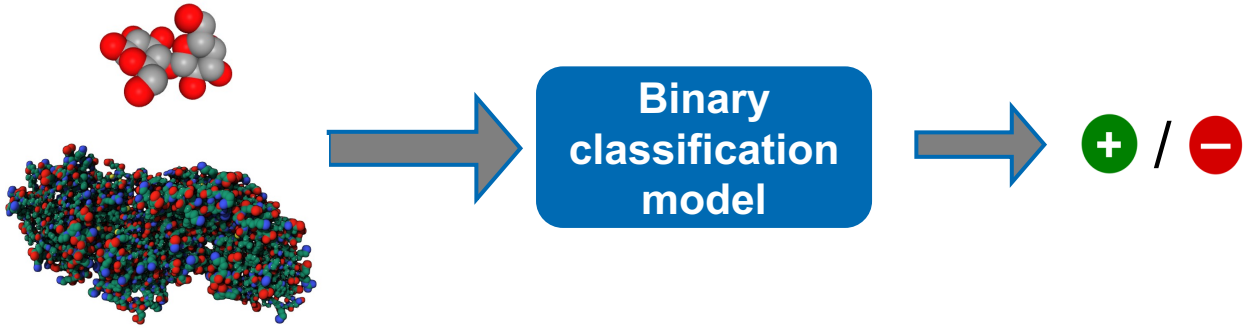
Heinrich Heine
Universität Düsseldorf

Challenges and best practices

Requirement of negative data

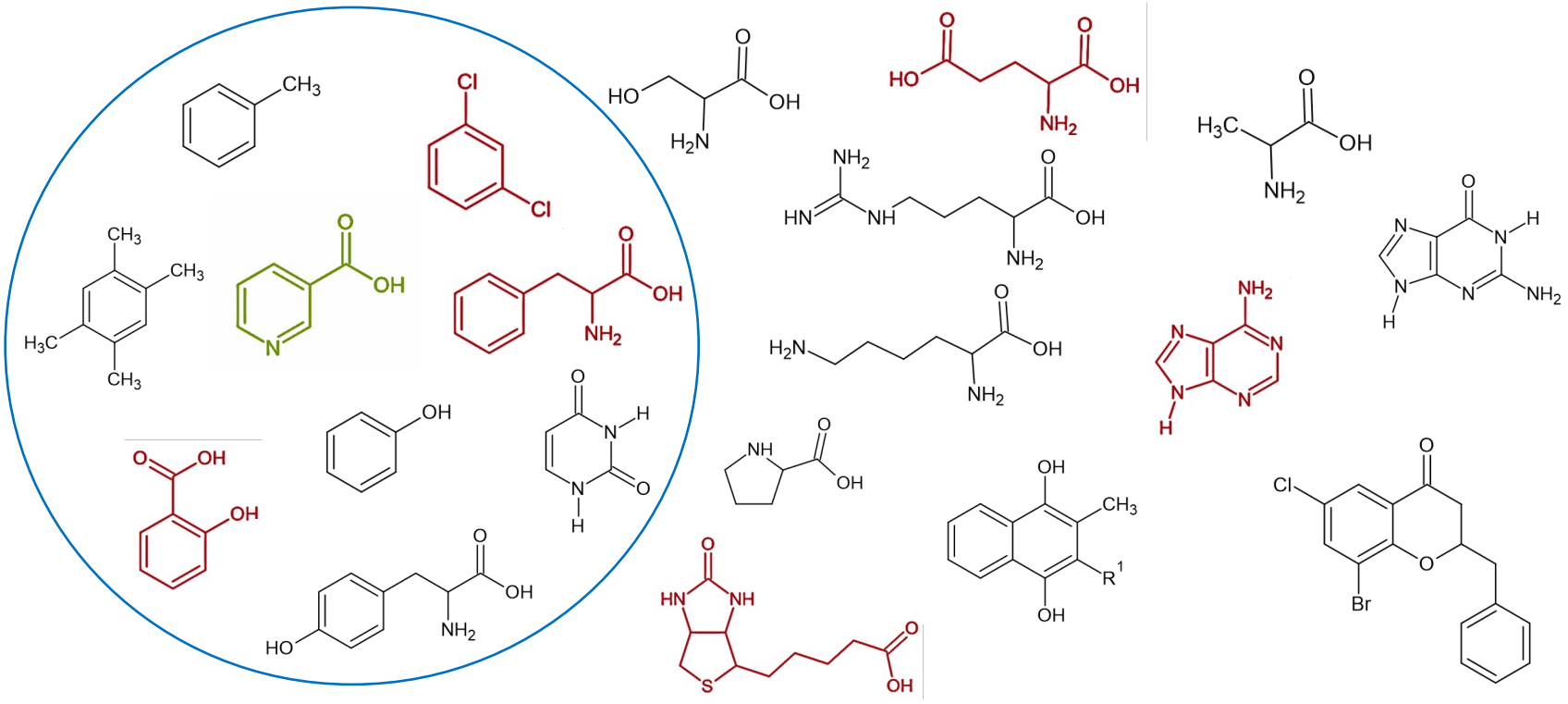
- When training binary classification models, we need positive and negative training data



- Example: Enzyme-substrate pair prediction
  - Experimentally verified enzyme-substrate pairs are stored in protein databases
  - We also need negative data, i.e., enzyme-non substrate pairs
    - Negative data is typically not stored in large biological databases

Small molecule image: pubchem.ncbi.nlm.nih.gov/compound/6134#section=3D-Conformer; Protein image generated with: https://www.rcsb.org/3d-view

How to add negative samples to the data set?

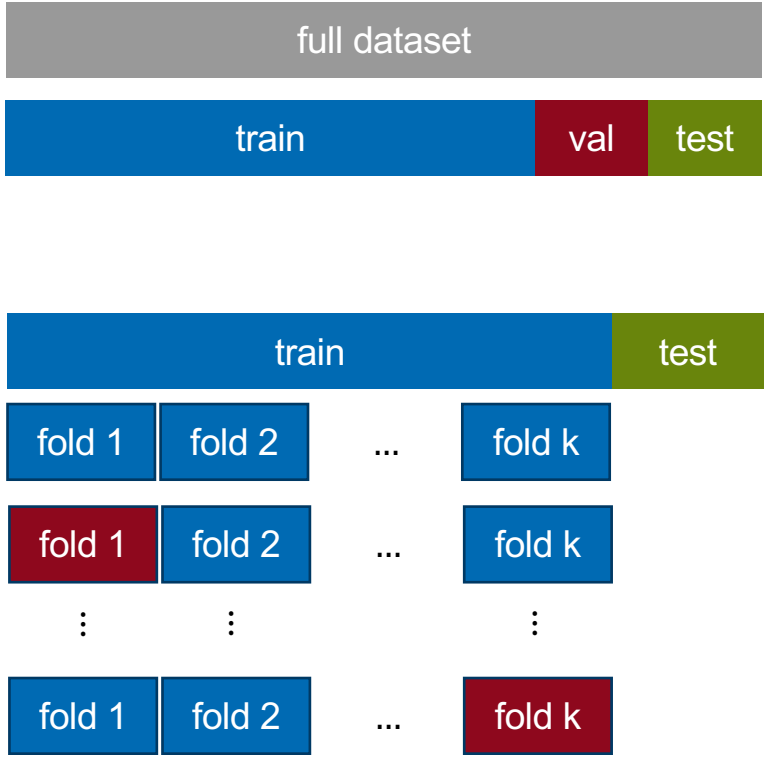


Summary – Sampling Negative Data Points

- When we can determine negative data points without experiments with a high likelihood, we can sample them
  - We might generate some false negative data points
- We can sample negative data points similar to the positive ones (hard negatives) to make the prediction task more difficult
- Negative-to-positive ratio can be treated as a hyperparameter

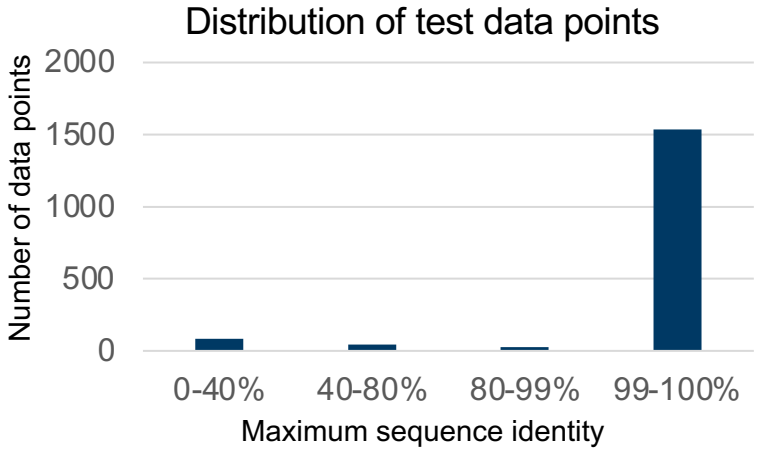
Splitting dataset (1)

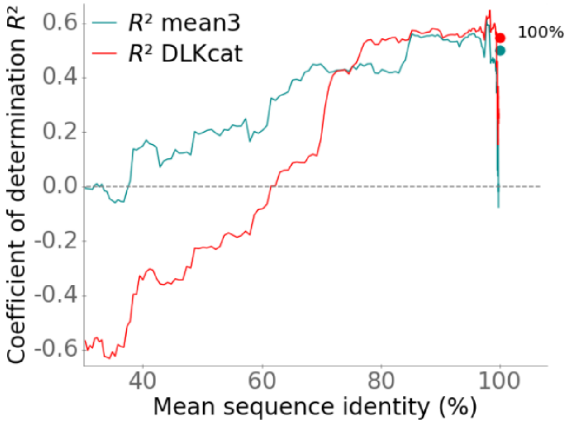
- Options for splitting the dataset:
  - Training, validation, and test set
    - Common fractions: 70/15/15 or 80/10/10
    - Larger datasets –> We can assign smaller fractions to validation and test sets
    - Hyperparameter optimization / model selection:
      - Train on training set & validate on validation set
      - Evaluate final model performance on test set
  - Perform k-fold cross-validation for smaller datasets:
    - Divide data in training and test set
    - Divide training data in k equally sized folds
    - Hyperparameter optimization / model selection:
      - Train on all folds except for one and validate on the remaining fold
      - Repeat k times and calculate average metric scores
      - Evaluate final model performance on test set



Splitting dataset (2)

- Standard in many ML domains to randomly split data
- When splitting protein datasets randomly, many test proteins can be identical or nearly identical to the training proteins
  - Prediction task becomes too easy
  - Evaluations on the test set are overly optimistic
- Example (turnover number k_{cat} prediction):





kcat study: https://www.nature.com/articles/s41929-022-00798-z; Right figure extracted from: https://doi.org/10.1101/2023.02.06.526991

Splitting dataset (3)

- How to split datasets protein dataset?
  - CD-HIT clustering algorithm
    - Creates clusters of sequences that have a sequence identity above a certain threshold

- Splitting datasets with small molecules
  - Jaccard distance between binary molecular fingerprints:
    - The proportion of elements that do not match, considering only those entries where at least one entry is non-zero.

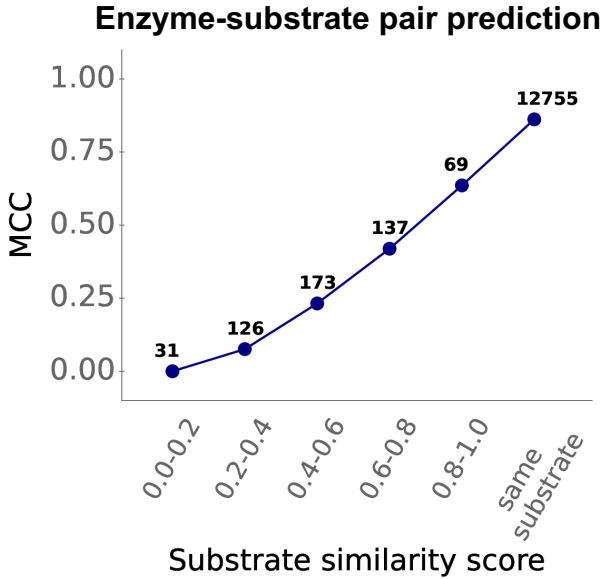


Image extracted from: https://doi.org/10.1371/journal.pcbi.1012100