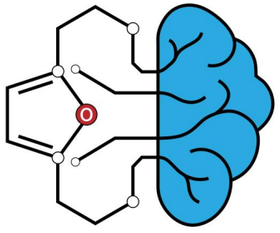
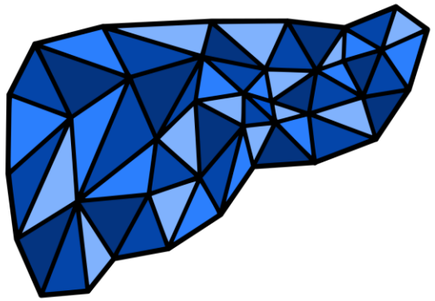
Pharmacovigilance Meets Demographics: Towards Personalized Cardiotoxicity Prediction

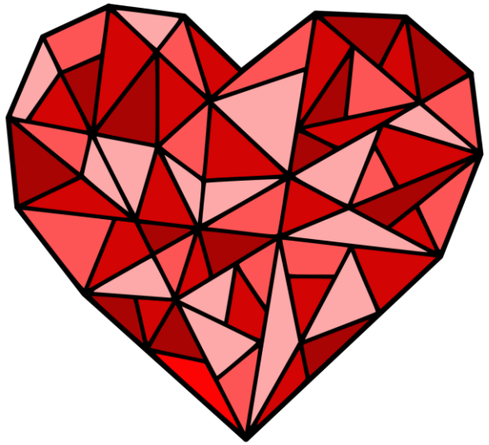
Mateusz Iwan



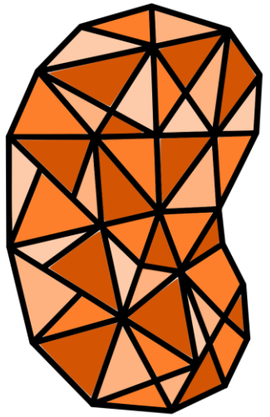
Advanced ML methods to predict and understand the toxicity of drugs



Hepatotoxicity

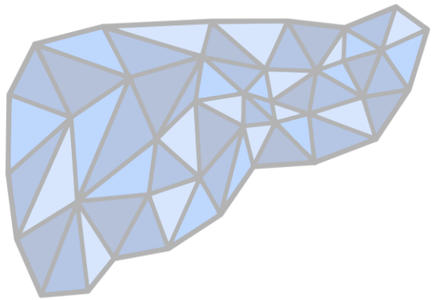


Cardiotoxicity

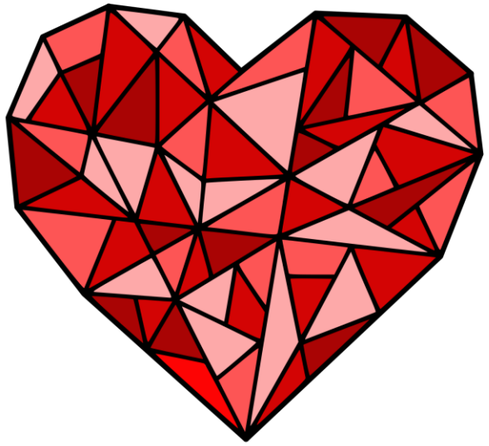


Nephrotoxicity

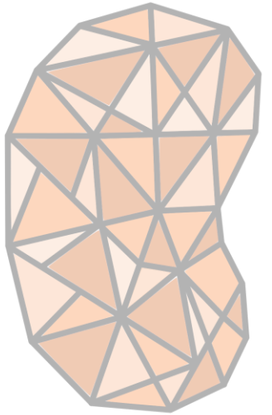
Advanced ML methods to predict and understand the toxicity of drugs



Hepatotoxicity



Cardiotoxicity



Nephrotoxicity

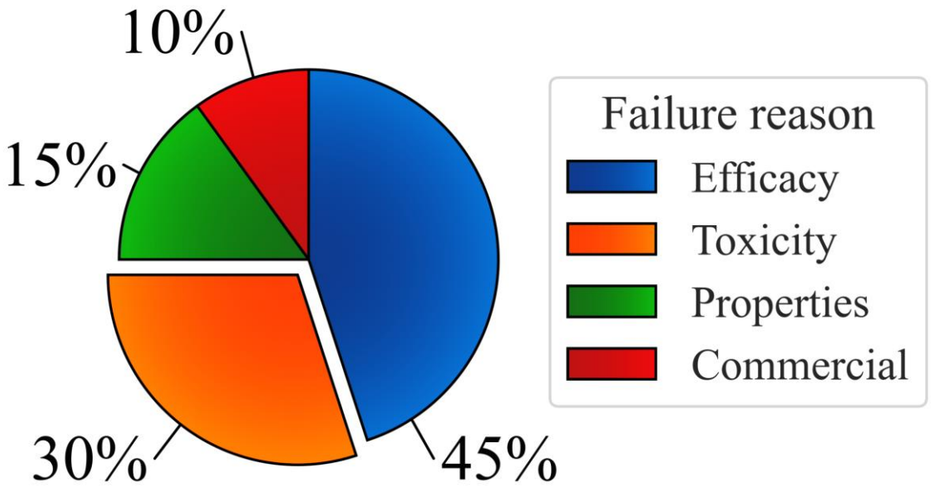
Rationale and related work[1-52]

Rationale and related work[1-52]

- Demographic factors influence drug-induced toxicity

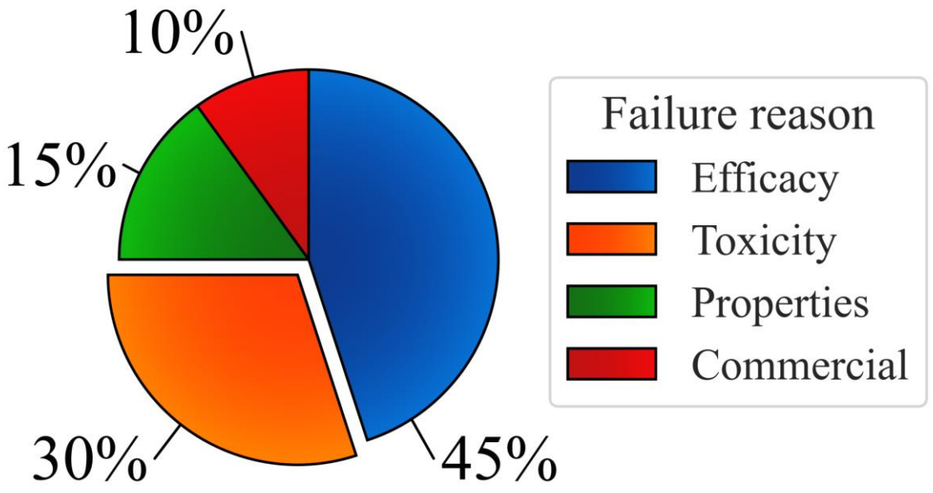
Rationale and related work[1-52]

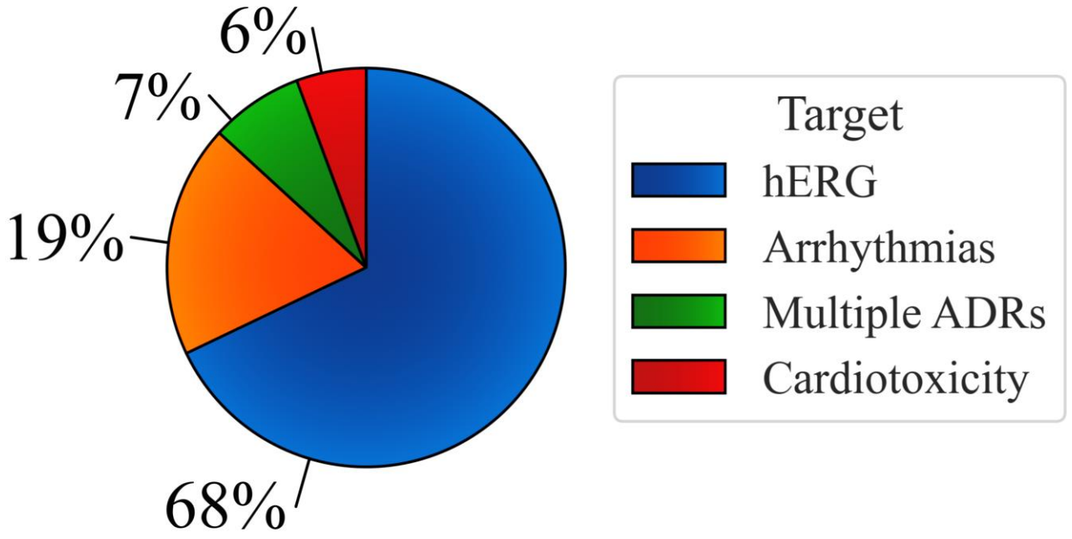
- Demographic factors influence drug-induced toxicity
- Multiple clinical trials fail due to problems with toxicity



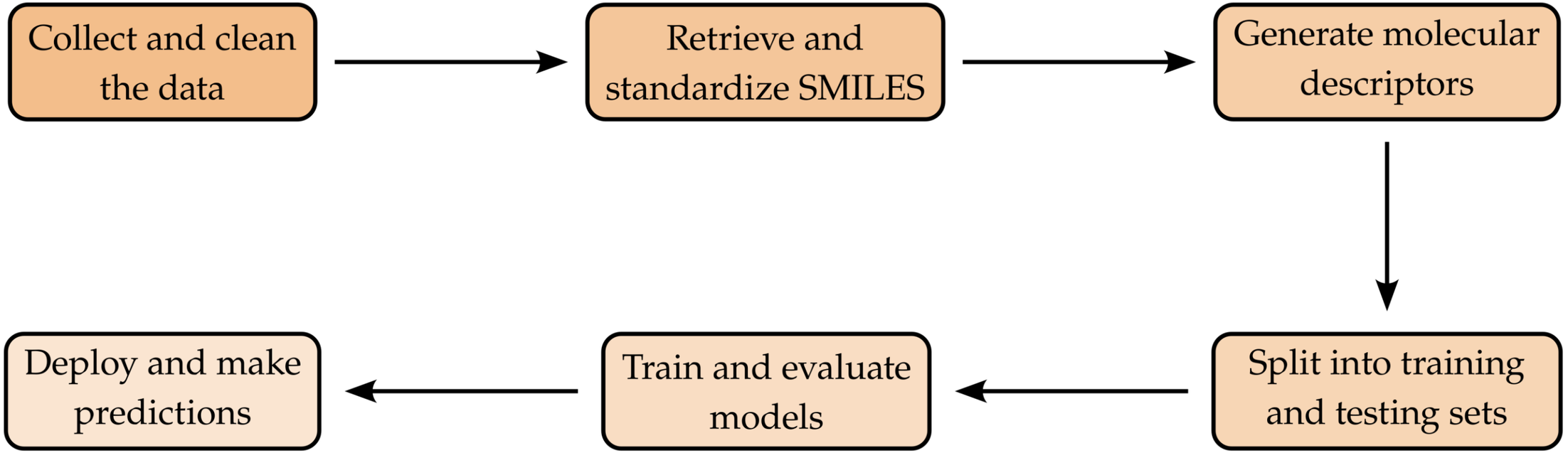
Rationale and related work[1-52]

- Demographic factors influence drug-induced toxicity
- Multiple clinical trials fail due to problems with toxicity
- Existing models focus mostly on the hERG channel and arrhythmias



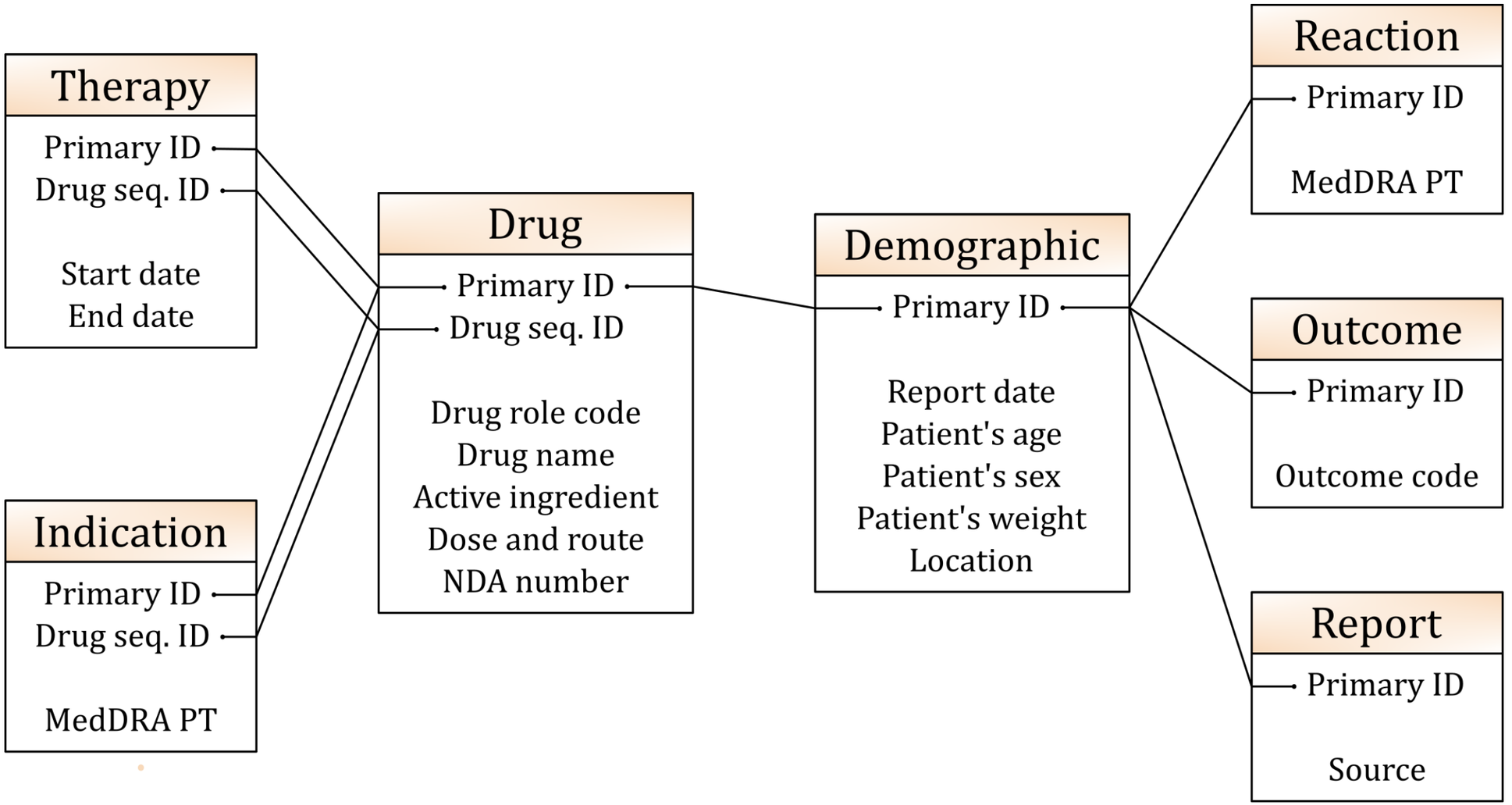


Typical ML workflow



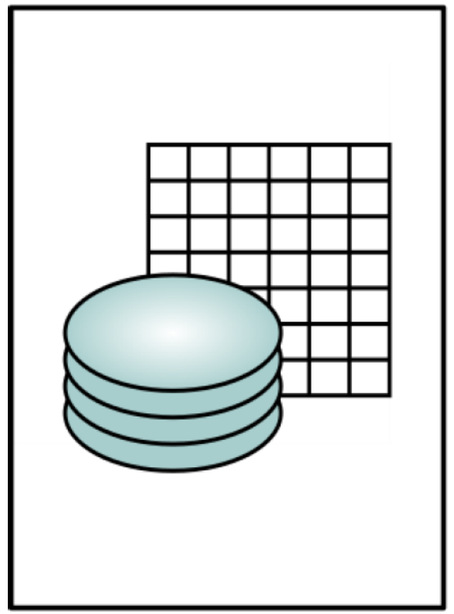
Dataset preparation

FAERS database[53]



A few numbers

- 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%



Initial dataset

Primary ID	Event date	Sex	Age [months]	Weight [kg]	ADR	Role code - Drug	Indication
226913651	2023-07-10	Male	516	98	Hospitalisation, Therapy Interrupted	PS - Humira	Rheumatoid arthritis, Ankylosing spondylitis
200966371	2021-11-22	Female	876	UNK	Off label use, Postperative wound infection	PS – Dexamethasone SS – Lenalidomide SS – Velcade	Plasma cell myeloma
114628731	2015-08-31	Female	UNK	UNK	Seizure	PS – Cymbalta C – Xanax C – Ambien C – Milnacipran	Depression
96157251	2013-10-11	Female	UNK	UNK	Headache	PS – Butrans	Product used for unknown indication
210530602	2022-07-07	UNK	UNK	UNK	Eye discharge, Ocular hyperaemia	PS - Xiidra	Product used for unknown indication

No.	Description	Issues
01	Sulfamethoxazole/trimethoprim ds	Varying separating character ('/')
02	Bi tildiem l.p. 90mg, comprim? Enrob? Lib?ration prolong?e	Missing hyphen between, irrelevant information, non-English entry and characters
03	Diphenidol	None
04	Depakine chrono 500mg,	Irrelevant information
05	Diazepam ^aps^	Irrelevant information
06	Amlodipine/irbesartan	Varying separating character ('/')
07	Urosdiol	Typo (ursodiol)
08	Exemestane (exemestane) (unknown)	Information in parentheses
09	Sodium chloride injection usp 0264?7800?09	Irrelevant information, non-standard characters
10	Quinapril pch	'pch' might be an abbreviation somewhere else
11	Lsartan	Typo (losartan)
12	Oxybutynim cl er	Abbreviated salt, abbreviated formulation
13	Dietary suplement\\ubidecarenone	General term, varying separating character('\\')
14	Vincristina teva italia	Non-English entry, irrelevant information
15	Cardio aspirin	'Cardio' might be a trade name somewhere elese
16	Fumarato de bisoprolol	Non-English entry
17	Adco nevirapine	Irrelevant information
18	Xeroquel lp 300mg, comprim? ? Lib?ration prolong?e	Irrelevant information, non-standard characters, non-English entry
19	Rizatriptan doc	Irrelevant information
20	Eptifibatid	Typo (eptifibatide)

Token processing[55-57]

- Storage of DrugBank, PME, and ChEMBL synonym dictionaries, and already seen drugs
- Automated queries to selected online resources while caching results
- Automated similarity searches and processing of already-seen entries
- Several options for string manipulation across the whole dataset: substitution, removal, extraction, and extension, using either string-based or regex-based solutions

Algorithm 1 Token processing

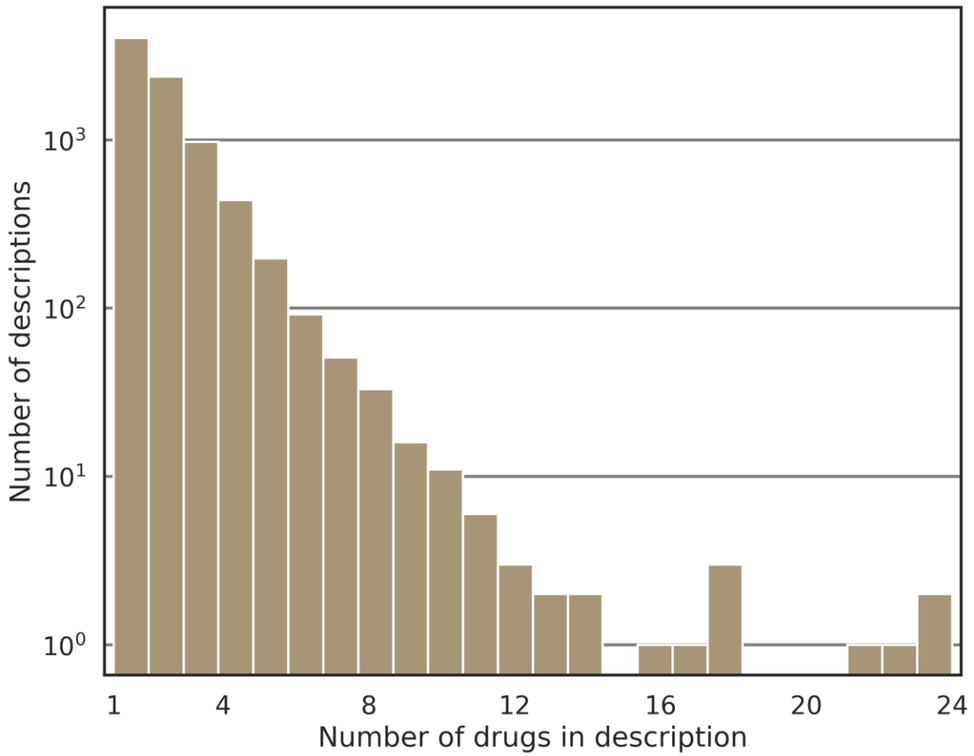
```
procedure PROCESS TOKENS(tokens)
    for token in tokens do
        if token in DrugBank then
            capture
        else
            calculate similarity to drugs and synonyms in DrugBank
            calculate similarity to previous tokens
            query external databases          ▷ e.g.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

Token processing

- The following groups were removed: vaccines, immunoglobulins, RNA-based drugs, peptides, proteins, polymers, probiotics, herbal and homeopathic formulations, infusion or dialysis fluids, multivitamins, foods, nutritional preparations, unclear abbreviations, and 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

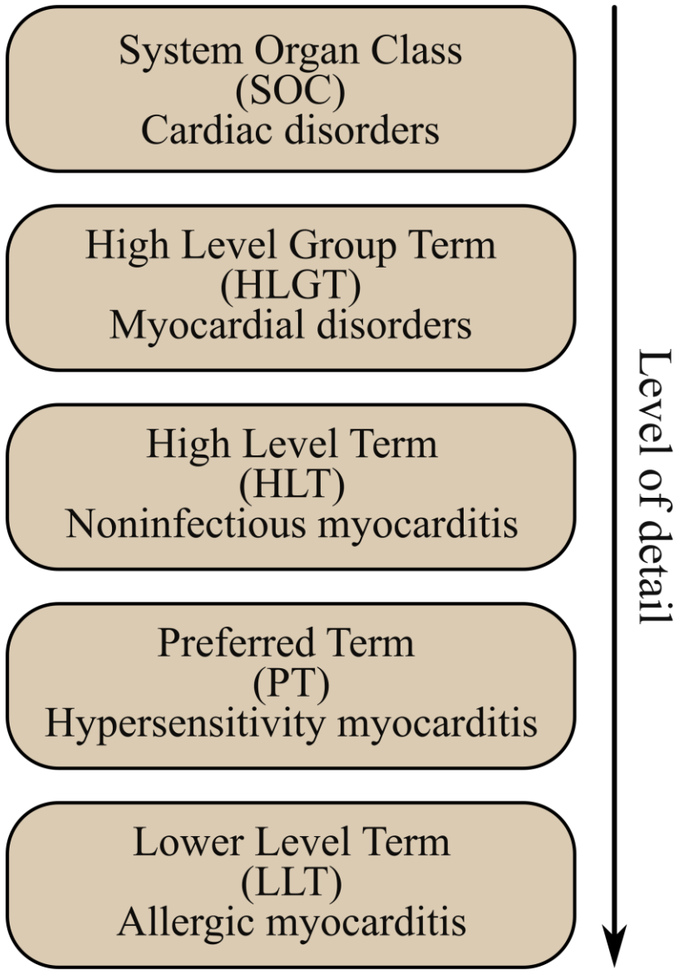


MedDRA – terms selection[54]

- Standardized medical terminology developed by the International Conference for Harmonisation (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

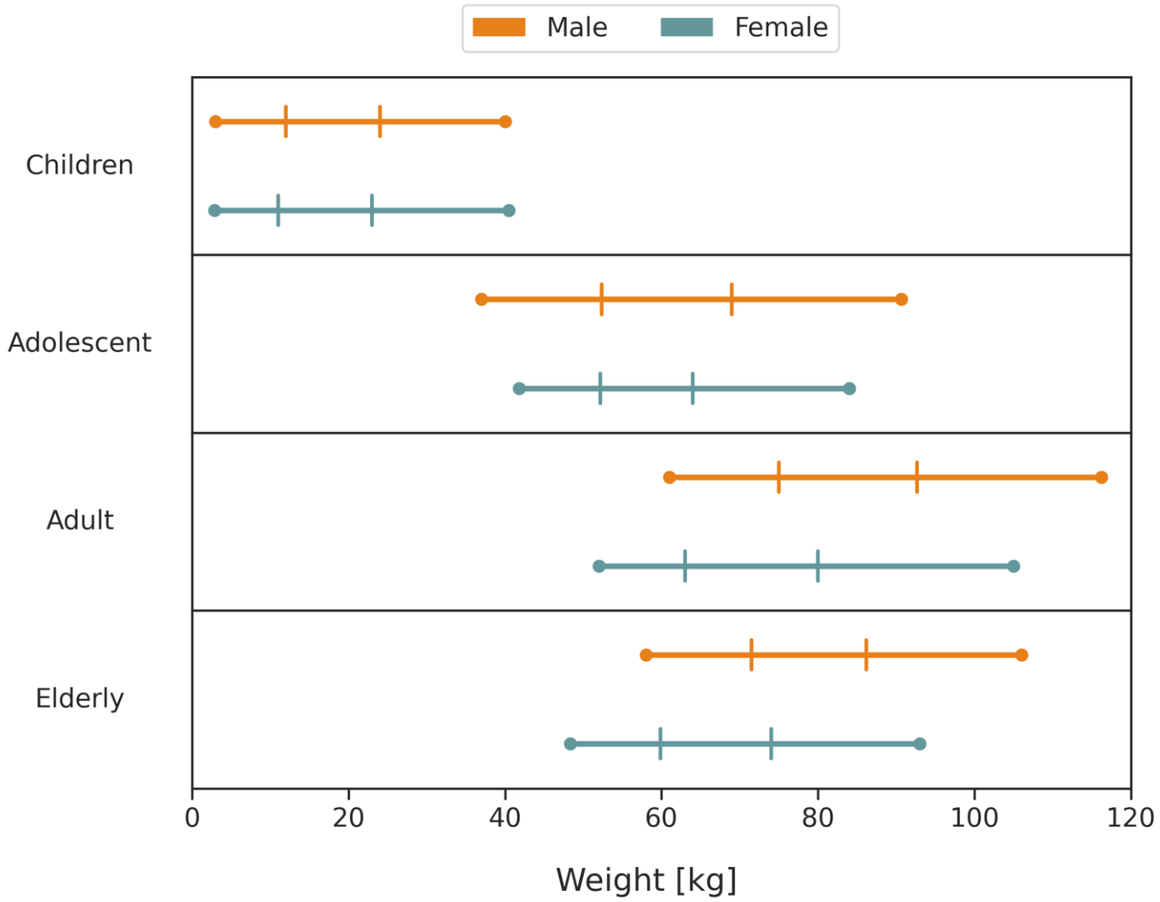


Demographic data processing

- Sex was used without further processing

- Age was binned following the FDA classification:
  - Children (birth - 12 years)
  - Adolescent (12 - 21 years)
  - Adult (21 - 65 years)
  - Elderly (65 - 100 years)

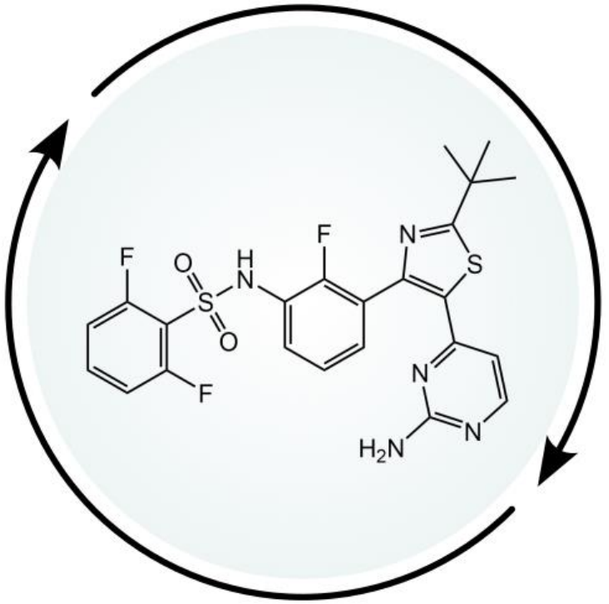
- Weight was binned based on quantiles:
  - Low ($Q_{0.05}$ - $Q_{0.33}$)
  - Average ($Q_{0.05}$ - $Q_{0.67}$)
  - High ($Q_{0.67}$ - $Q_{0.95}$)



SMILES processing[58-65]

❖ Drug name to SMILES mapping:
- Chemical Identifier Resolver
- PubChem API
- Manual check

❖ Standardization:
- Stripping of salts
- Charge neutralization
- Removal of stereochemistry



❖ Calculated descriptors:
- Mordred molecular descriptors
- RDKIT molecular descriptors
- Klekota & Roth fingerprints
- ChemBERTa embeddings
- Morgan fingerprints
- MACCS fingerprints
- CDDD embeddings

❖ Final statistics:
- 7 types of descriptors
- 3,759 drugs with valid SMILES
- 3,618 drugs after standardization

Disproportionality Analysis

Disproportionality Analysis[66-68]

- Disproportionality Analysis (DPA):
  - Based on a statistical analysis of the number of reported drug-reaction cases vs the expected number
  - Used for the early detection of potential adverse drug reactions
  - A signal does not equal a causal relationship

	Event	¬ Event
Drug	a	b
¬ Drug	c	d

- Frequently used metrics:
  - Proportional Reporting Rate (PRR)
  - Reporting Odds Ratio (ROR)
  - Information Component (IC)

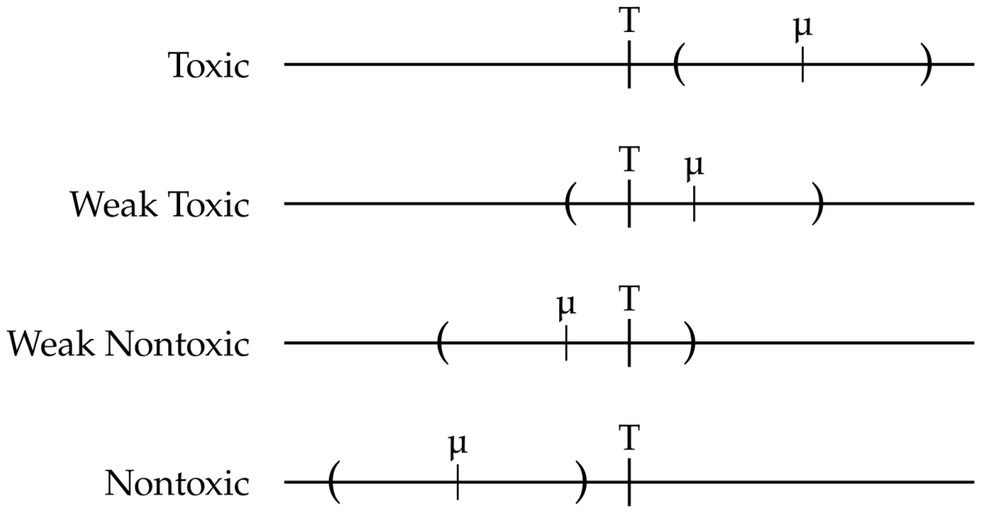
$$PRR = \frac{a\,/\,(a+b)}{c\,/\,(c+d)} \qquad ROR = \frac{a\,/\,b}{c\,/\,d}$$

- Three PT sets to describe Cardiotoxicity:
  - Cardiovascular (Vasc)
  - Cardiac (Card)
  - Cardiac Reduced (CRed)

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

Disproportionality Analysis

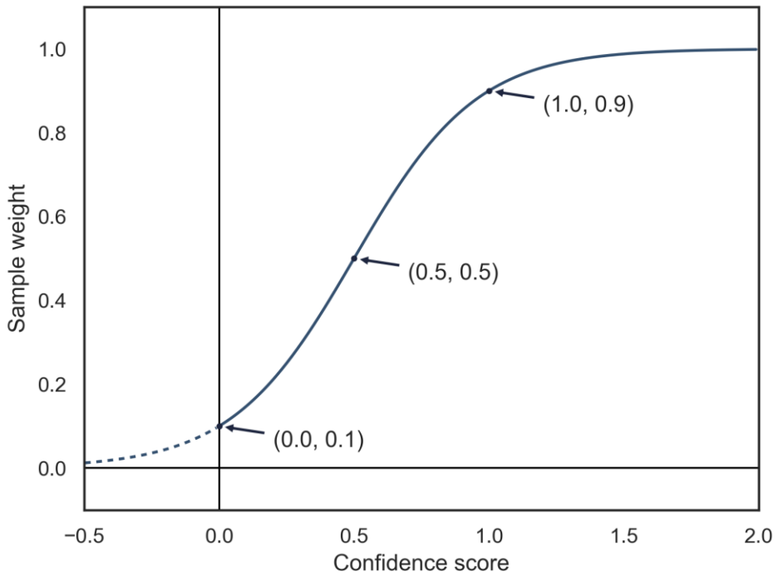
- Four-class assignment based on the Confidence Interval (CI)
- Used thresholds:
  - PRR and ROR: 1.0
  - IC: 0.0



- Additional label confidence score:

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

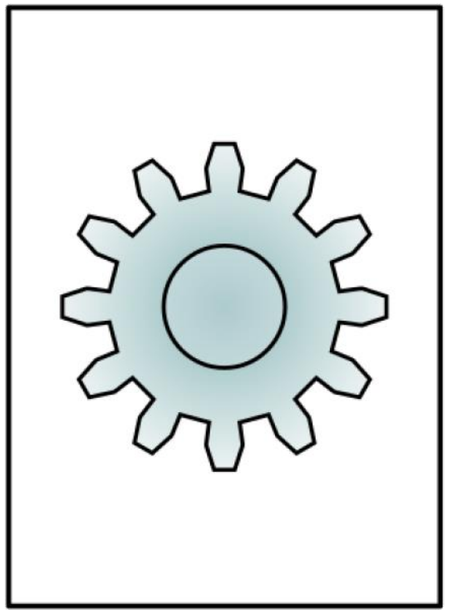
- Further transformed using a modified sigmoid function:



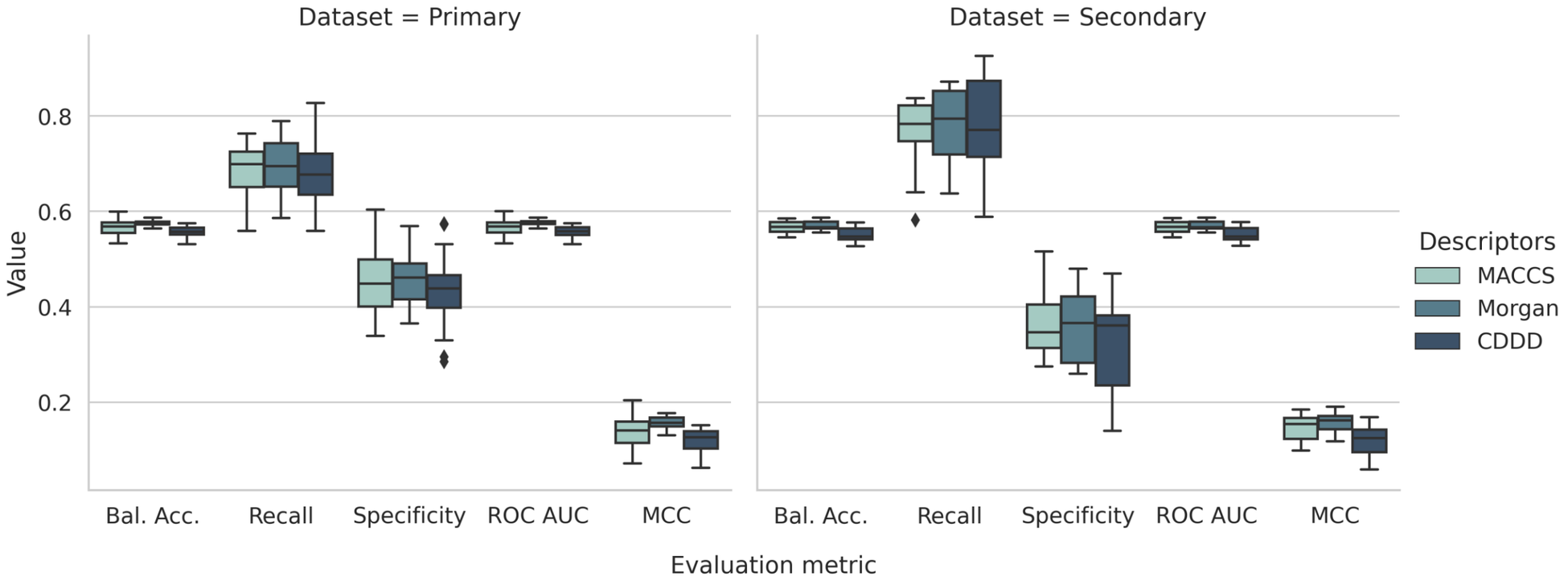
Datasets evaluation

Dataset evaluation - setup

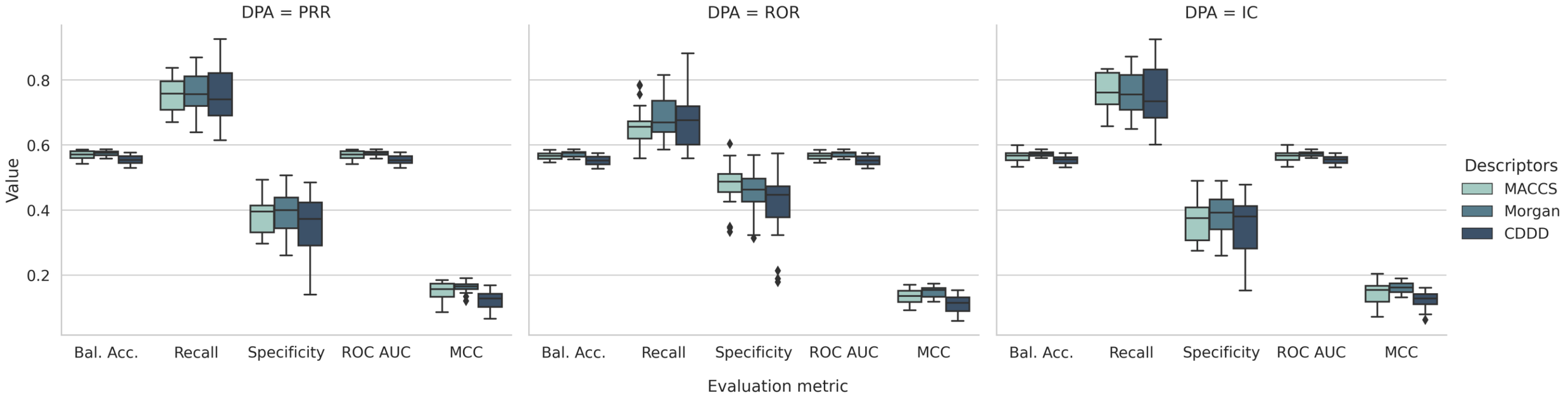
- 5-fold cross-validation using stratified random split
- Stratification on molecule types, target, and demographic factors
- Four classes of models
- Three descriptor types
- Evaluated conditions:
  - Inclusion of 'Secondary suspects'
  - Three PT sets describing cardiotoxicity
  - Three DPA metrics describing signal strength
- Multi-instance dataset



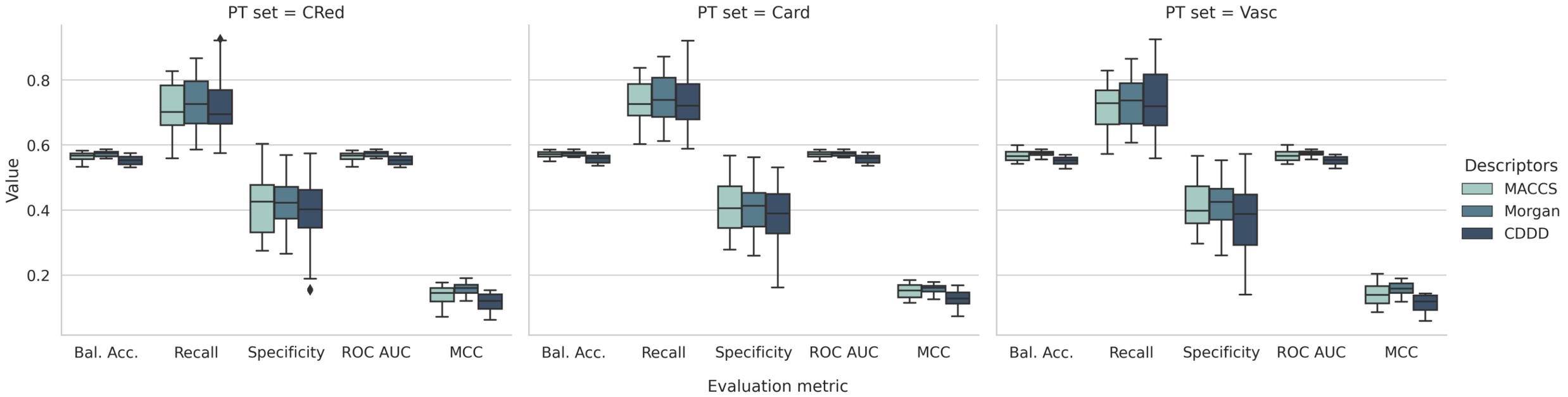
Inclusion of secondary suspects



Differences in DPA metrics



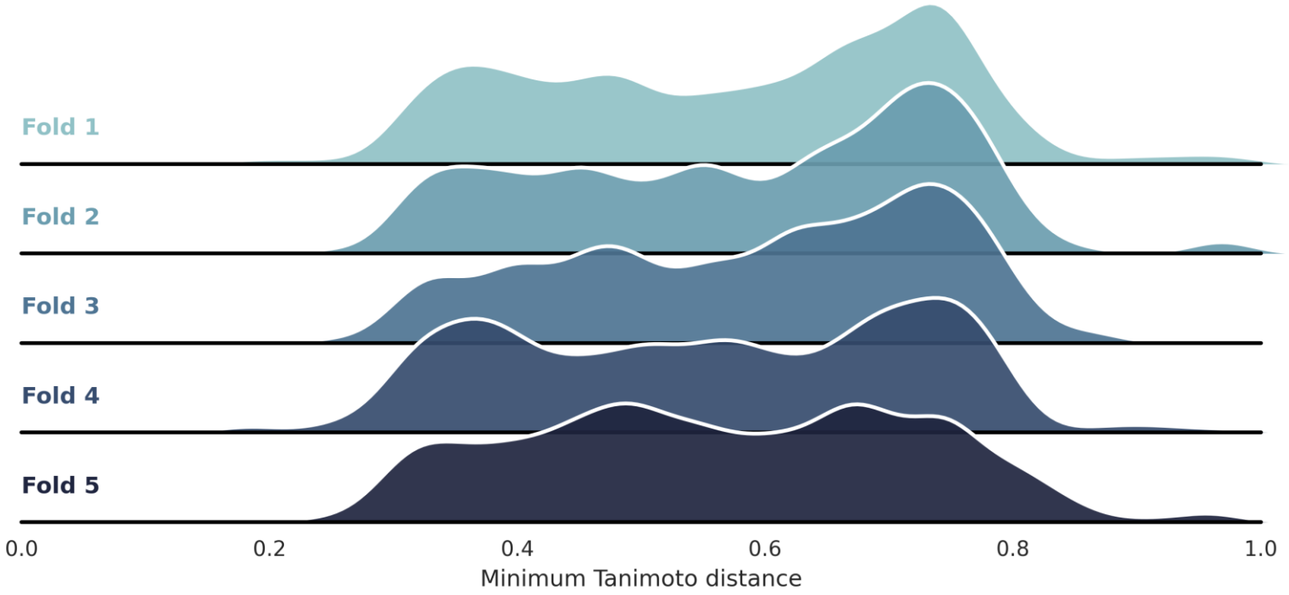
Differences in Preferred Term sets

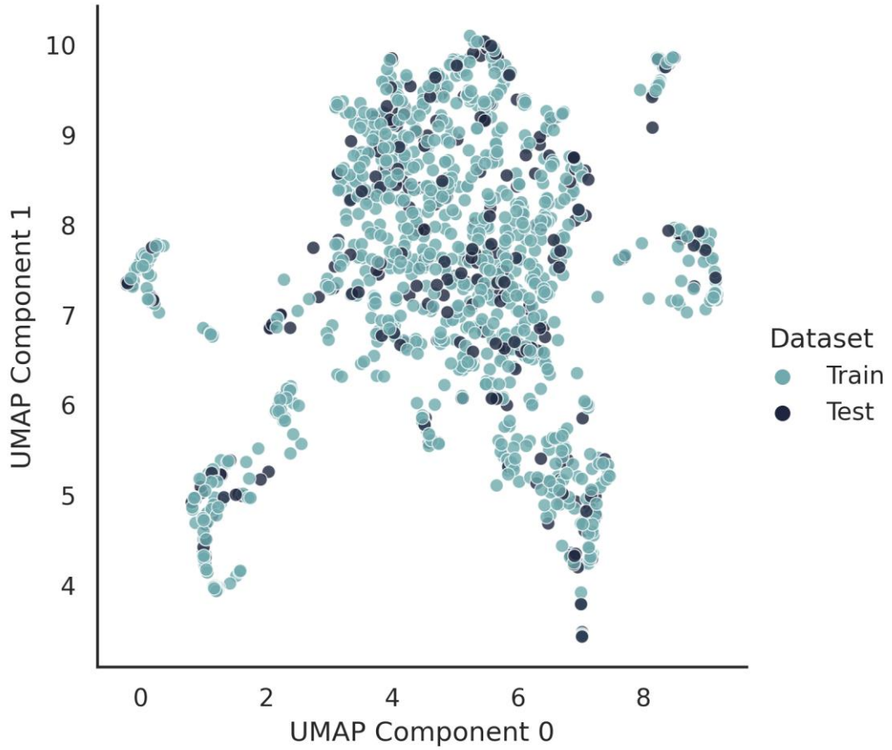


Initial models

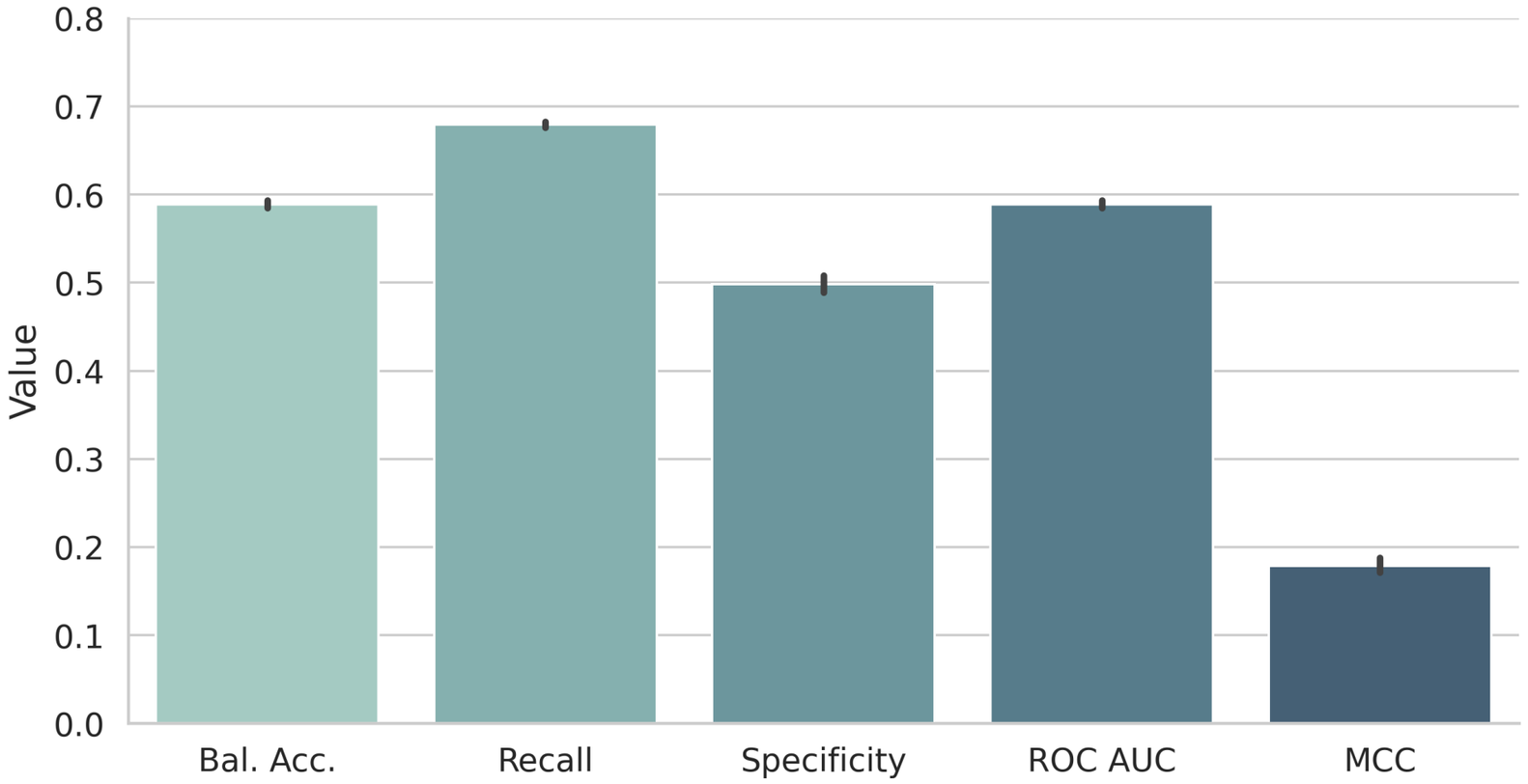
Dataset splitting

- Butina clustering using Tanimoto distance
- Morgan fingerprints (radius=2, nBits=2048)
- Test set - fold with the highest Mean Minimum Tanimoto distance

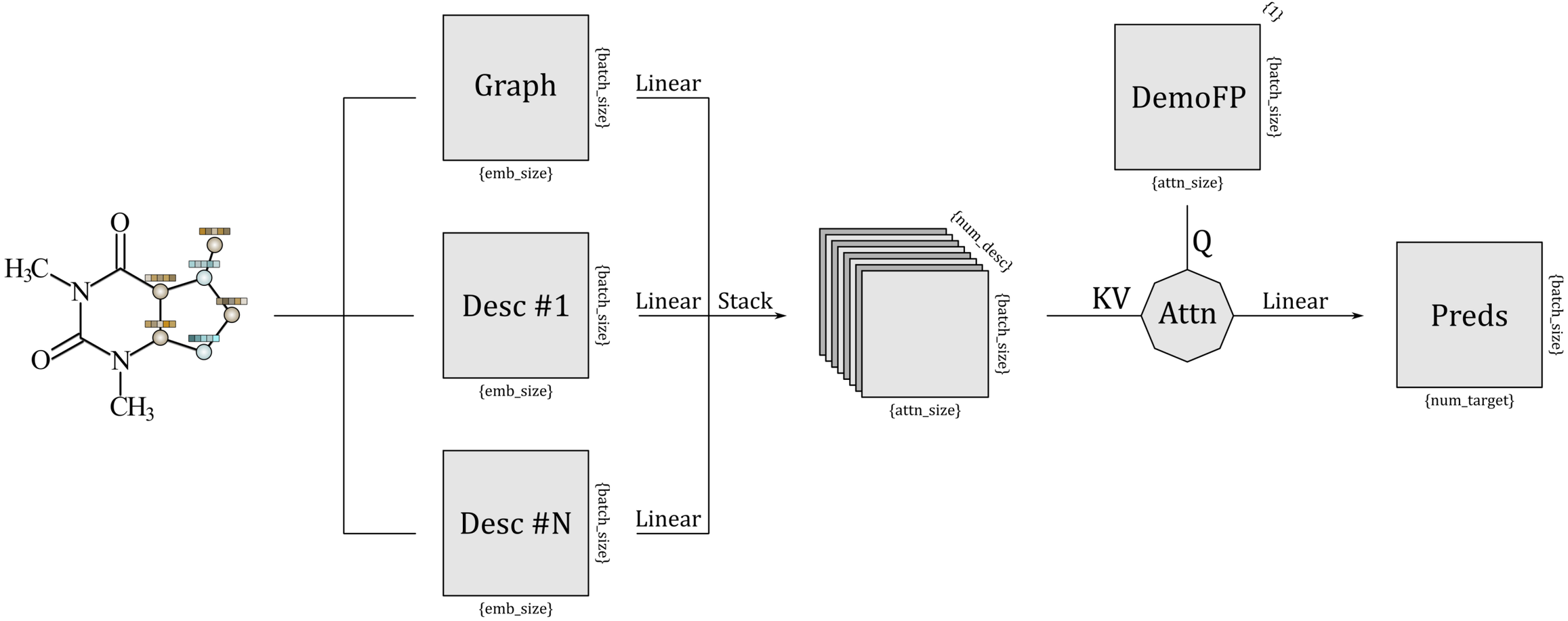




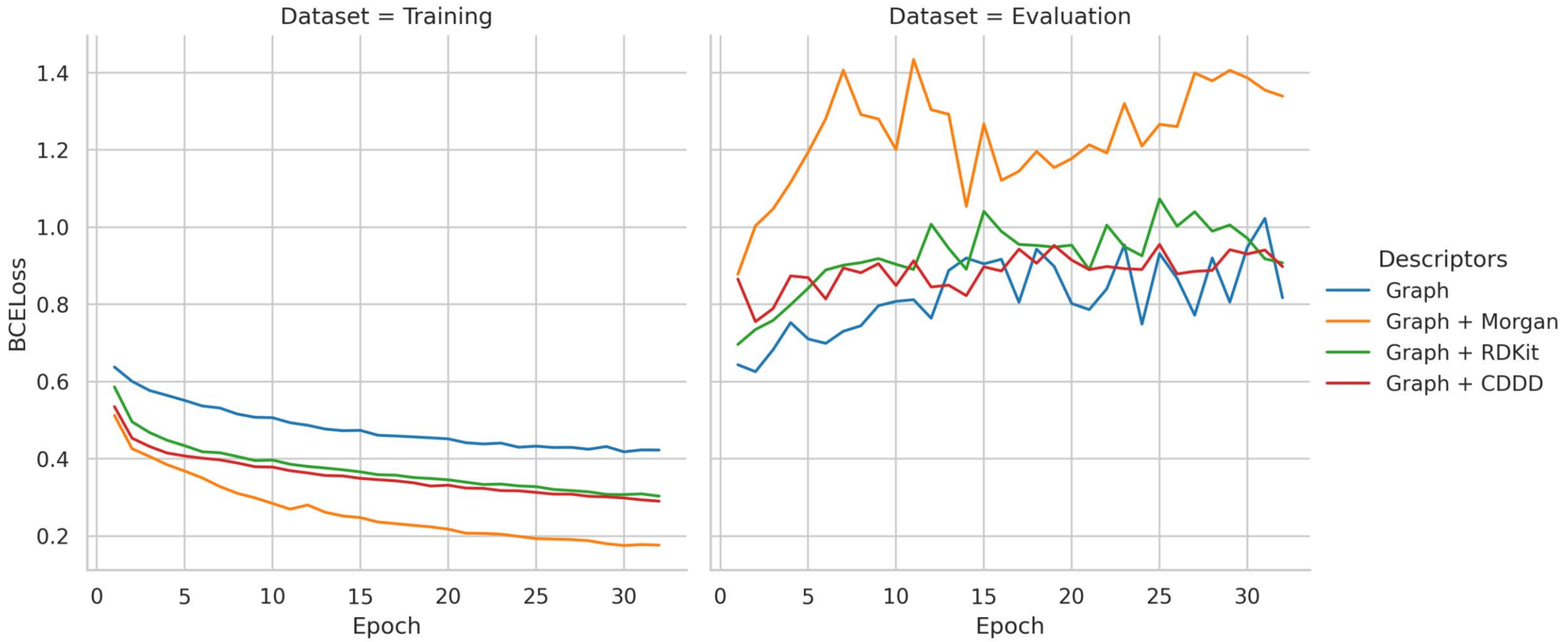
Baseline models - results



Deep learning architecture

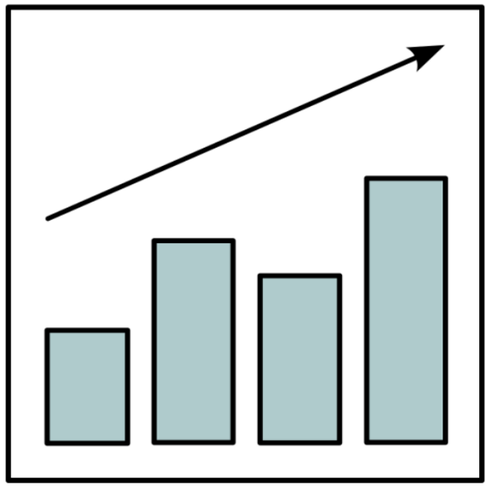


DL model – initial results



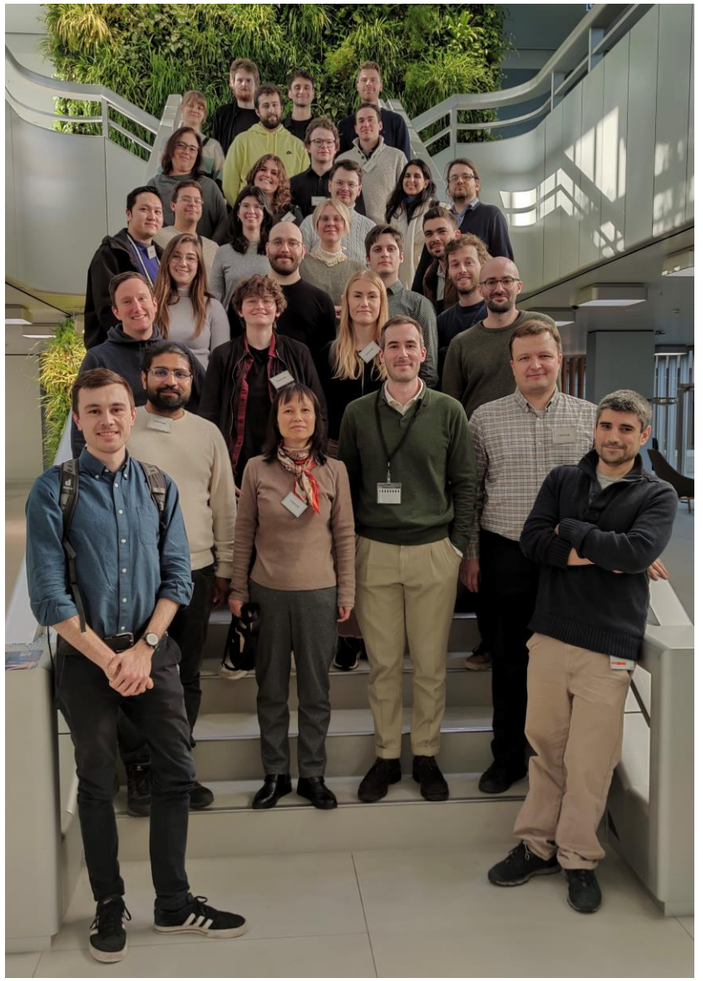
Future plans

- Alternative standardization of data
- Evaluation of DPA parameters
- Testing of different models, descriptors, and DL architectures
- Explanation of predictions
- Applying the same pipeline to Hepatotoxicity and Nephrotoxicity

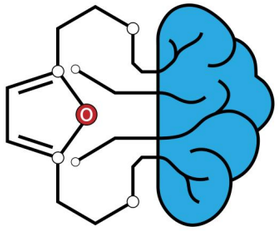


Financing

This study was partially 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!




BAYER


TU/e


ISTITUTO DI RICERCHE
FARMACOLOGICHE
MARIO NEGRI · IRCCS

References

[1] https://doi.org/10.3390/toxics12010087
[2] https://doi.org/10.3389/fphar.2019.01631
[3] https://doi.org/10.1186/s13321-021-00541-z
[4] https://doi.org/10.1186/s12967-019-1918-z
[5] https://doi.org/10.1021/acs.jcim.0c00884
[6] https://doi.org/10.1021/acs.jcim.8b00769
[7] https://doi.org/10.1093/bioinformatics/btaa075
[8] https://doi.org/10.1021/mp700124e
[9] https://doi.org/10.1016/j.comtox.2019.100089
[10] https://doi.org/10.1016/j.crtox.2023.100121
[11] https://doi.org/10.1101/2023.10.15.562398
[12] https://doi.org/10.1021/acs.jcim.7b00641
[13] https://doi.org/10.1016/j.chemolab.2023.104829
[14] https://doi.org/10.1016/j.comtox.2017.05.001
[15] https://doi.org/10.1002/minf.201500040
[16] https://doi.org/10.1021/acs.molpharmaceut.6b00471
[17] https://doi.org/10.1016/j.toxlet.2020.07.003
[18] https://doi.org/10.1002/minf.201200039
[19] https://doi.org/10.1039/d1ra07956e
[20] https://doi.org/10.1273/cbij.21.70
[21] https://doi.org/10.1021/acs.jcim.1c00744
[22] https://doi.org/10.1038/s41598-019-47536-3
[23] https://doi.org/10.2174/1568026614666140506124442
[24] https://doi.org/10.1016/j.tox.2021.153018
[25] https://doi.org/10.48550/arXiv.2112.13467
[26] https://doi.org/10.1021/acs.jcim.8b00150
[27] https://doi.org/10.3390/molecules25112615
[28] https://doi.org/10.1186/s12859-019-2814-5
[29] https://doi.org/10.1016/j.chemolab.2020.104213
[30] https://doi.org/10.1021/acs.jcim.2c00822
[31] https://doi.org/10.3389/fphar.2022.951083
[32] https://doi.org/10.3389/fphar.2020.00639
[33] https://doi.org/10.1021/acs.jcim.3c01301
[34] https://doi.org/10.1016/j.compbiomed.2022.106491
[35] https://doi.org/10.4155/fmc-2020-0156
[36] https://doi.org/10.1016/j.vascn.2020.106895
[37] https://doi.org/10.1177/1074248421995348
[38] https://doi.org/10.1002/cpt.367
[39] https://doi.org/10.1038/aps.2014.35
[40] https://doi.org/10.1039/c5tx00294j
[41] https://doi.org/10.1093/bib/bbac211
[42] https://doi.org/10.1007/s10822-016-9898-z
[43] https://doi.org/10.1002/minf.201700074
[44] https://doi.org/10.1021/tx200099j
[45] https://doi.org/10.1021/mp300023x
[46] https://doi.org/10.3389/fphys.2023.1266084
[47] https://doi.org/10.48550/arXiv.2210.04151
[48] https://doi.org/10.1093/eurheartj/ehab588
[49] https://doi.org/10.3389/fphys.2023.1156286
[50] https://doi.org/10.1016/j.compbiolchem.2020.107286

More references

[51] Dahlöf, B.: Cardiovascular disease risk factors: Epidemiology and risk assessment. The American Journal of Cardiology 105, 3A–9A (2010) https://doi.org/10.1016/j.amjcard.2009.10.007

[52] Sun, D., Gao, W., Hu, H., Zhou, S.: Why 90% of clinical drug development fails and how to improve it? Acta Pharmaceutica Sinica B 12, 3049–3062 (7 2022) https://doi.org/10.1016/j.apsb.2022.02.002

[53] U.S. Food and Drug Administration: FAERS: FDA Adverse Event Reporting System, https://fis.fda.gov/extensions/FPD-QDE-FAERS/FPD-QDE-FAERS.html

[54] Brown, E.G., Wood, L., Wood, S.: The medical dictionary for regulatory activities(MedDRA). Drug Saf. 20(2), 109–117 (Feb 1999)

[55] Drugbank 6.0: the drugbank knowledgebase for 2024. Nucleic Acids Research 52,D1265–D1275 (1 2024). https://doi.org/10.1093/nar/gkad976

[56] Andrew, W.: Front matter. In: Pharmaceutical Manufacturing Encyclopedia, p. iii.Elsevier (2007)

[57] Zdrazil, B., Felix, E., Hunter, F., Manners, E.J., Blackshaw, J., Corbett, S., de Veij,M., Ioannidis, H., Lopez, D.M., Mosquera, J.F., Magarinos, M.P., Bosc, N., Arcila,R., Kizilören, T., Gaulton, A., Bento, A.P., Adasme, M.F., Monecke, P., Landrum,G.A., Leach, A.R.: The ChEMBL Database in 2023: a drug discovery platformspanning multiple bioactivity data types and time periods. Nucleic Acids Research 52(D1), D1180–D1192 (11 2023). https://doi.org/10.1093/nar/gkad1004 https://doi.org/10.1093/nar/gkad1004

[58] NCI CDDD Group: Chemical Identifier Resolver, https://cactus.nci.nih.gov/chemical/structure

[59] Kim, S., Chen, J., Cheng, T., Gindulyte, A., He, J., He, S., Li, Q., Shoe-maker, B.A., Thiessen, P.A., Yu, B., Zaslavsky, L., Zhang, J., Bolton, E.E.: Pub-Chem 2023 update. Nucleic Acids Research 51(D1), D1373–D1380 (10 2022).https://doi.org/10.1093/nar/gkac956, https://doi.org/10.1093/nar/gkac956

[60] Moriwaki, H., Tian, Y.S., Kawashita, N., Takagi, T.: Mordred: a molecular de-scriptor calculator. J. Cheminform. 10(1), 4 (Feb 2018)

Even more references

[61] Landrum, G., Tosco, P., Kelley, B., Rodriguez, R., Cosgrove, D., Vianello, R.,sriniker, gedeck, Jones, G., NadineSchneider, Kawashima, E., Nealschneider, D.,Dalke, A., Swain, M., Cole, B., Turk, S., Savelev, A., Vaucher, A., Wójcikowski, M.,Take, I., Scalfani, V.F., Walker, R., Ujihara, K., Probst, D., guillaume godin, Pahl,A., Lehtivarjo, J., Berenger, F., jasondbiggs, strets123: rdkit/rdkit: 2024_03_1(q1 2024) release (May 2024). https://doi.org/10.5281/zenodo.11102446

[62] Klekota, J., Roth, F.P.: Chemical substructures that enrich for biological activity. Bioinformatics 24(21), 2518–2525 (2008)

[63] Morgan, H.L.: The generation of a unique machine description for chemical structures technique developed at Chemical Abstracts Service. J. Chem. Doc. 5(2),107–113 (May 1965)

[64] Durant, J.L., Leland, B.A., Henry, D.R., Nourse, J.G.: Reoptimization of MDLkeys for use in drug discovery. J. Chem. Inf. Comput. Sci. 42(6), 1273–1280 (Nov2002)

[65] Winter, R., Montanari, F., Noé, F., Clevert, D.A.: Learning continuous and data-driven molecular descriptors by translating equivalent chemical representations. Chemical Science 10, 1692–1701 (2019). https://doi.org/10.1039/C8SC04175J

[66] 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). https://doi.org/https://doi.org/10.1002/pds.677

[67] 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). https://doi.org/https://doi.org/10.1002/pds.1001

[68] 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). https://doi.org/10.1007/s002280050466

[69] Fusaroli, M.: pvda. https://github.com/fusarolimichele/pvda/ (2025)

[70] By Dgamma25 - Own work, CC BY-SA 4.0, https://commons.wikimedia.org/w/index.php?curid=159752276

[71] FAERS Public Dashboard - FAQ — fis.fda.gov.https://fis.fda.gov/extensions/FPD-FAQ/FPD-FAQ.htmlT oc514144622, [Accessed 03–03 – 2025]