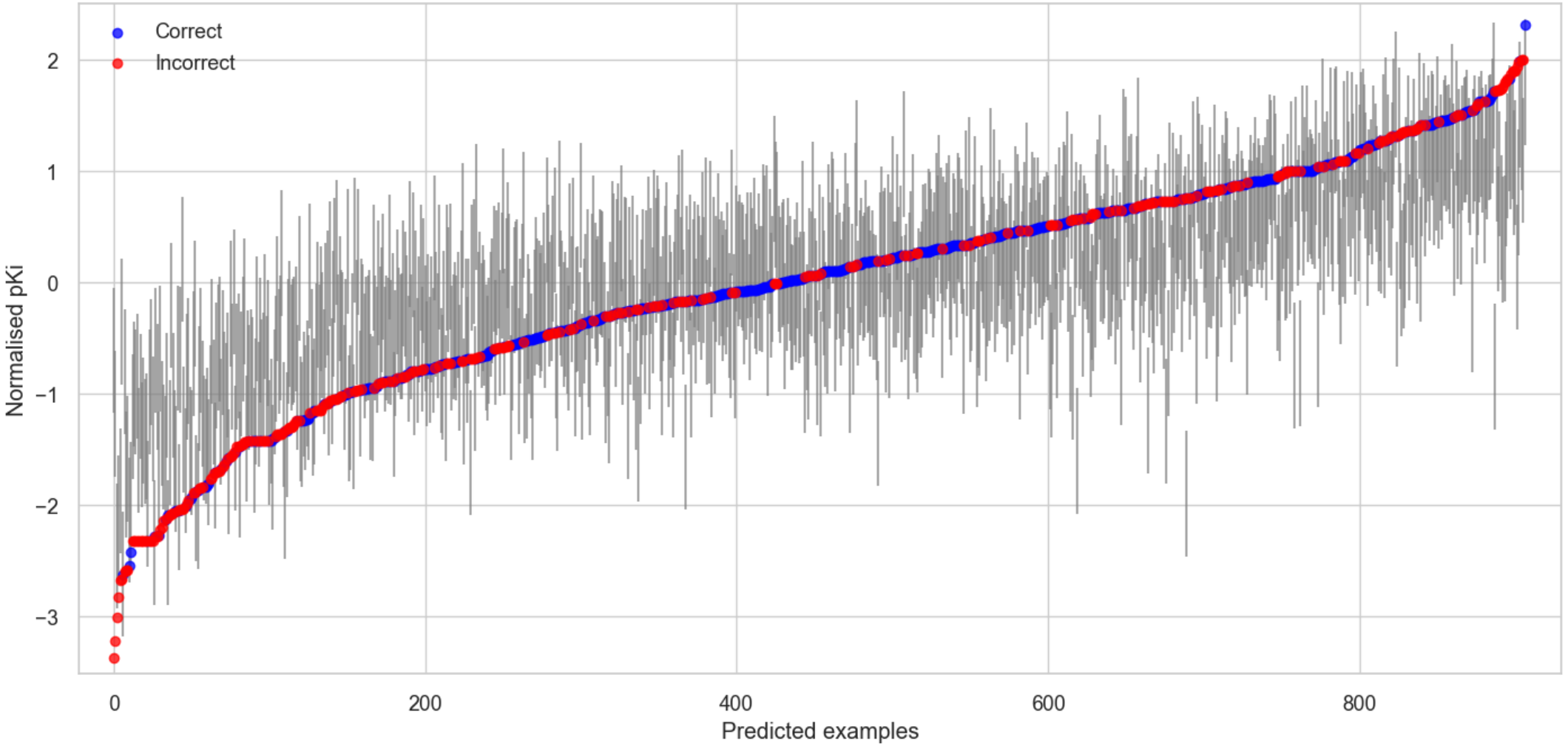
XGBoost with Fingerprint





REGRESSION MODELS

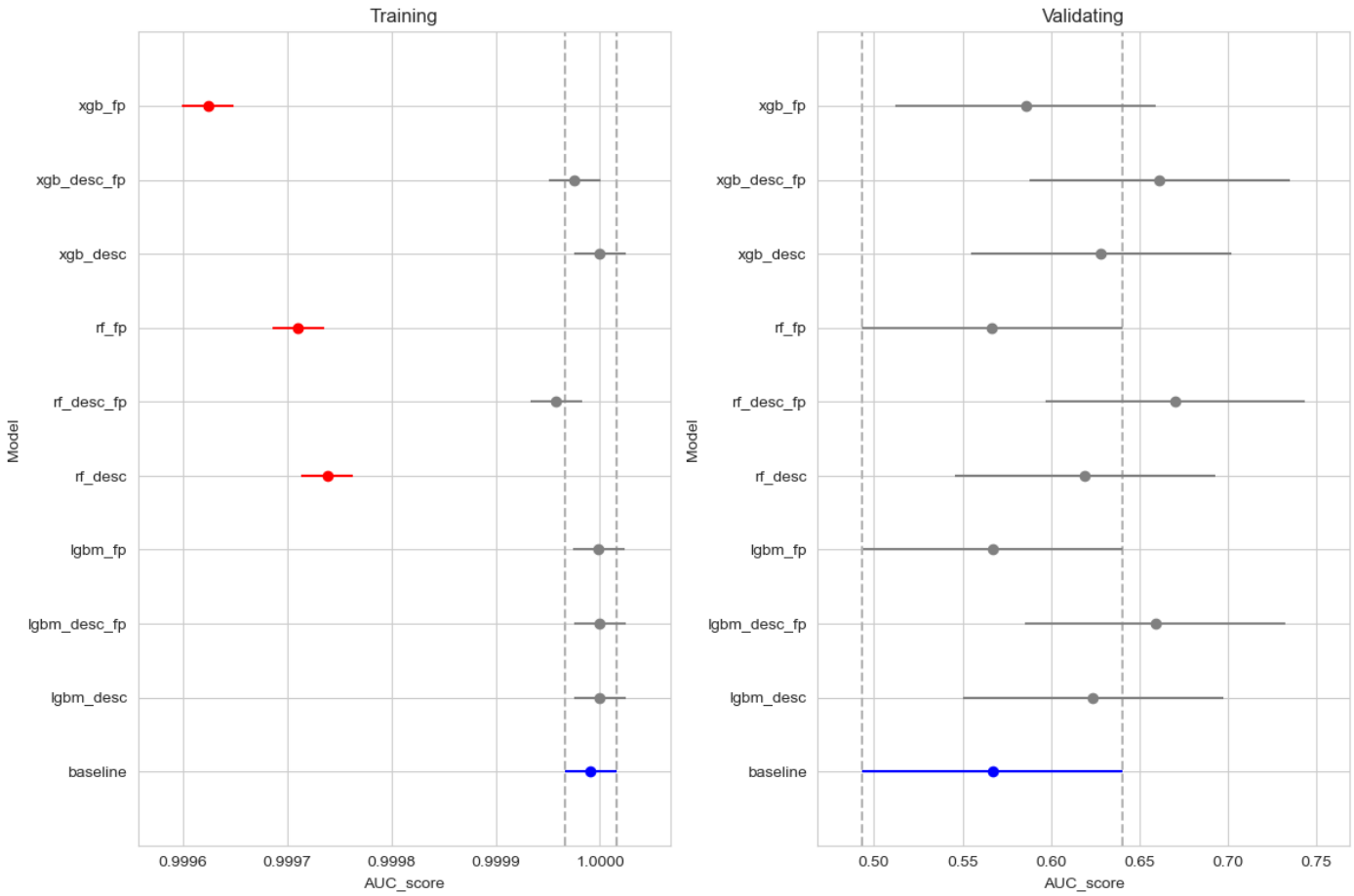
XGBoost with Fingerprint



Median interval width: 1.12

CLASSIFICATION MODELS

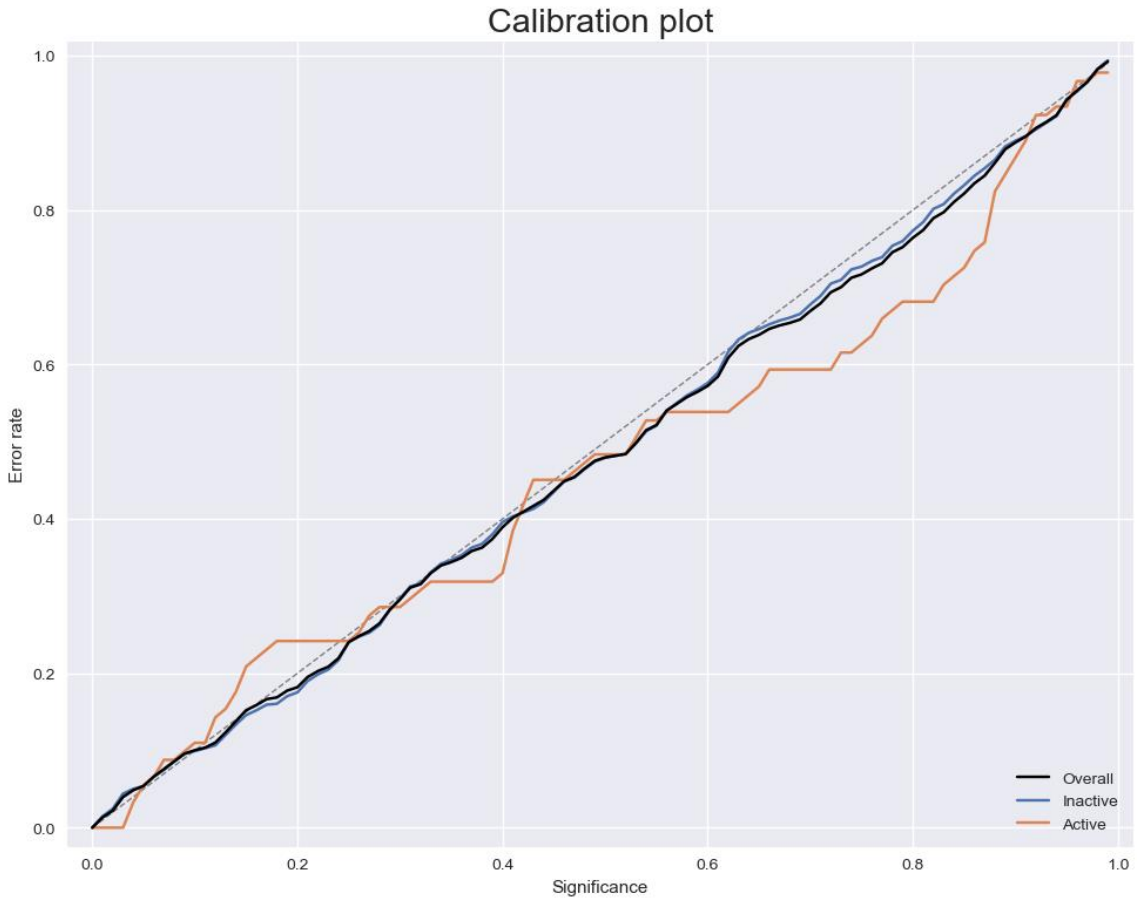
Tracking metrics: AUC-PR

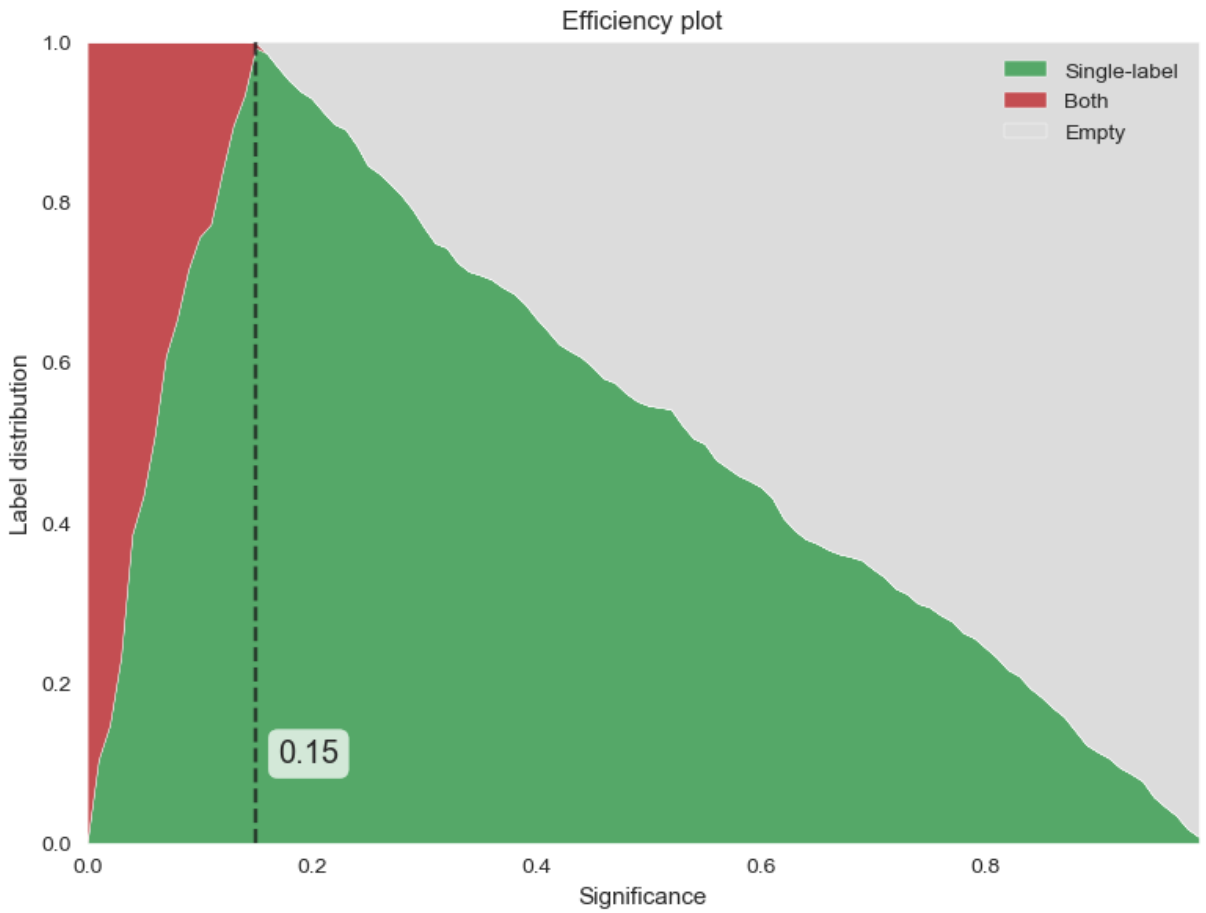


Interval from Turkey-HSD with correction

CLASSIFICATION MODELS

Random Forest with Descriptors and Fingerprint





CLASSIFICATION MODELS

Random Forest with Descriptors and Fingerprint

Count-based

Normalized (recall-oriented)

TAKE HOME MESSAGE

Data Cleaning and Context is important.

- Understand the data source and domain-specific nuances.
- Handle duplicates, missing values, and inconsistent records carefully.

Data Splitting is important

- Random splits often yield better-calibrated conformal predictors.
- Alternative methods like cluster-based splitting may harm calibration due to distribution shift.
- Always evaluate the impact of your split strategy on model calibration.

Multiple optimization is always needed

- Trade-off between validity and efficiency in CP
- Trade-off between recall and precision in classification model
- Trade-off between loss_value and overfitting in any kind of model

REFERENCE

Arvidsson McShane, S., Norinder, U., Alvarsson, J., Ahlberg, E., Carlsson, L., & Spjuth, O. (2024). **CPSign: conformal prediction for cheminformatics modeling**. *Journal of Cheminformatics, 16*(75). https://doi.org/10.1186/s13321-024-00870-9

Vovk, V., Gammerman, A., & Shafer, G. (2005). *Algorithmic learning in a random world*. Springer Science & Business Media.

Johansson, U., Ahlberg, E., Boström, H., Carlsson, L., Linusson, H., & Sönströd, C. (2015). Handling small calibration sets in Mondrian Inductive Conformal Regressors. In *Proceedings of the 3rd International Symposium on Statistical Learning and Data Sciences (SLDS 2015)*. (In press).

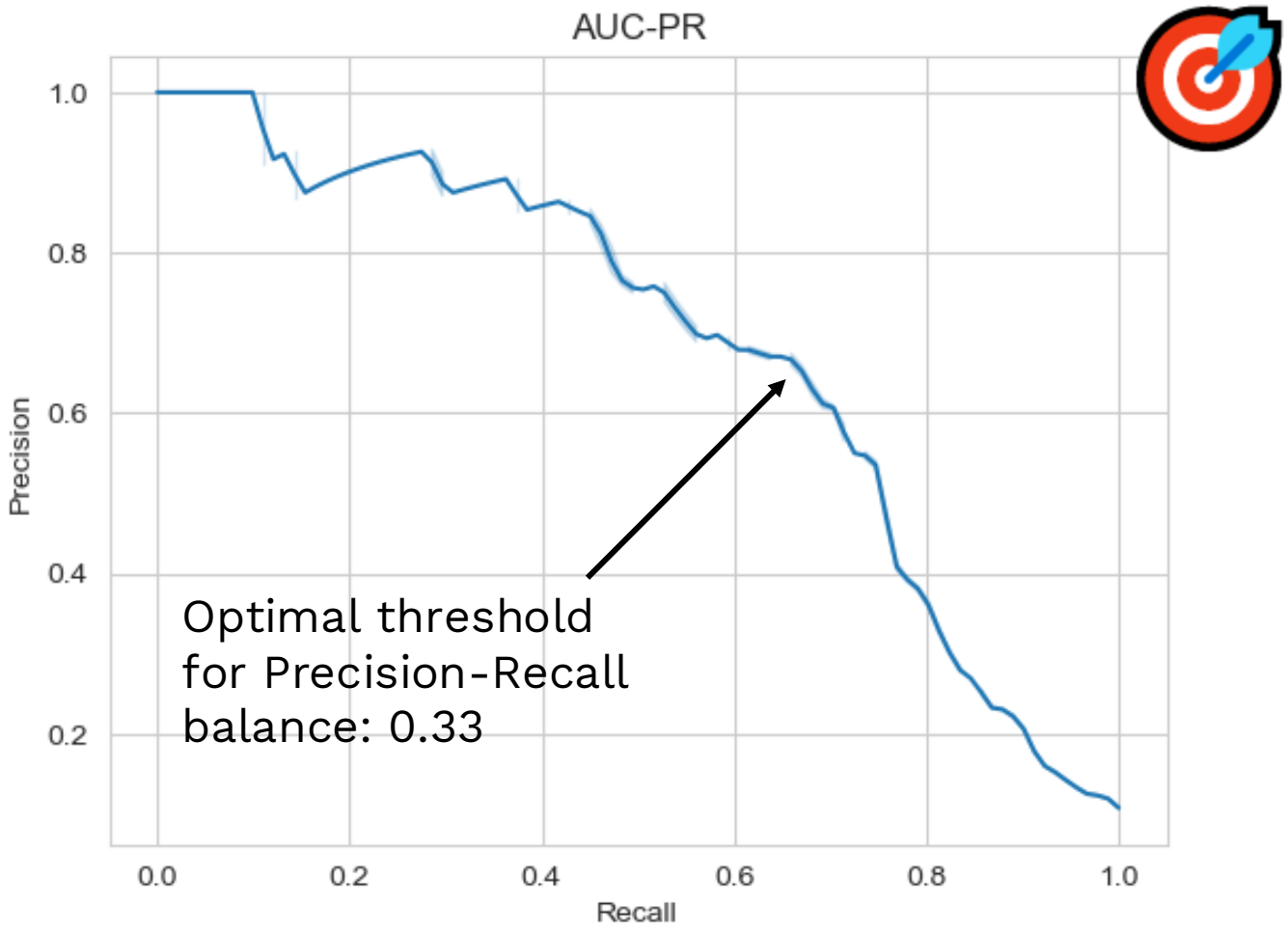
Pharmbio/plot_utils (n.d.). *plot_utils: Utility scripts for plotting*. GitHub repository. https://github.com/pharmbio/plot_utils

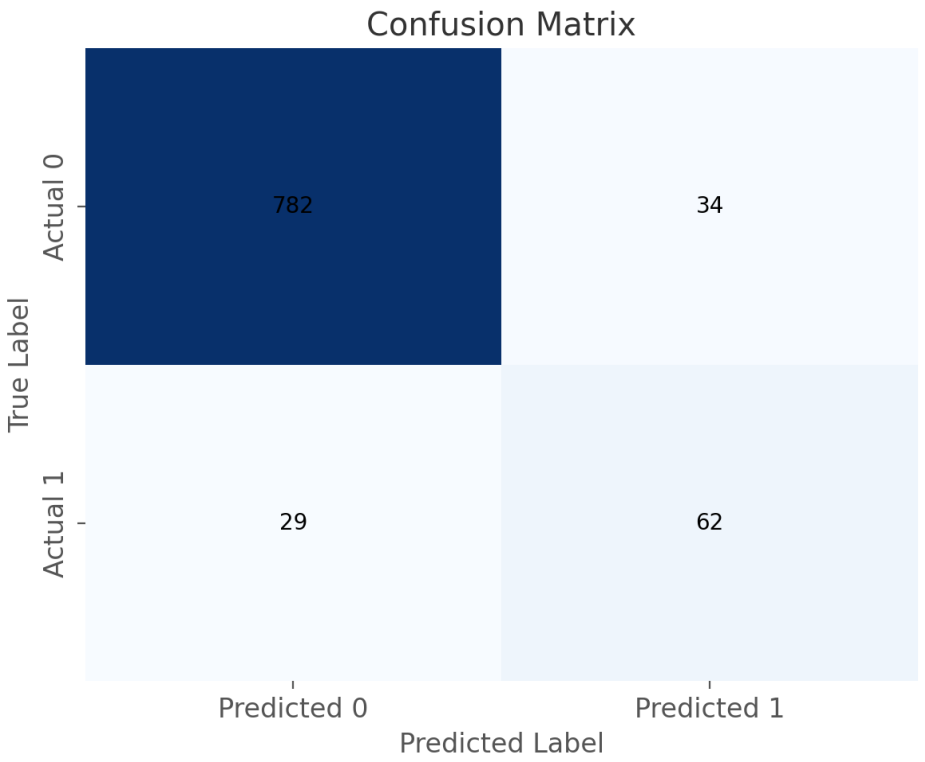
Thank you for listening

Feedbacks and Questions

CLASSIFICATION MODELS

Random Forest with Descriptors and Fingerprint





At threshold 0.33