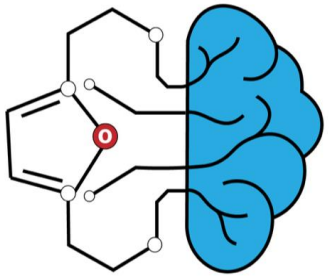
Demographic-Sensitive Cardiotoxicity Prediction

Lausanne, 25.04.2025

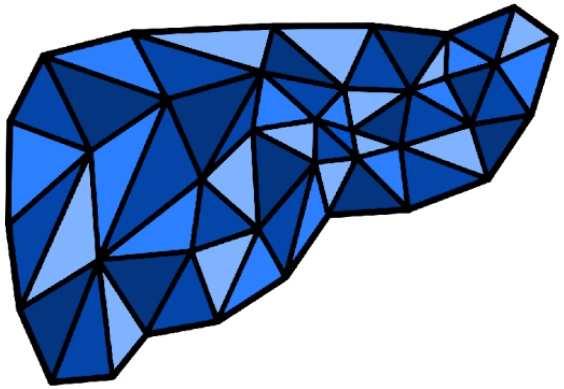


TU/e

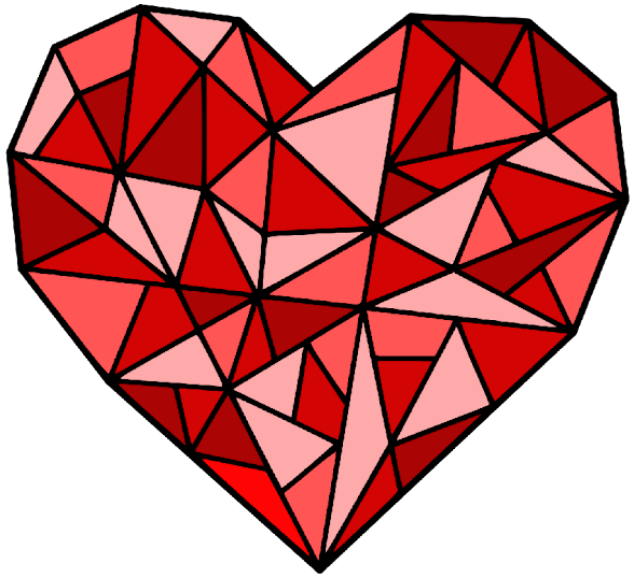
ISTITUTO DI RICERCHE FARMACOLOGICHE MARIO NEGRI · IRCCS

BAYER

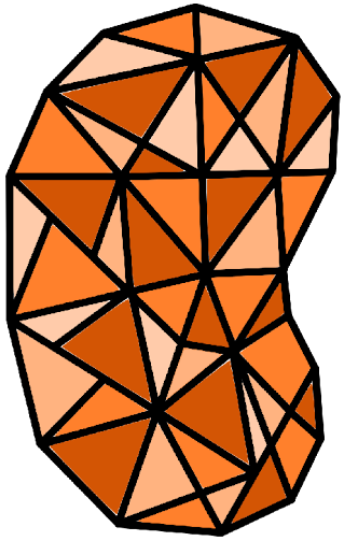
Advanced ML methods to predict and understand the toxicity of drugs



Hepatotoxicity



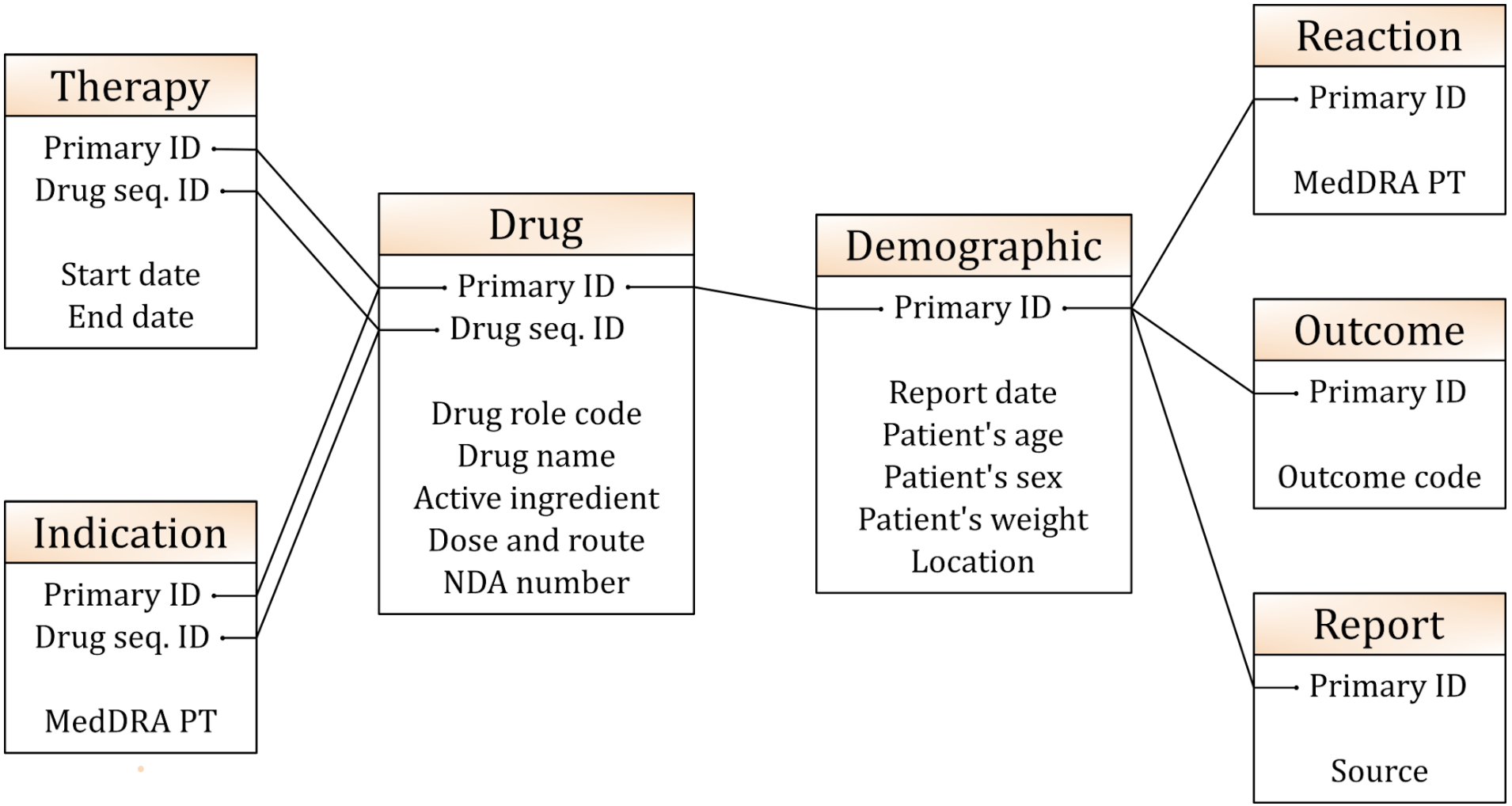
Cardiotoxicity



Nephrotoxicity

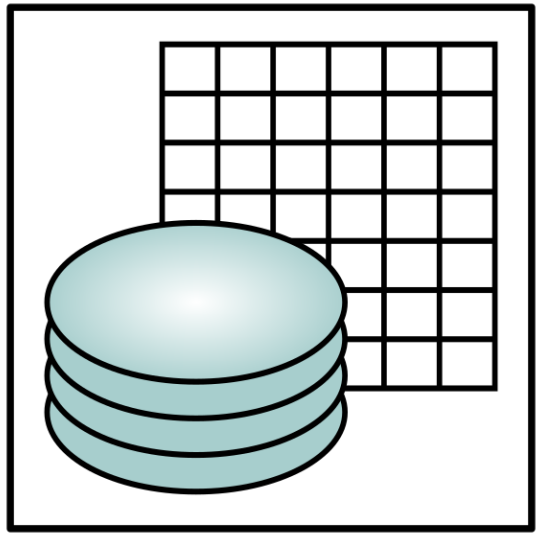
Dataset preparation

FAERS database[1]



Some statistics

- Data collection: Q4 2012 – Q3 2024
- Number of unique reports: 17,687,672
- Number of unique drug descriptions: 591,402
- Number of unique adverse effects: 35,966
- Data completeness:
  - Sex: 87.2%
  - Age: 57.2%
  - Weight: 18.9%



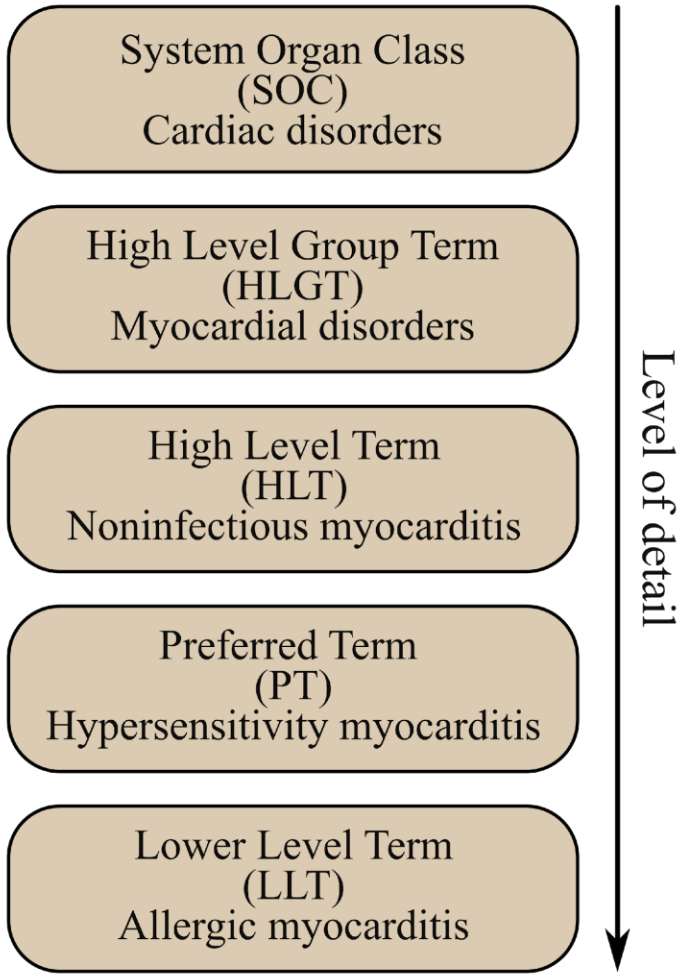
MedDRA terms selection[2]

- Standardized medical terminology developed by International Conference for Harmonization (ICH)

- Selected HLGT:
  - Cardiac arrhythmias
  - Myocardial disorders
  - Heart Failures
  - Pericardial / Endocardial disorders
  - Coronary artery disorders
  - Cardiac disorders, signs, and symptoms NEC

- Removed PTs:
  - Mechanical injuries / complications
  - Congenital / infectious conditions

- Total number of descriptions: 123,958



Drug descriptions processing

- Unified within a class
- Offline access to DrugBank[3], PME[4], ChEMBL[5]
- Automated queries to several online databases and resources
- Revertible changes and history storage
- Several options for regex-based string processing, and string similarity searches

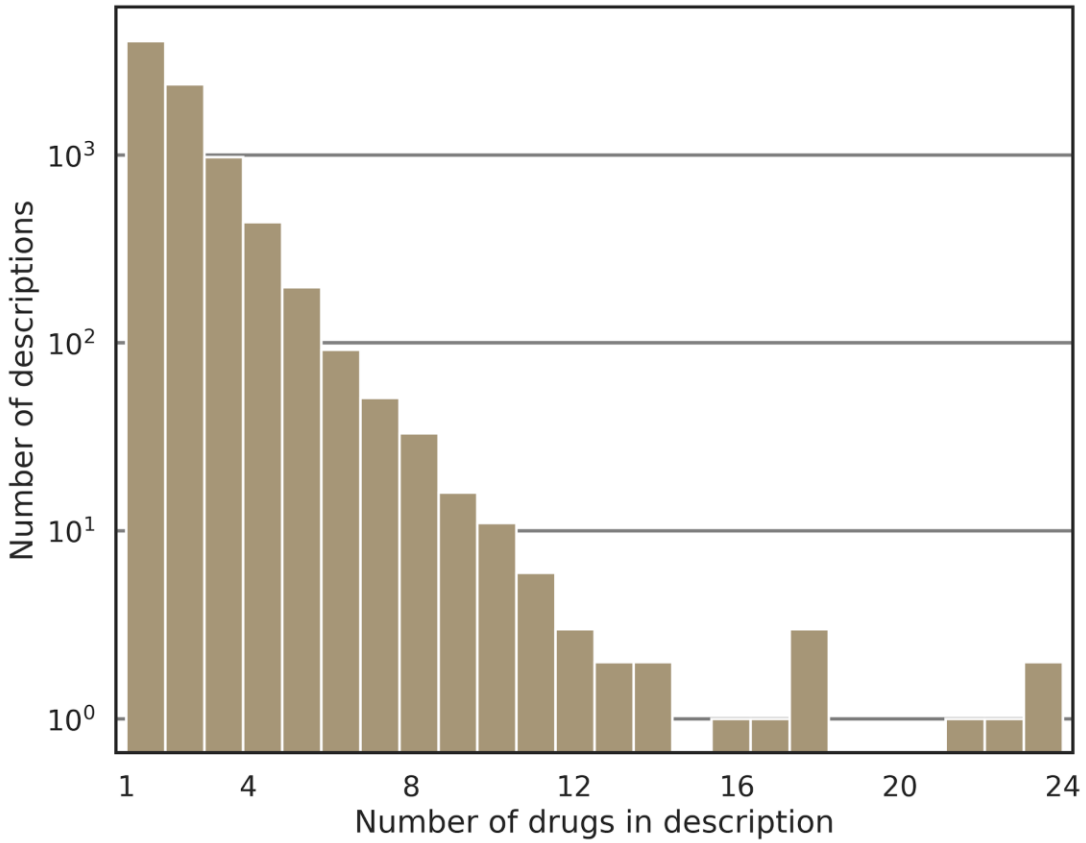
Algorithm 1 Token processing

```
procedure PROCESS TOKENS(tokens)
    for token in tokens do
        if token in DrugBank then
            capture
        else
            calculate similarity to drugs and synonyms in DrugBank
            calculate similarity to previous tokens
            query external databases            ▷ e.g. PubChem, PME, RxReasoner

            decision ← input()
            if decision == 'remove' then
                remove token
            else if decision == 'substitute' then
                replace token with user-provided string
            else if decision == 'update' then
                add new information to the token
            else if decision == 'capture' then
                capture
            else
                skip
            end if
        end if
    end for
end procedure
```

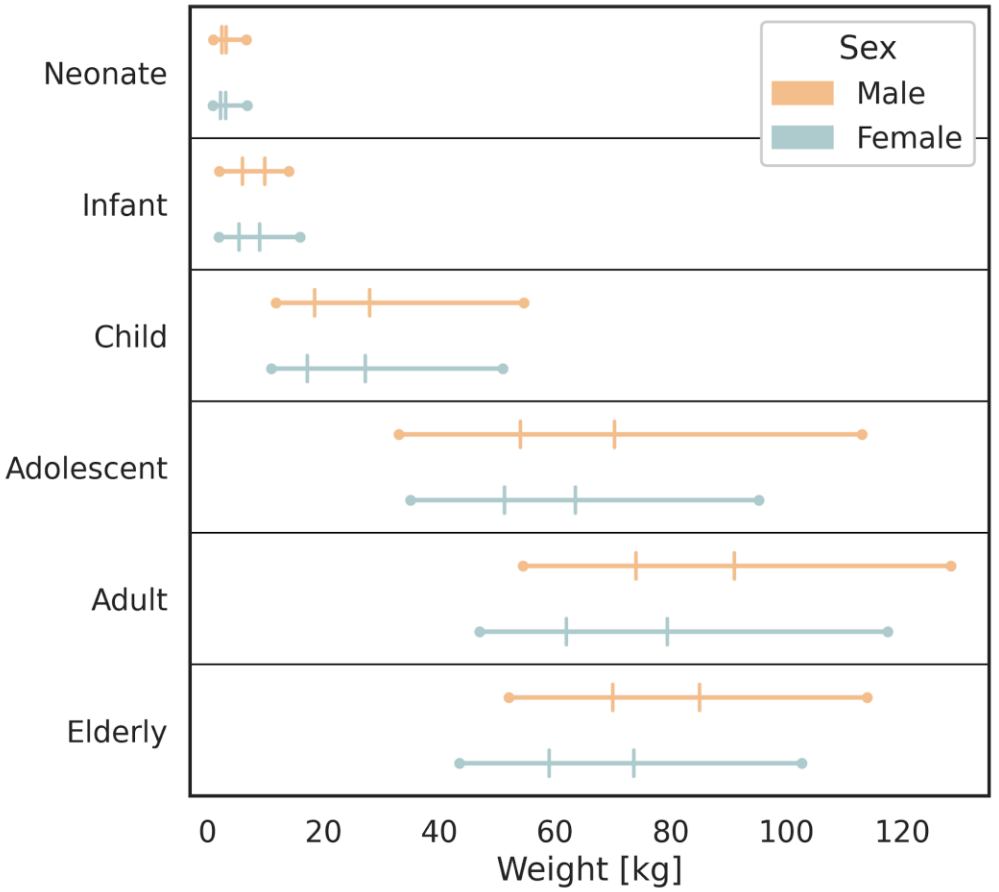
Drug descriptions processing

- Approximately 85% of drug descriptions were assigned active ingredient(s)
- Removed: vaccines, immunoglobulins, RNA-based drugs, peptides, proteins, polymers, probiotics, herbal and homeopathic formulations, infusion or dialysis fluids, multivitamins, foods, nutritional preparations, unclear abbreviations, entries with contradictory results
- Additional string similarity based full record linkage using prepared mapping and remaining entries
- Final drug descriptions – active ingredients mapping statistics:
  - 311,451 drug descriptions with assigned actives
  - 8,260 drug combinations
  - 4,333 unique drugs



Demographic data processing

- Sex was used without further processing
- Age was binned using FDA classification:
  - Neonate (birth - 1 month)
  - Infant (1 month – 2 years)
  - Child (2-12 years)
  - Adolescent (12-21 years)
  - Adult (21-65 years)
  - Elderly (65-100 years)
- Weight was binned based on quantiles:
  - Low ($Q_{0.05}$ - $Q_{0.33}$)
  - Average ($Q_{0.05}$ - $Q_{0.67}$)
  - High ($Q_{0.67}$ - $Q_{0.95}$)

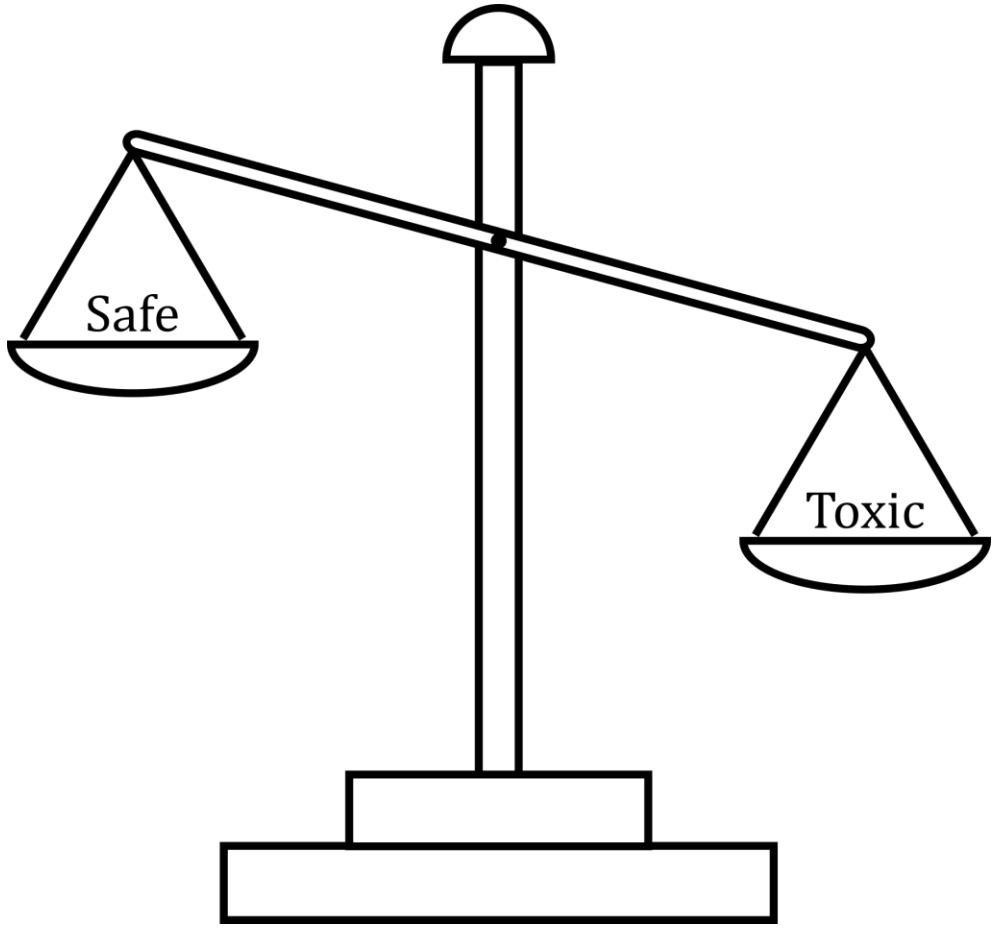


Disproportionality Analysis

Disproportionality Analysis

- Disproportionality Analysis:
  - Used for early detection of potential ADRs
  - Statistical analysis of number of reported drug-reaction cases vs expected number

- Frequently used metrics:
  - Proportional Reporting Rate[6] (PRR)
  - Reporting Odds Ratio[7] (ROR)
  - Information Component[8] (IC)



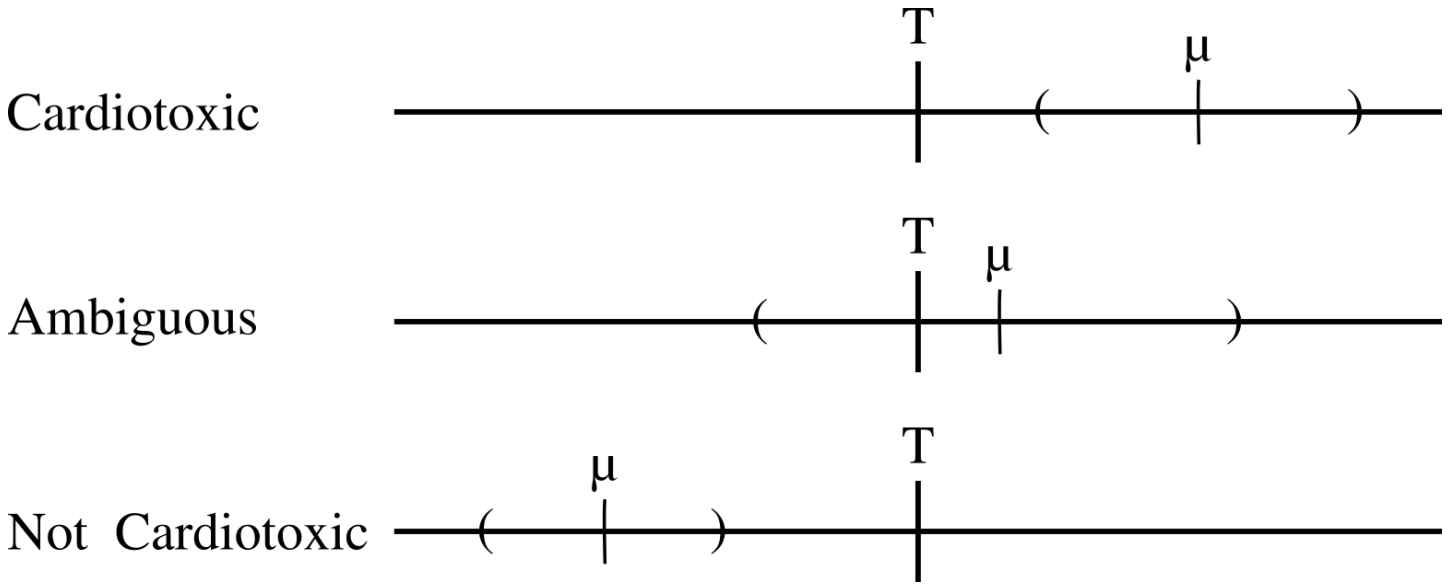
DPA metrics and label assignment

$$PRR = \frac{a\ /\ (a+b)}{c\ /\ (c+d)}$$

$$ROR = \frac{a\ /\ b}{c\ /\ d}$$

$$IC = \log_2\left(\frac{a+\kappa}{N_{exp}+\kappa}\right)$$

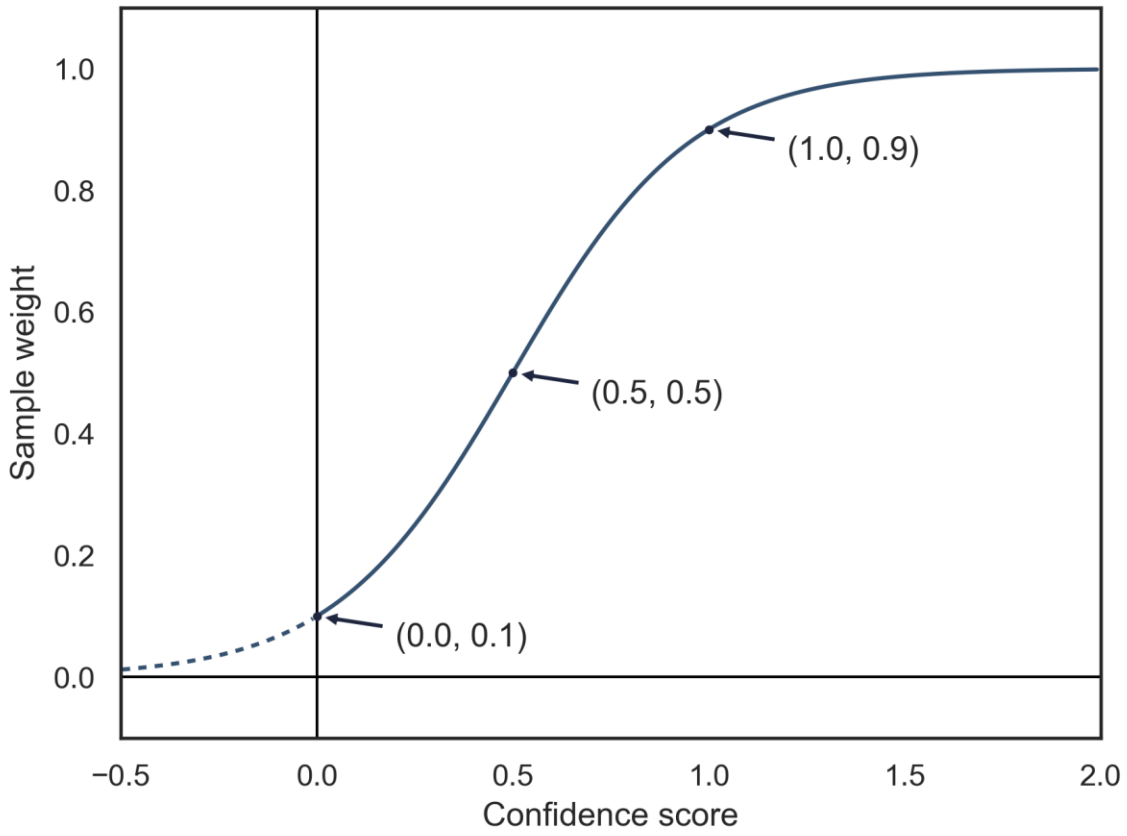
	Event Present	Event Missing
Drug Present	a	b
Drug missing	c	d



Label confidence scores

$$\text{Confidence score} = \frac{\mid \mu - T \mid}{CI_{upper} - CI_{lower}}$$

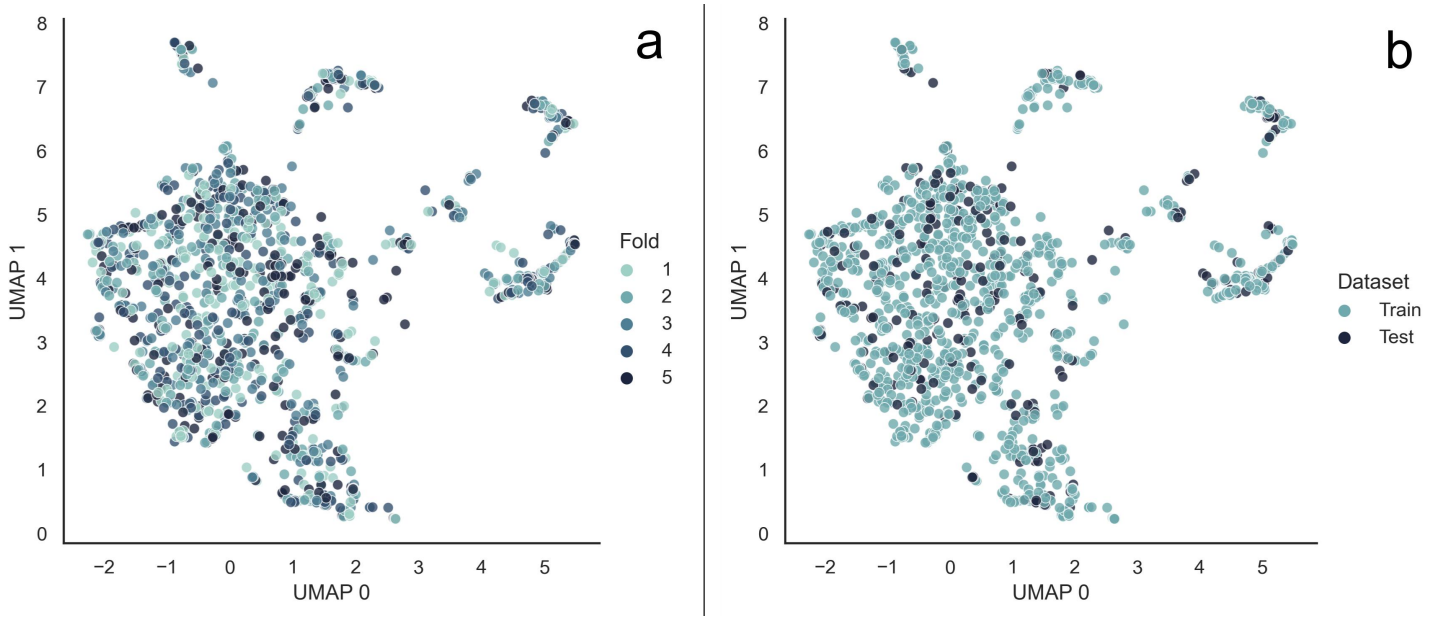
$$y = f(x) = \frac{1}{1 + e^{-2 \times ln9 \times (x - \frac{1}{2})}}$$

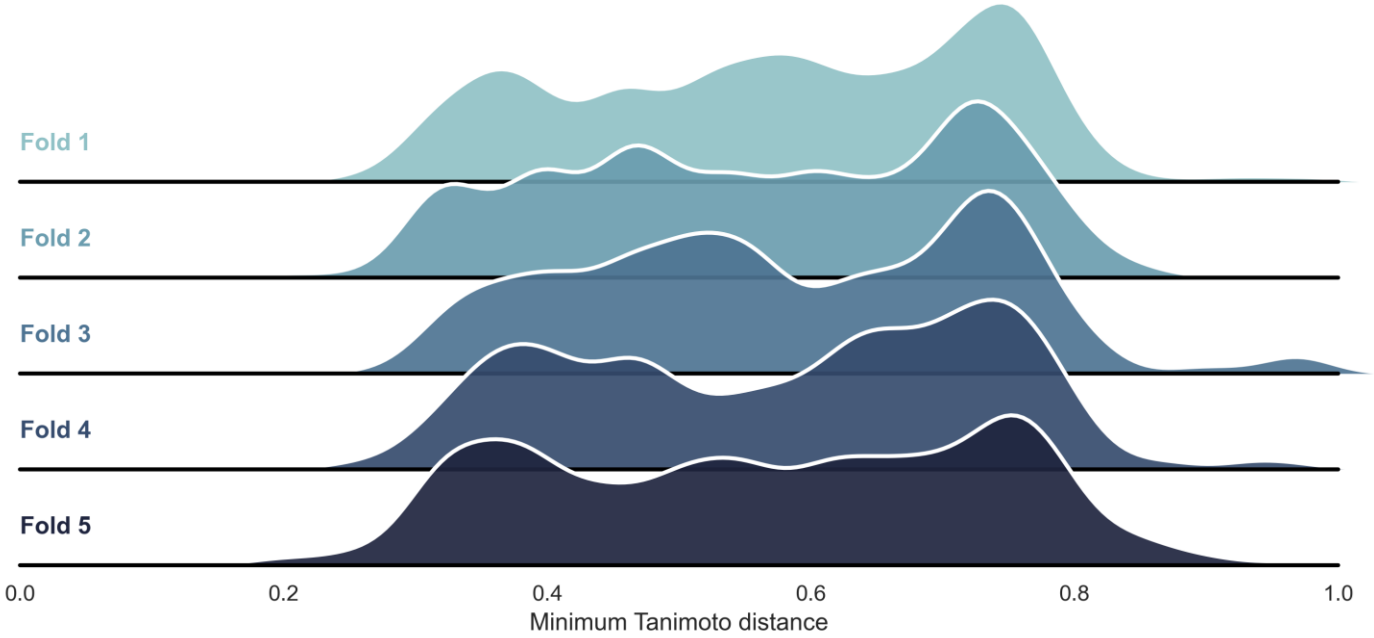


Initial models

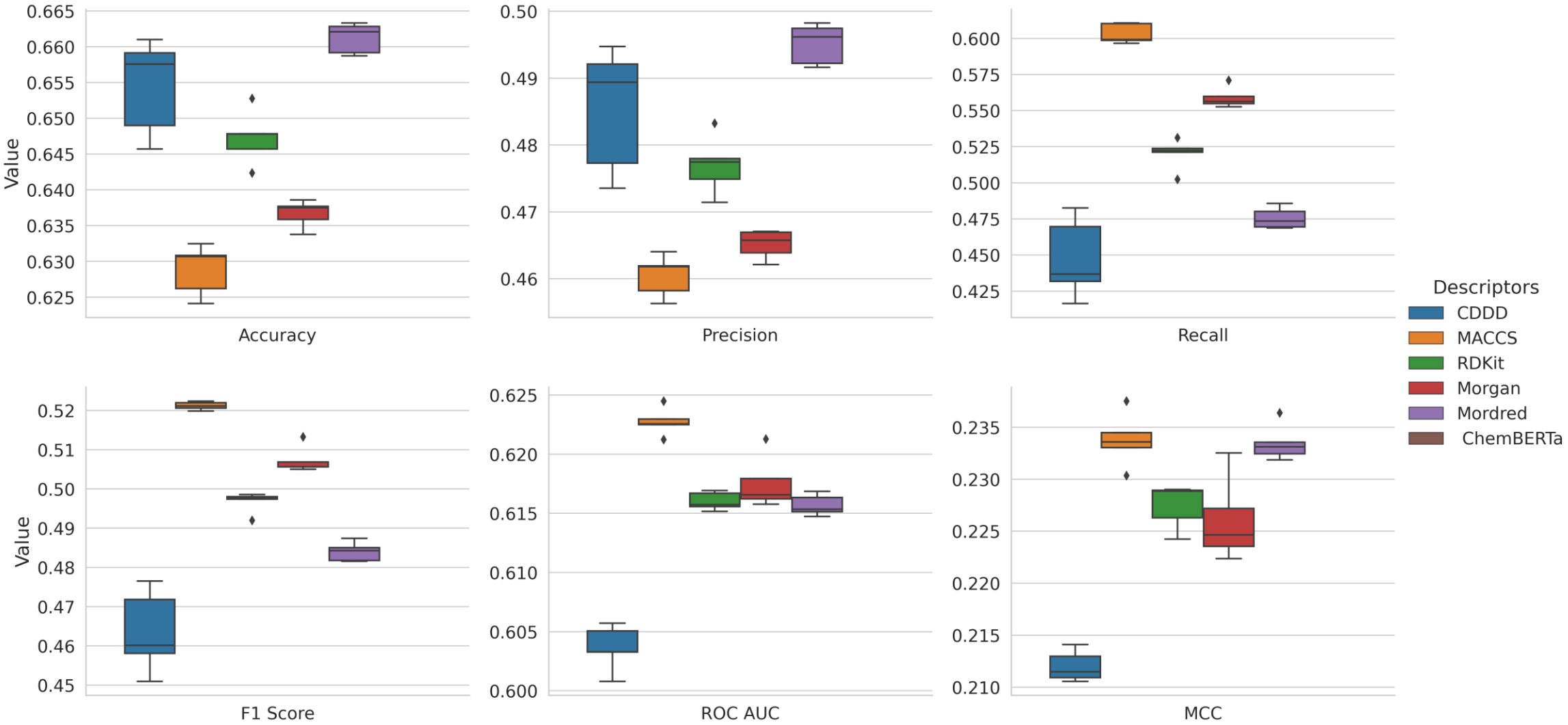
Train-test split

- Stratified Group 5-Fold split[9] :
  - Based on Butina Clustering[10]
  - Using Morgan Fingerprints[11]
  - Stratified on Class and demographic factors

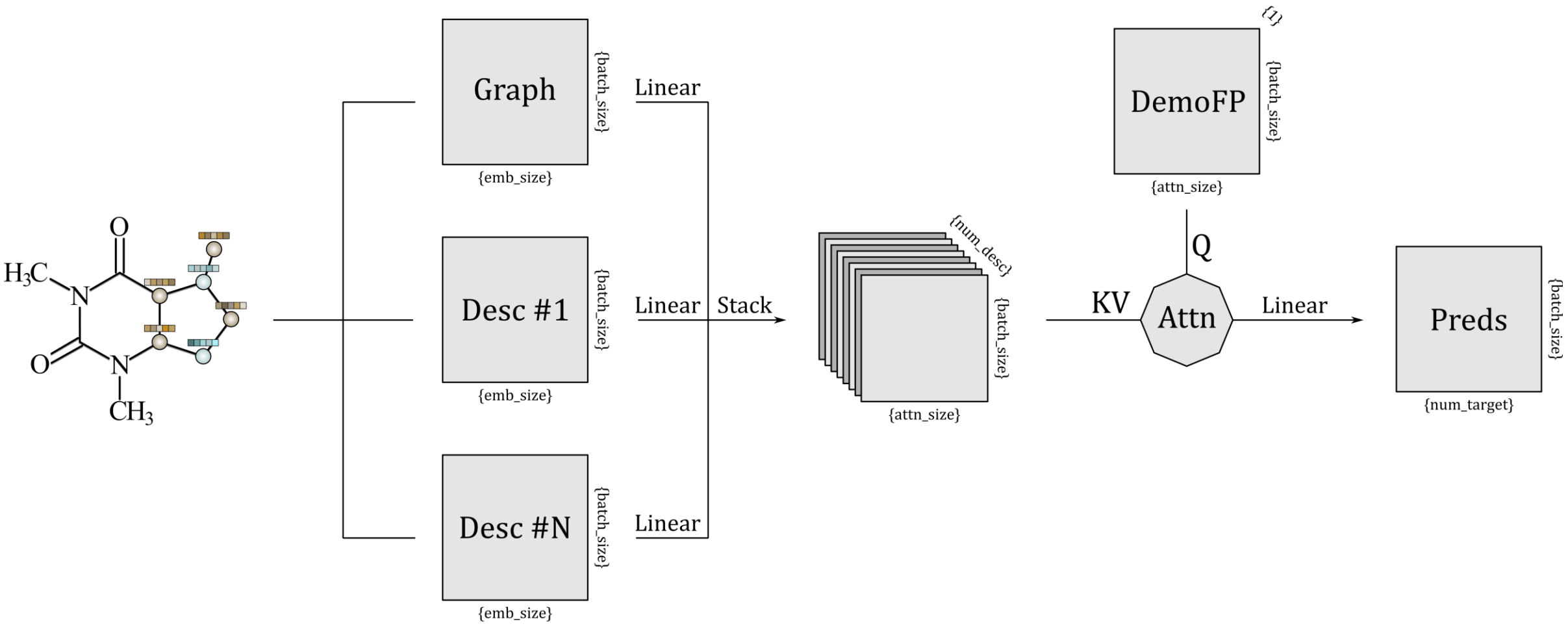




Classical Models

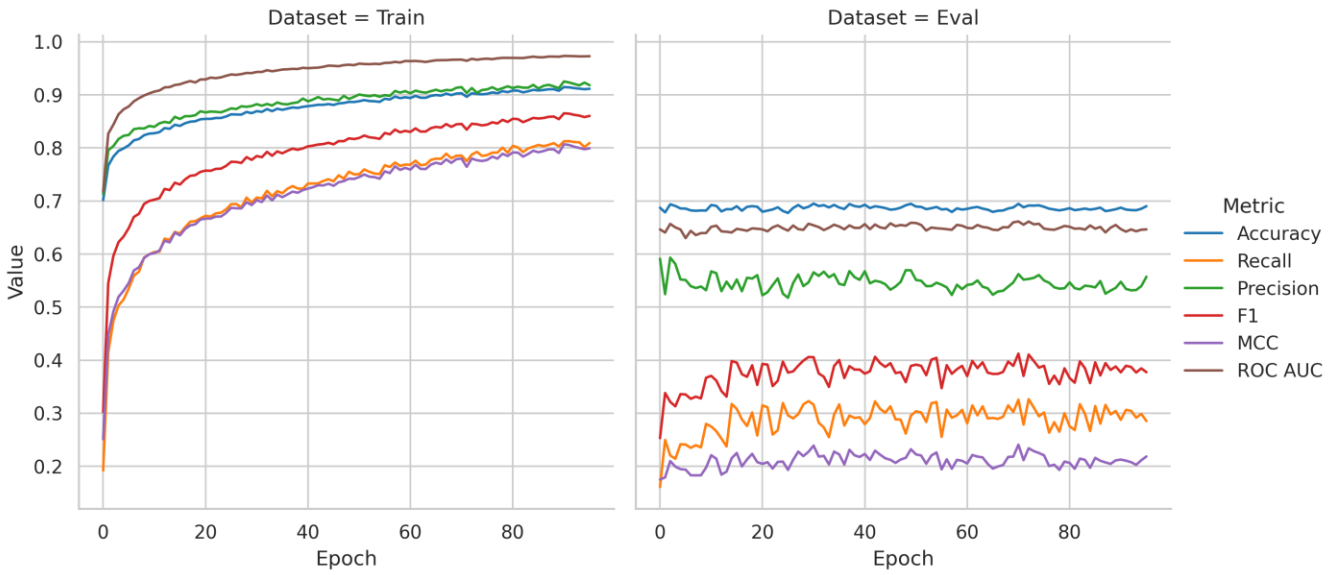


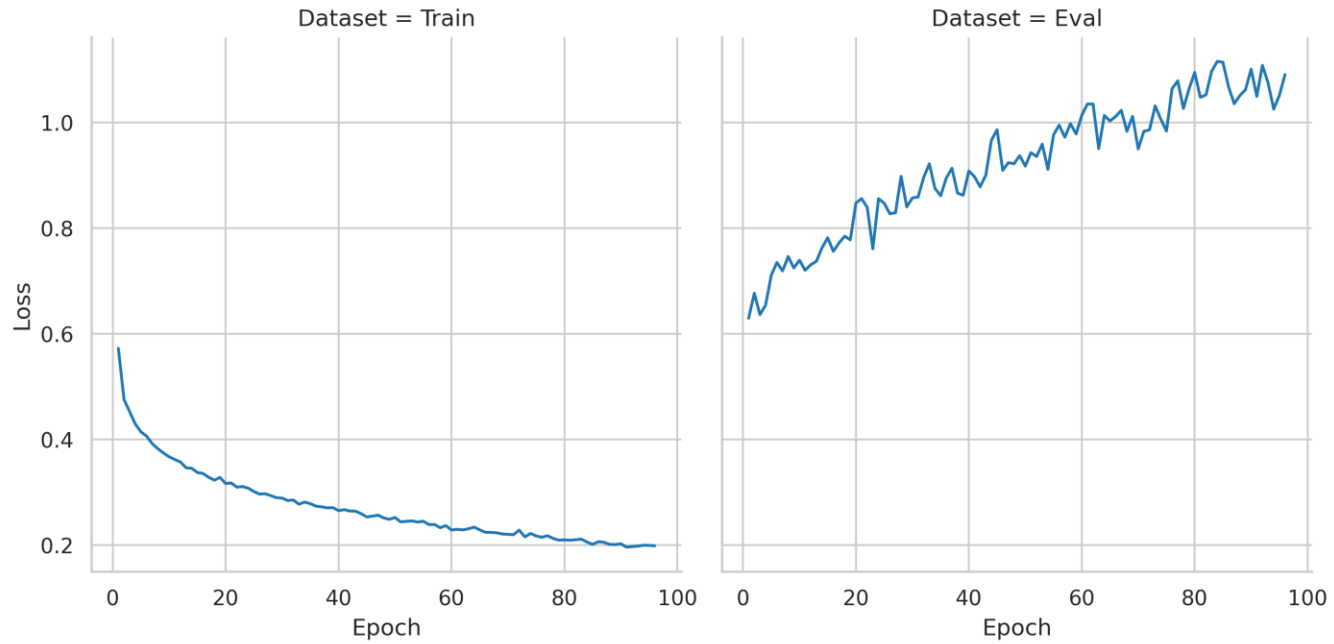
General DL architecture



Results and future

- Classical ML models are weakly predictive
- DL models overfit to the training test and don't generalize
- Next steps:
  - Add sample weighting
  - Add a second task – predicting confidence scores/weights directly
  - Move from Multi-Instance to Multi-task setting or build separate models





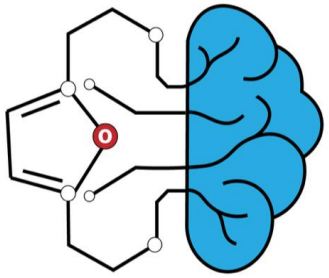
References

- [1] U.S. Food and Drug Administration. FAERS: FDA Adverse Event Reporting System.
- [2] Brown, E.G., Wood, L., Wood, S. The Medical Dictionary for Regulatory Activities (MedDRA). *Drug Safety*. 20(2), 109–117 (Feb 1999).
- [3] DrugBank 6.0. The DrugBank Knowledgebase for 2024. *Nucleic Acids Research*. 52, D1265–D1275 (Jan 2024).
- [4] Andrew, W. Front Matter. In: *Pharmaceutical Manufacturing Encyclopedia*, p. iii. Elsevier (2007).
- [5] Zdrazil, B., et al. The ChEMBL Database in 2023: A Drug Discovery Platform Spanning Multiple Bioactivity Data Types and Time Periods. *Nucleic Acids Research*. 52(D1), D1180–D1192 (Nov 2023).
- [6] Evans, S.J.W., Waller, P.C., Davis, S. Use of Proportional Reporting Ratios (PRRs) for Signal Generation from Spontaneous Adverse Drug Reaction Reports. *Pharmacoepidemiology and Drug Safety*. 10(6), 483–486 (2001).
- [7] Rothman, K.J., Lanes, S., Sacks, S.T. The Reporting Odds Ratio and Its Advantages Over the Proportional Reporting Ratio. *Pharmacoepidemiology and Drug Safety*. 13(8), 519–523 (2004).
- [8] Bate, A., Lindquist, M., Edwards, I.R., Olsson, S., Orre, R., Lansner, A., De Freitas, R.M. A Bayesian Neural Network Method for Adverse Drug Reaction Signal Generation. *European Journal of Clinical Pharmacology*. 54(4), 315–321 (Jul 1998).
- [9] Pedregosa, F., et al. Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research*. 12, 2825–2830 (2011).
- [10] Butina, D. Unsupervised Database Clustering Based on Daylight's Fingerprint and Tanimoto Similarity: A Fast and Automated Way to Cluster Small and Large Data Sets. *Journal of Chemical Information and Computer Sciences*. 39, 747–750 (1999).
- [11] Morgan, H.L. The Generation of a Unique Machine Description for Chemical Structures: A Technique Developed at Chemical Abstracts Service. *Journal of Chemical Documentation*. 5(2), 107–113 (May 1965).

Financing

- This study was funded by the Horizon Europe funding programme, under the Marie Skłodowska-Curie Actions Doctoral Networks grant agreement “Explainable AI for Molecules - AiChemist” no. 101120466.

Thank you for your attention!



TU/e

ISTITUTO DI RICERCHE FARMACOLOGICHE MARIO NEGRI · IRCCS

BAYER

Some more slides

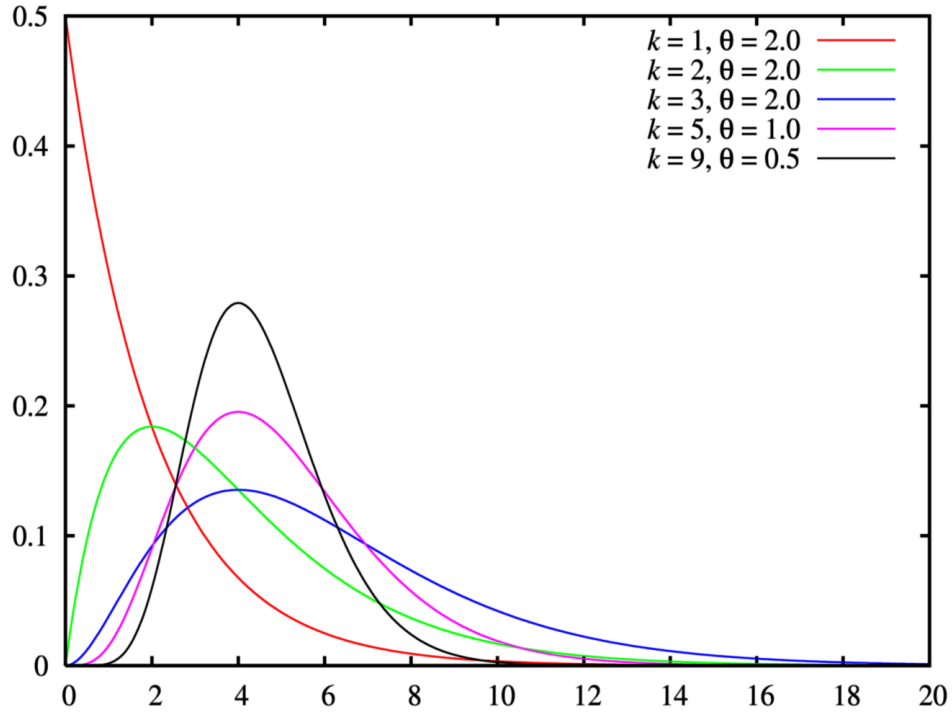
Information Component

- Information Component (IC):
  - Adapted from Bayesian Neural Network
  - Includes stabilizing factor
  - Equations taken from M.Fusaroli
  - CI based on gamma distribution

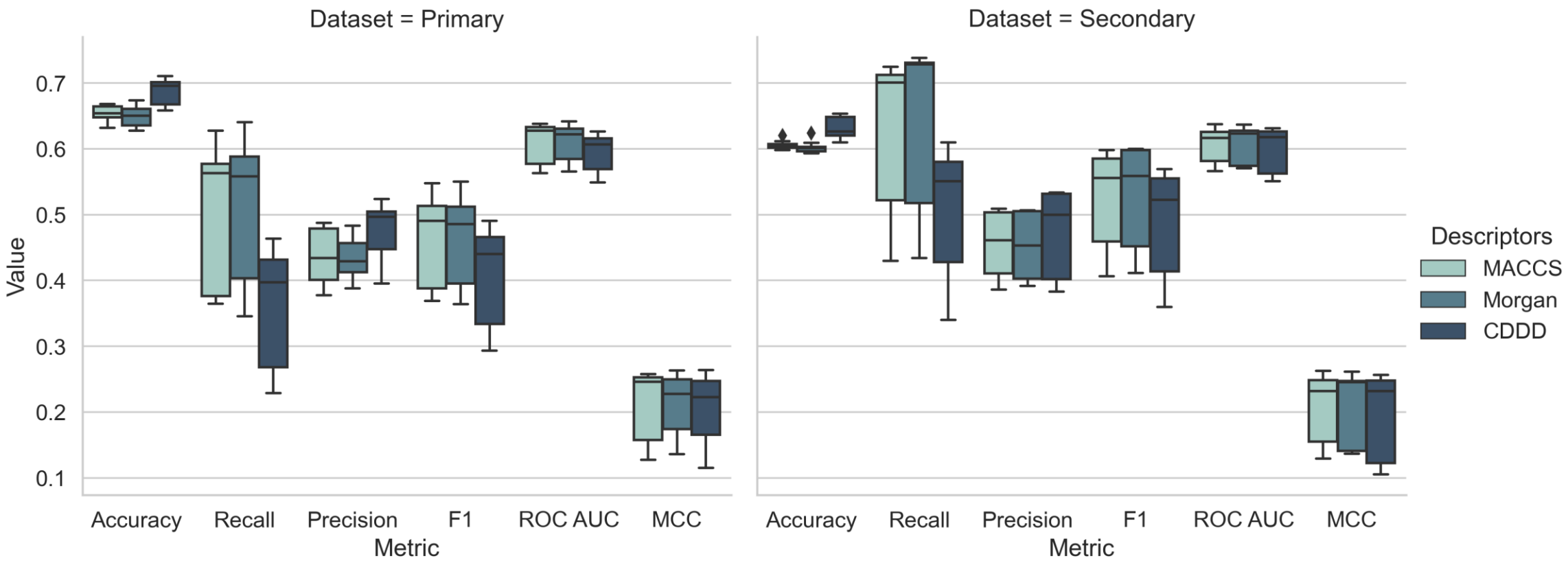
- Signal criteria:
  - At least 3 drug-ADR reports
  - $IC_{0.025} \geq 0$ for cardiotoxic label

$$IC = \log_2\left(\frac{a+\kappa}{N_{exp}+\kappa}\right) \qquad IC_{\alpha/2} = \log_2\left(Q_\Gamma(\alpha/2)\right)$$

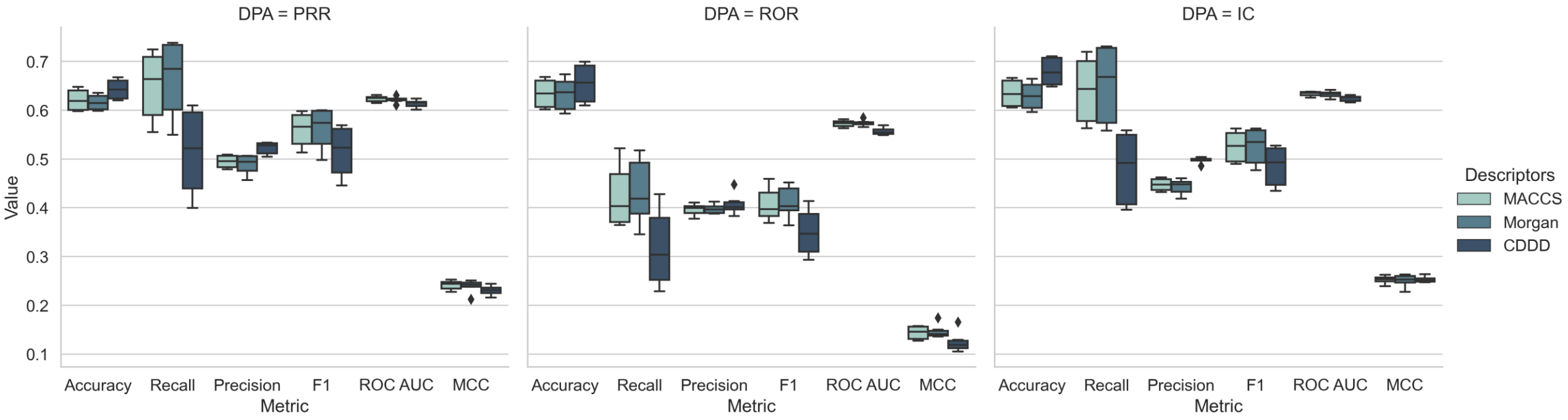
$$\Gamma = \Gamma\left(k = a+\kappa,\ \theta = \frac{1}{N_{exp}+\kappa}\right)$$



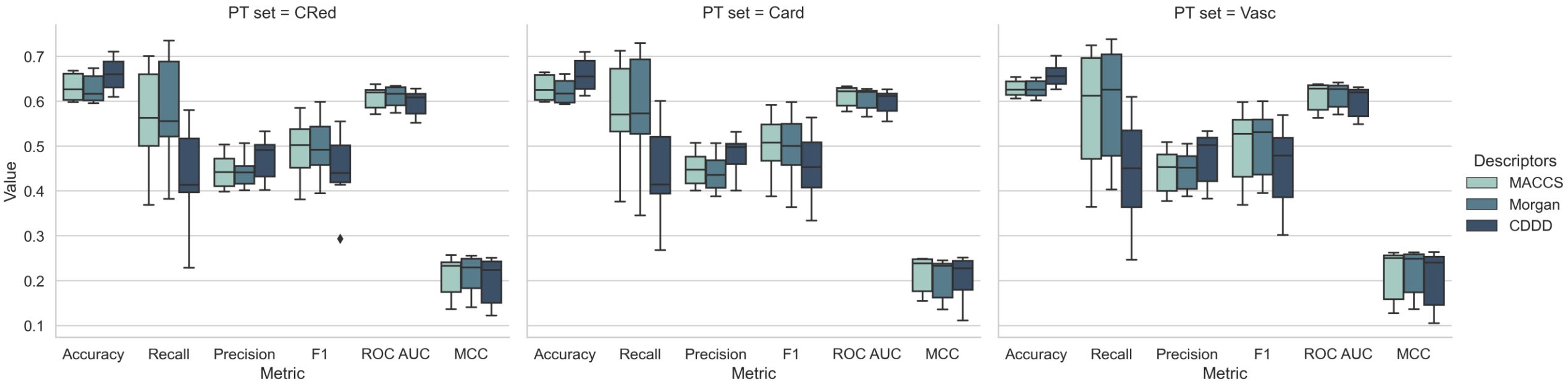
Dataset selection - Primary vs Secondary



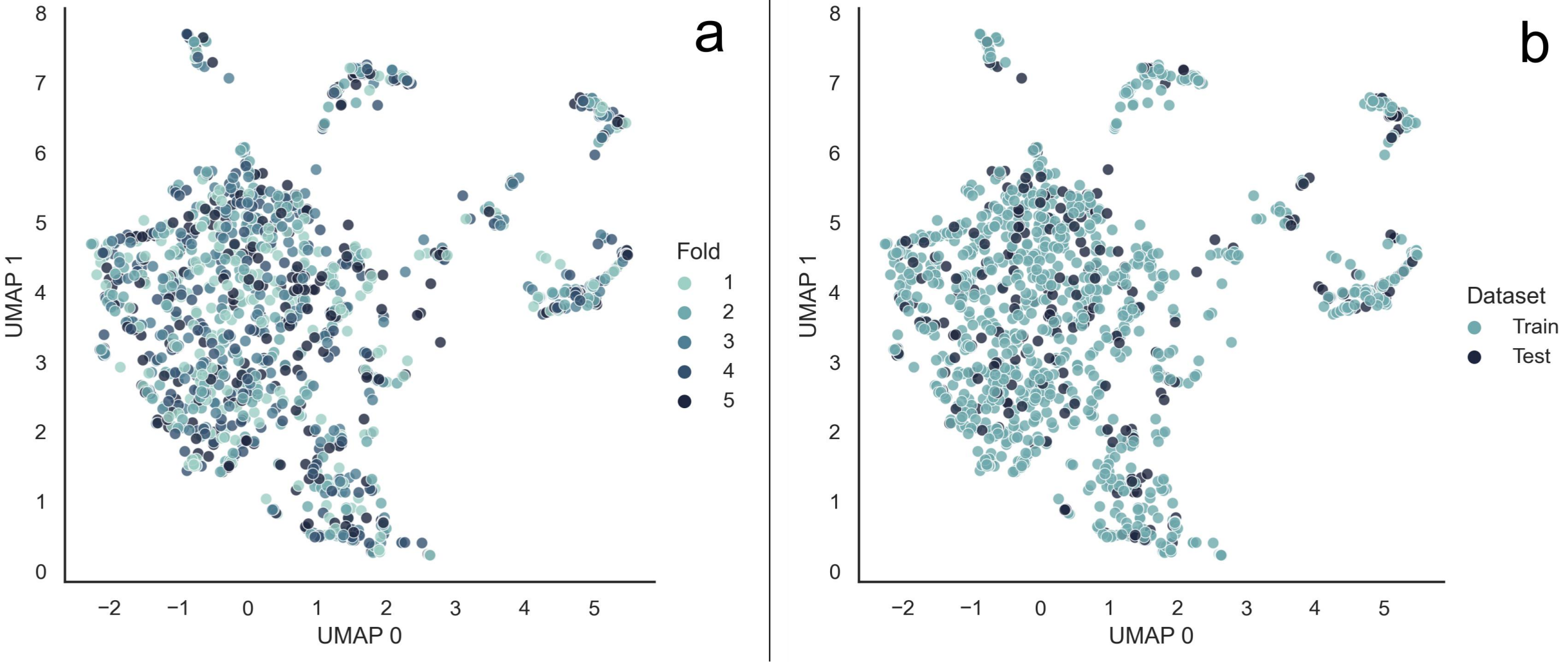
Dataset selection - DPA metrics



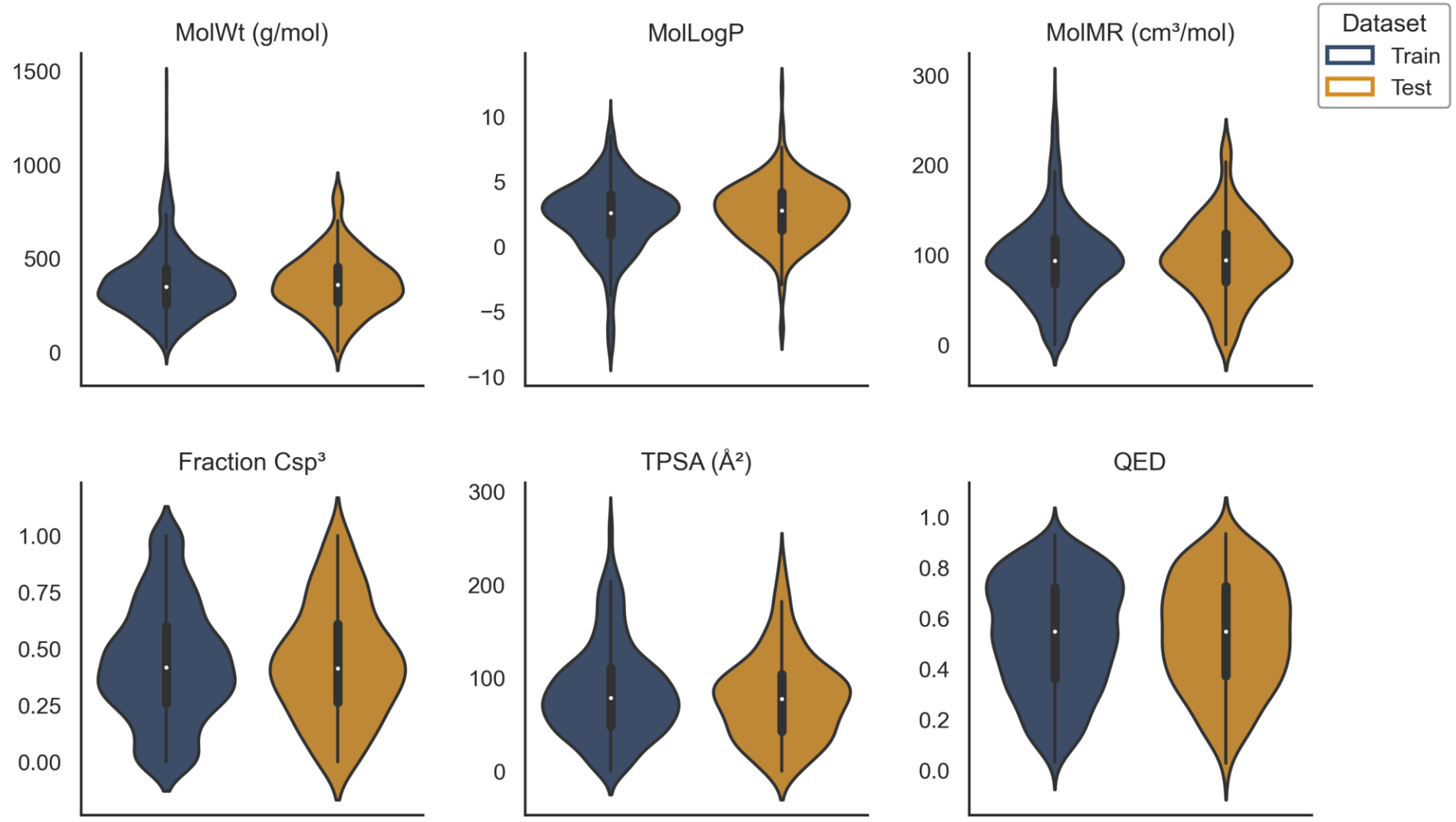
Dataset selection - PT sets



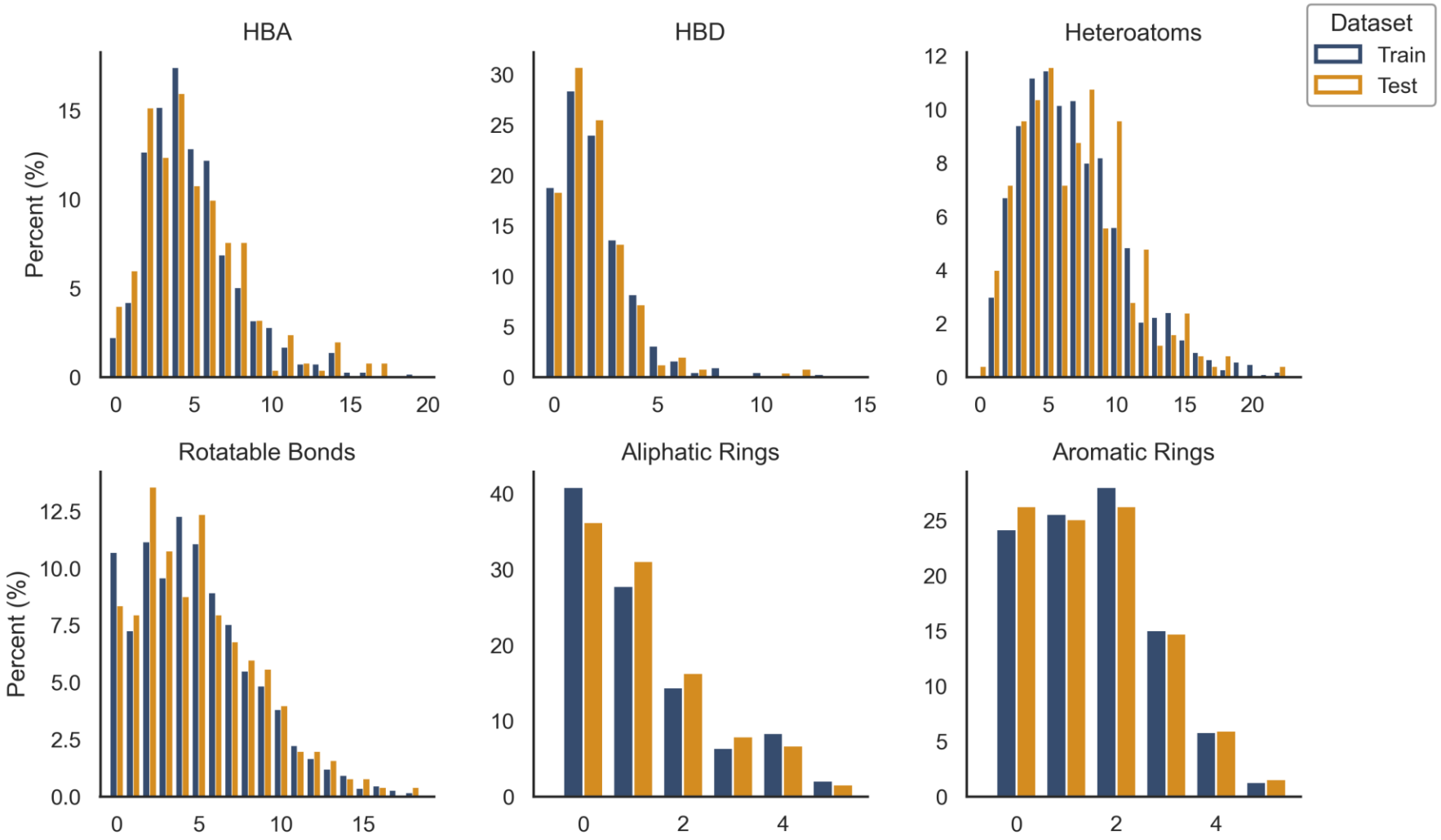
Dataset split



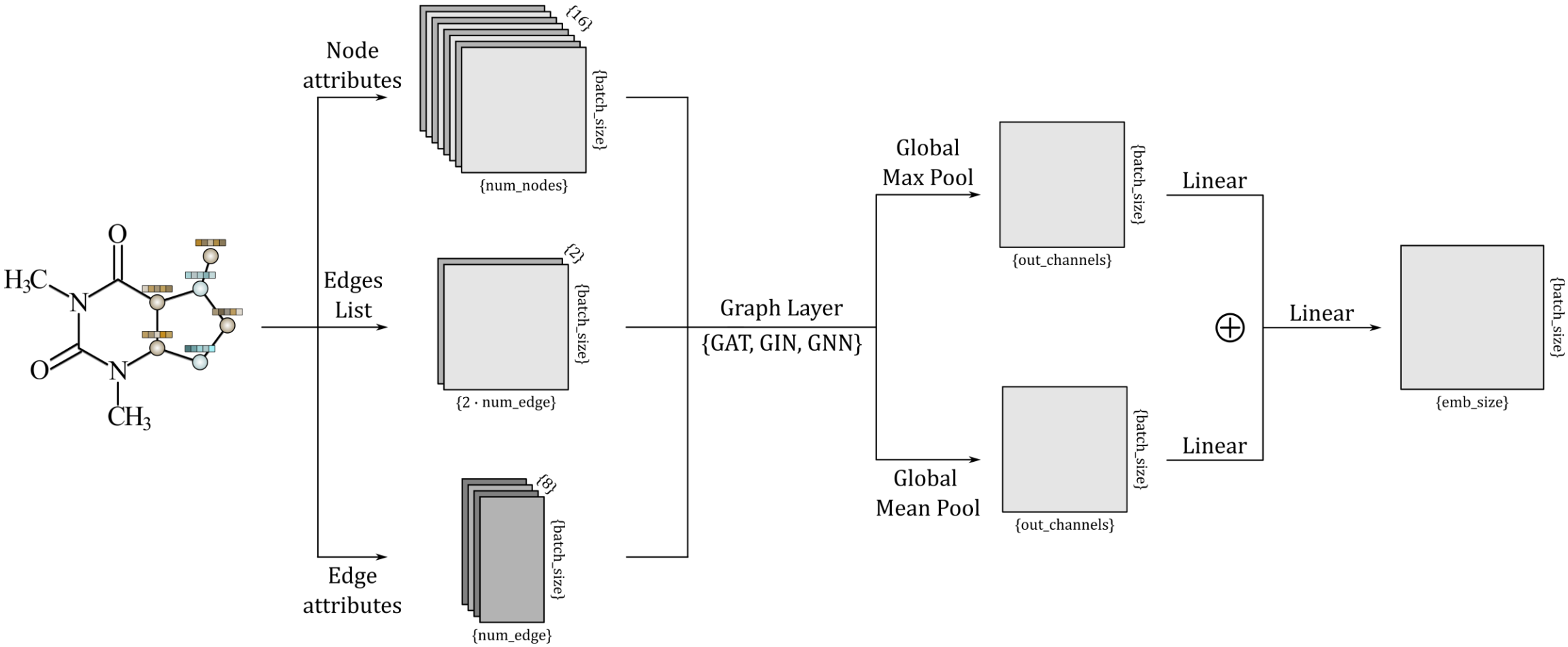
Train and Test properties



Train and Test properties



GNN


Node attributes
{16}
{batch_size}
{num_nodes}
Edges List
{2}
{batch_size}
{2 · num_edge}
Edge attributes
{8}
{batch_size}
{num_edge}
Graph Layer
{GAT, GIN, GNN}
Global Max Pool
{batch_size}
{out_channels}
Global Mean Pool
{batch_size}
{out_channels}
Linear
Linear
Linear
{batch_size}
{emb_size}