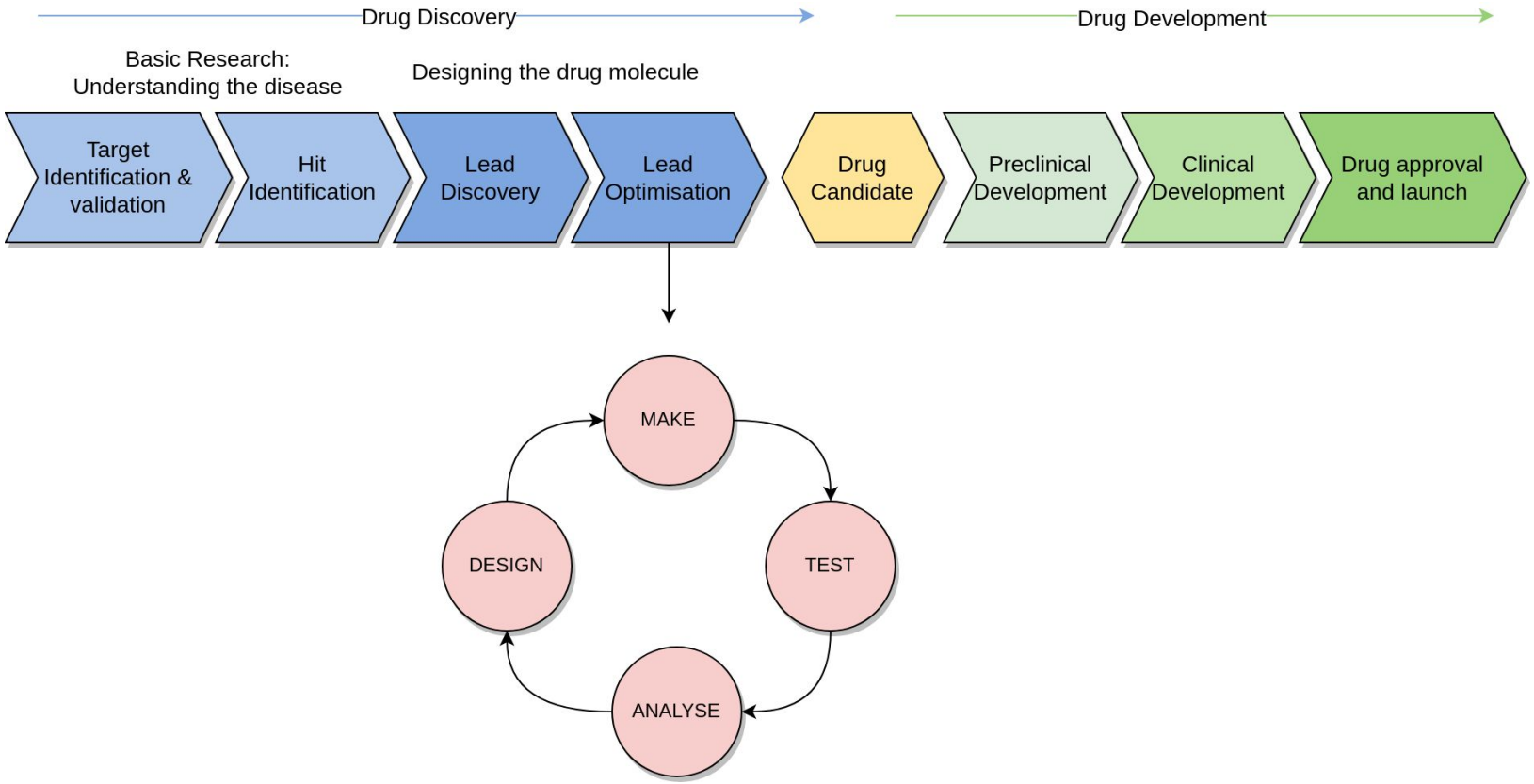
Predicting CYP inhibitions

Olatunde James Omoya
Jonna Marie Matthiesen

Drug Discovery and Development



Predicting CYP inhibitions

Goals

- Reduce cost of drug discovery process
- Reduce failure rate in later stages

Project: Prediction of Cytochrome P450s inhibitions

- Single-task binary classification using classical machine learning approaches
- CYP2C19, CYP2D6, CYP3A4, CYP1A2, CYP2C9
- Both, balanced and imbalanced datasets

Features

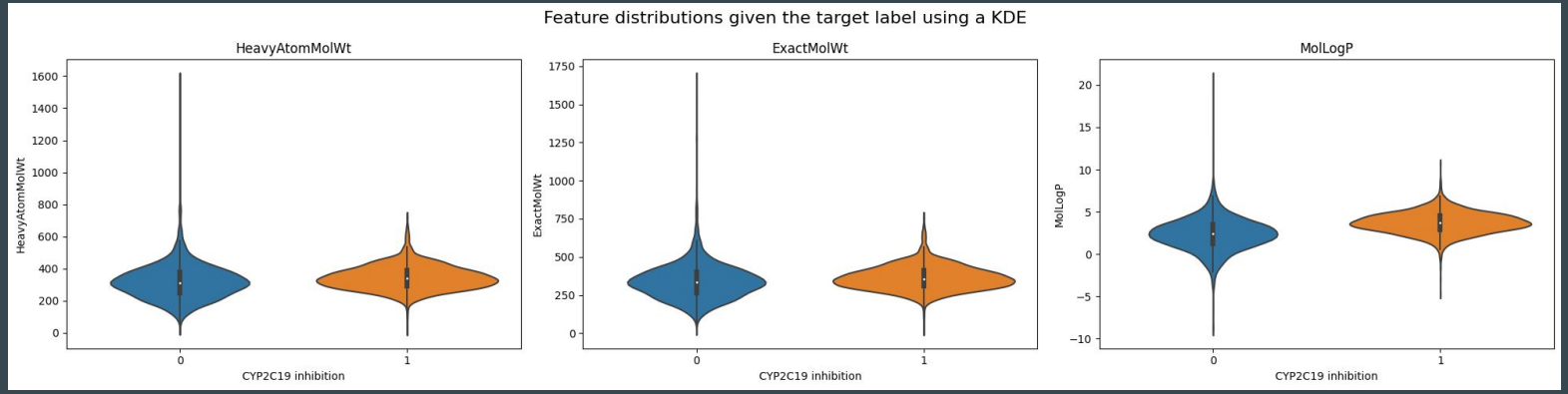
Molecular Descriptors

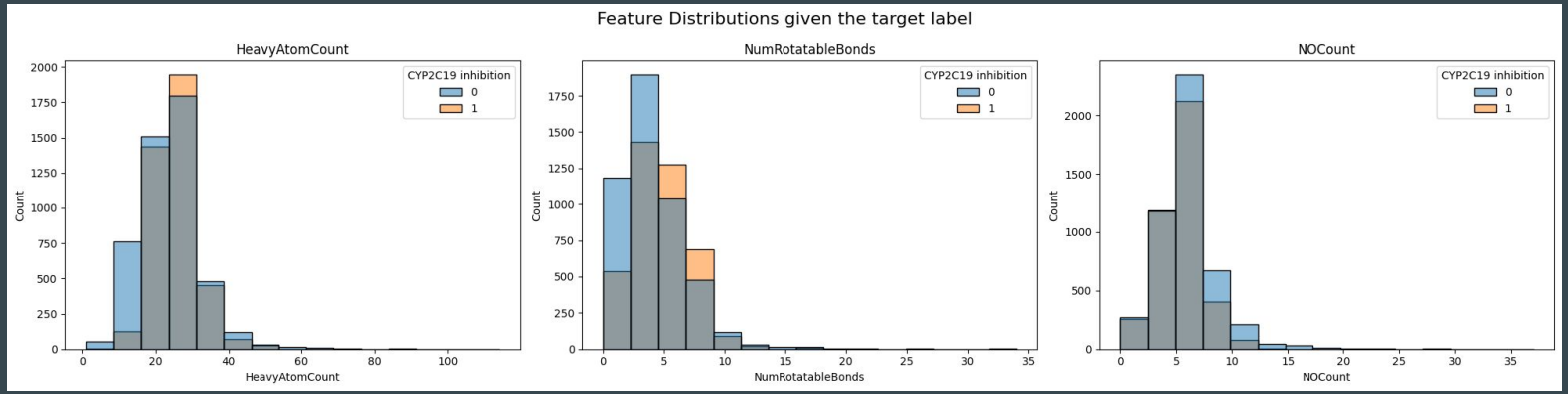
- 208 1D/2D descriptors calculated using RDKit
- HeavyAtomMolWt, HeavyAtomCount, ExactMolWt, MolLogP, NumRotatableBonds, NOCount

Molecular Fingerprints

- First: ECFP-4 (Morgan Fingerprints)
- Later: Exploration and comparison of other fingerprints
    - E.g., MACCS, Atom-Pair and 2D Pharmacophore Fingerprints

Feature Exploration

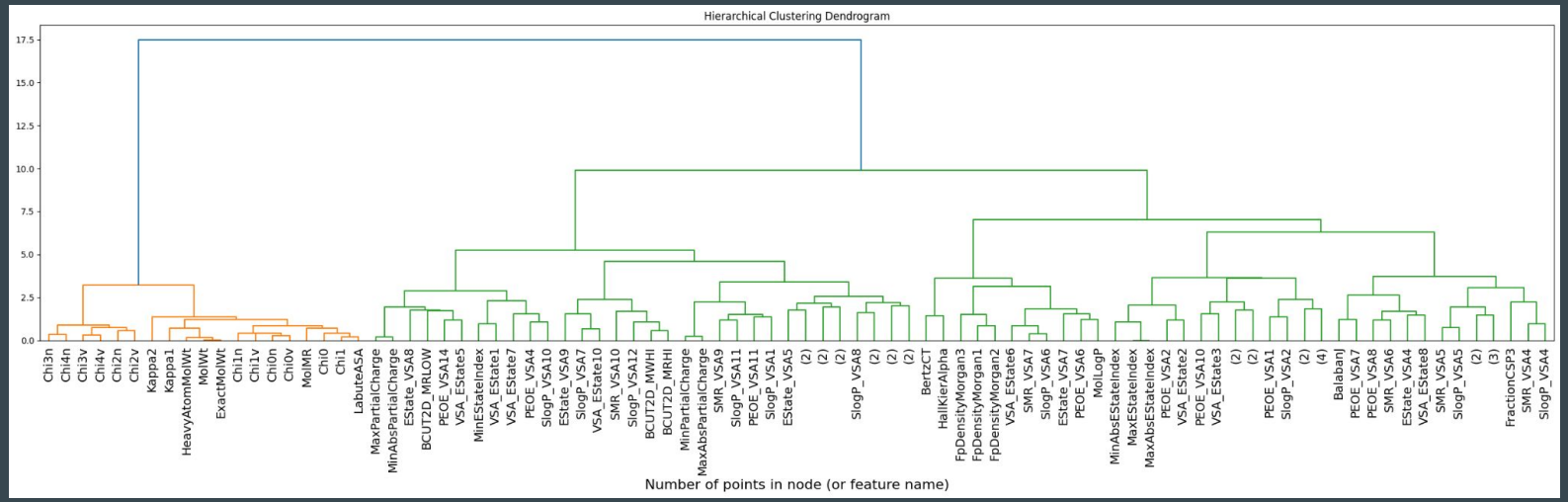




Feature Selection

- (Zero-)Variance Threshold - Removing (quasi-)constant features
- Correlation based feature selection for **continuous descriptors**
- Univariate statistical tests for **discrete descriptors and fingerprints**
  - mutual information
- Model-based feature selection
  - L1-metric
  - Tree-based model

Correlation-based Feature Selection



Feature Selection

- (Zero-)Variance Threshold - Removing (quasi-)constant features
- Correlation based feature selection for **continuous descriptors**
- Univariate statistical tests for **discrete descriptors and fingerprints**
  - mutual information
- Model-based feature selection
  - L1-metric
  - Tree-based model

Feature Normalization

- MinMaxNormalization - Scaling features to the range of [0, 1]
- StandardScaler for **continuous features**

Feature Selection

- (Zero-)Variance Threshold - Removing (quasi-)constant features
- Correlation based feature selection for **continuous descriptors**
- Univariate statistical tests for **discrete descriptors and fingerprints**
  - mutual information
- Model-based feature selection
  - L1-metric
  - Tree-based model

Feature Normalization

- MinMaxNormalization - Scaling features to the range of [0, 1]
- StandardScaler for **continuous features**

Dimensionality Reduction

- [Kernel] Principal component analysis (PCA) for **continuous descriptors**

Predicting CYP inhibitions

Olatunde James Omoya
Jonna Marie Matthiesen