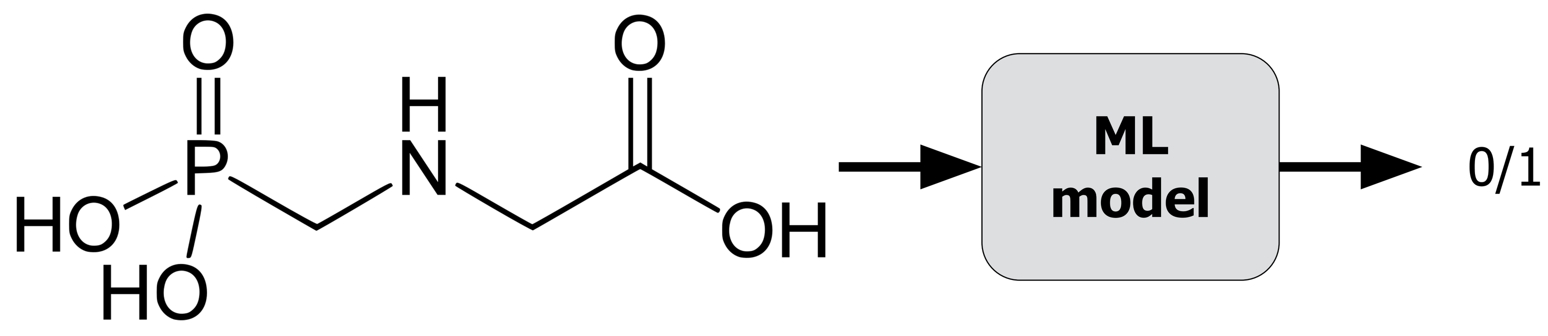
ML in agrochemistry and ecotoxicology

Jakub Adamczyk • ML & Chemoinformatics Lab (MLCIL), Faculty of Computer Science, AGH

Introduction

Molecular property prediction:

- classification / regression on molecular graphs
- e.g. solubility, bioactivity, toxicity
- commonly used in novel drug design and pharmacology



Agrochemistry:

- pesticides, fertilizers, plant growth hormones etc.
- surprisingly outdated - still based on field and lab experiments
- important due to legislative, ecological, and health concerns

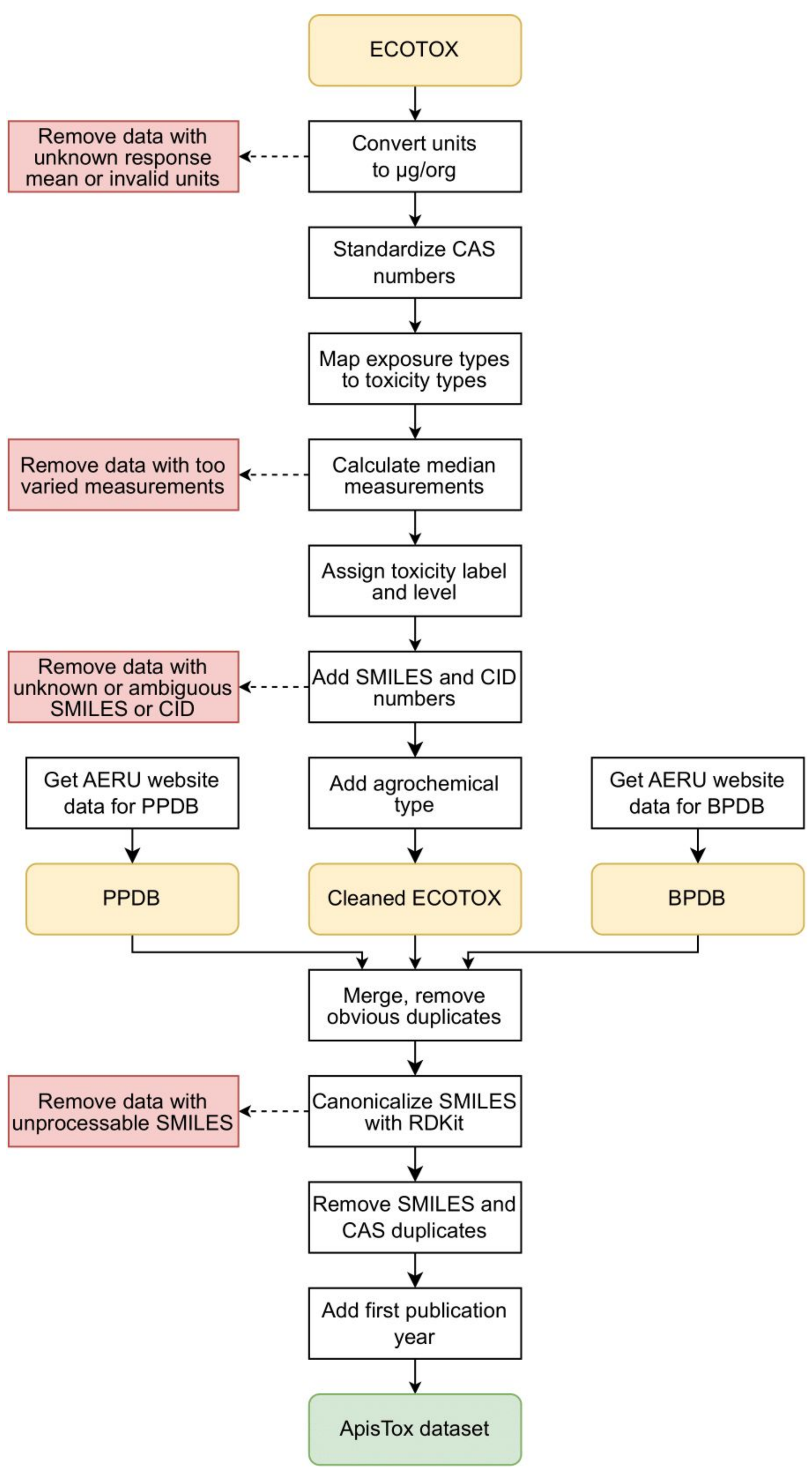
Ecotoxicology:

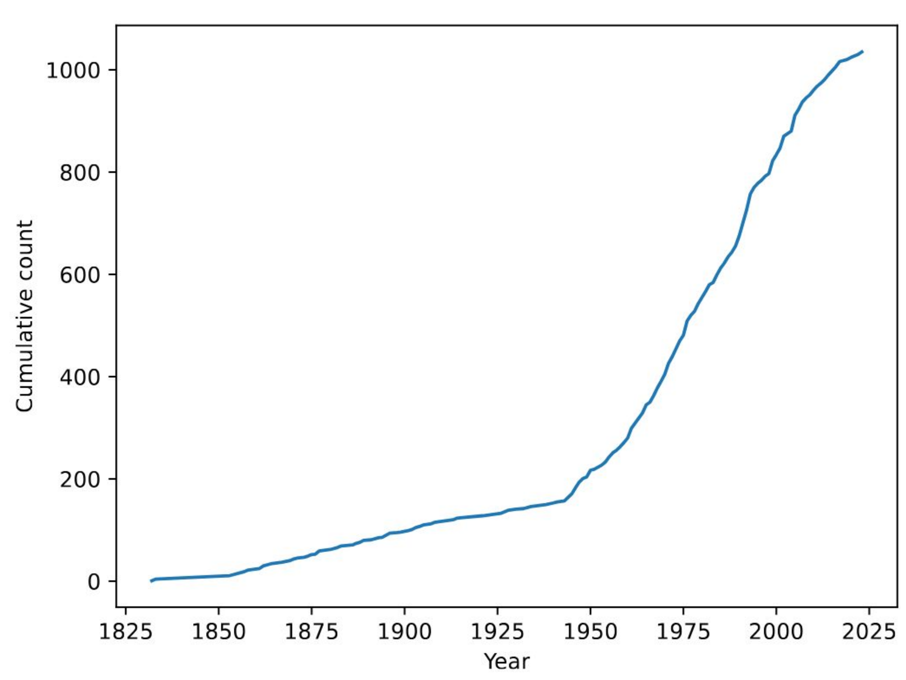
- measuring, modeling & predicting toxicology for animals and plants
- regulatory area for pesticides
- e.g. algae, small mammals, bees
- **really** need predictive models - measuring e.g. LD50 (median lethal dose) requires killing 50% of test organisms

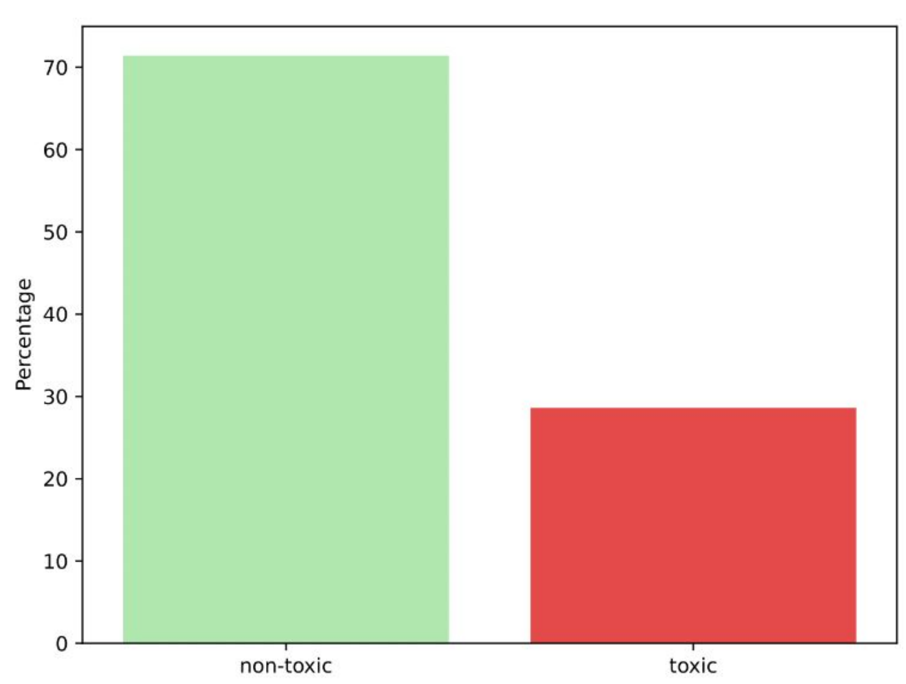
ApisTox dataset

J. Adamczyk, J. Poziemski, P. Siedlecki
"ApisTox: a new benchmark dataset for the classification of small molecules toxicity on honey bees"
Scientific Data

- dataset of pesticides toxicity for honey bees (*Apis Mellifera*)
- **largest:** 1035 compounds
- **best quality:**
  - standardized structures
  - deduplicated
  - unified labels
- **binary classification:**
  - binarized LD50 values
  - regression → classification
- **workflow:**
  - 3 data sources
  - cleaned & merged
  - additional metadata
- **challenging:**
  - atypical structures, salts etc.
  - differs from medicinal chem.
  - time & maximum diversity train-test splits

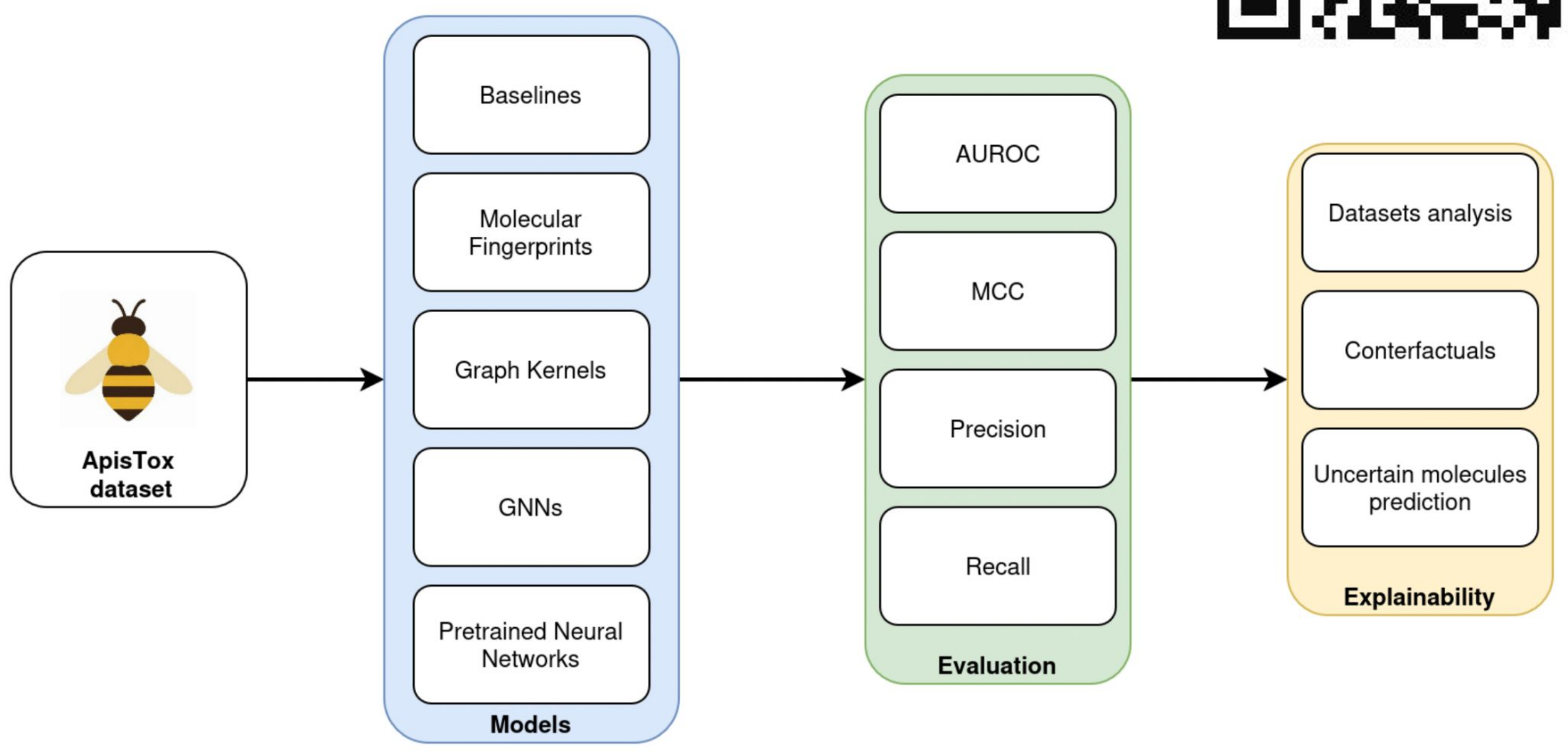






ML models

J. Adamczyk, J. Poziemski, P. Siedlecki
"Evaluating machine learning models for predicting pesticides toxicity to honey bees"
ArXiv preprint



Model groups:

- molecular fingerprints
- feature extraction baselines
- graph kernels + SVM
- graph neural networks (GNNs)
- embeddings from pretrained models

Evaluation:

- time split
- Matthews Correlation Coefficient

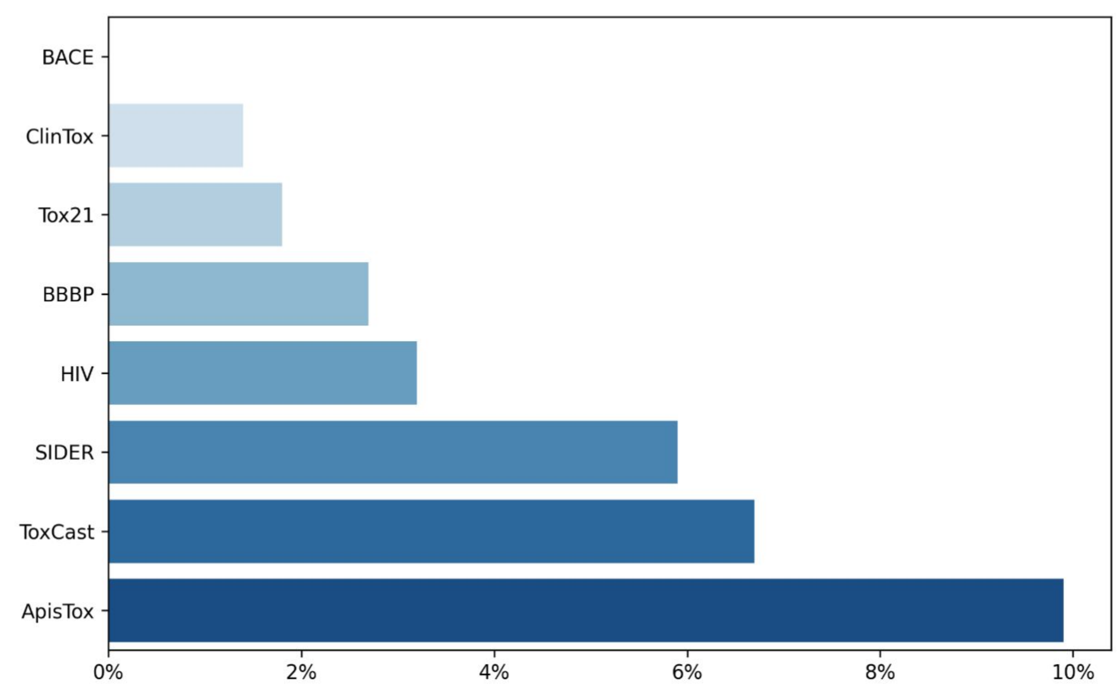
Results:

- molecular fingerprints, particularly ECFP, win by a wide margin
- graph kernels are strong
- GNNs and pretrained embeddings did not even outperform baselines

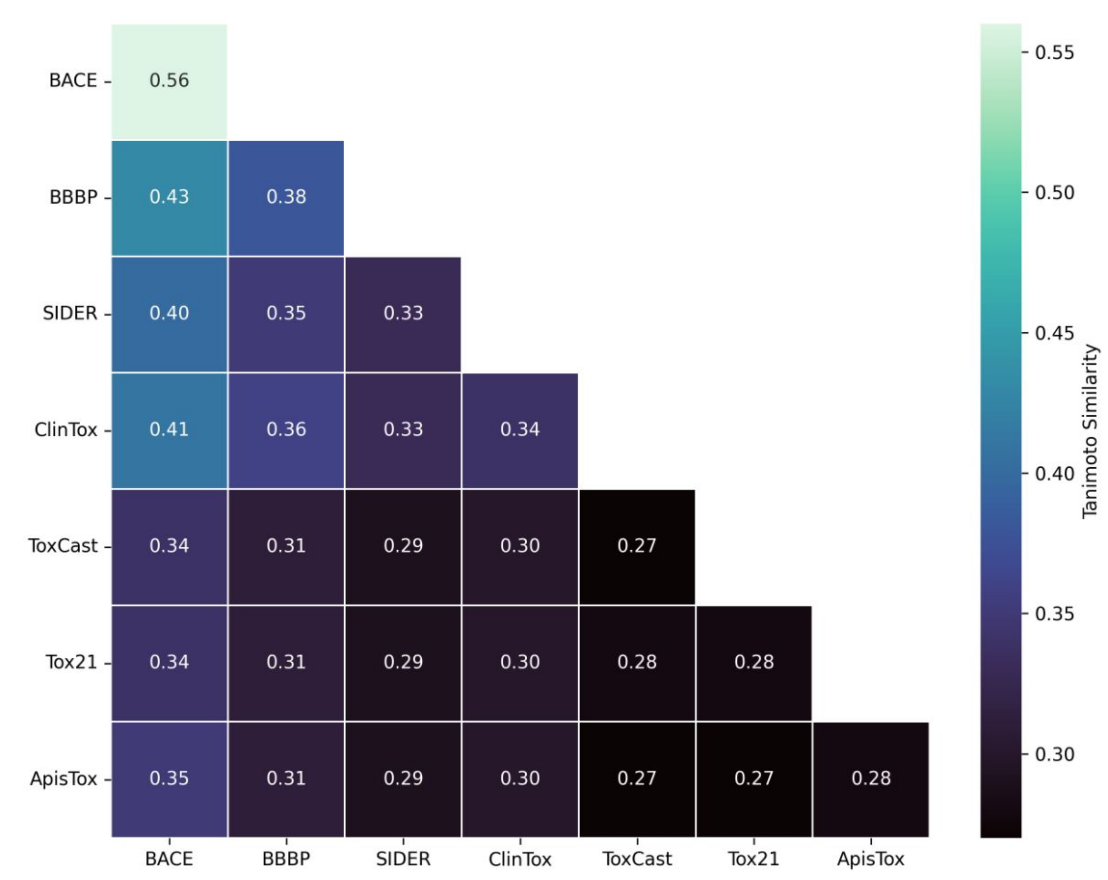
Group	Method	MCC
Fingerprints	ECFP	**0.48 ± 0.02**
	Layered	0.46 ± 0.02
	RDKit	0.46 ± 0.02
	SECFP	0.45 ± 0.03
	Topological Torsion	0.44 ± 0.02
Baselines	Atom counts	0.29 ± 0.04
	LTP	0.23 ± 0.01
	MOLTOP	**0.33 ± 0.01**
Graph kernels	Propagation	0.36
	Shortest paths	0.31
	WL	0.41
	WL-OA	**0.43**
GNNs	GCN	0.30 ± 0.04
	GraphSAGE	**0.33 ± 0.04**
	GIN	0.32 ± 0.06
	GAT	0.26 ± 0.05
	AttentiveFP	0.29 ± 0.06
Pretrained neural networks	MAT	0.25
	R-MAT	0.29
	GROVER	0.05
	ChemBERTa	0.27
	Mol2Vec	**0.31**

Why?

- agrochemistry is **very** different from medicinal chemistry
- very **diverse** and complex domain of chemistry
- existing neural models are **overtuned** for small, overused benchmarks



- a lot of molecules with **"non-medicinal" atoms**
- e.g. metals, toxic functional groups
- many models for medicinal chemistry ignore those



- how to evaluate diversity?
- average pairwise similarity between molecules
- ECFP + Tanimoto similarity
- ApisTox is **very diverse** internally
- it is also **orthogonal** to existing benchmark