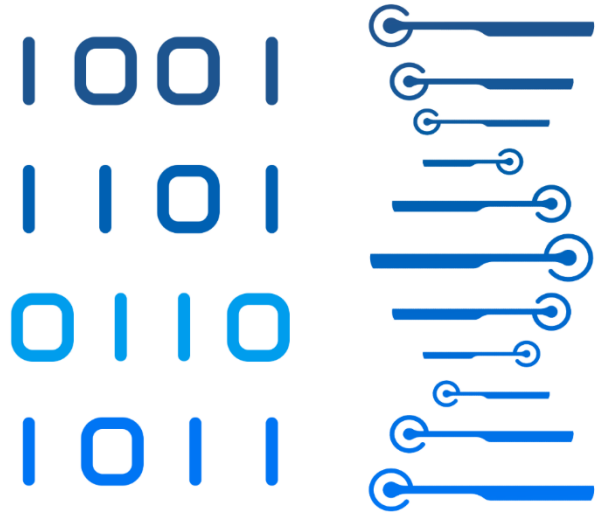
IBMC
Institute of Biomedical Chemistry

Way2Drug PREDICTIVE SERVICES
Understanding Chemical-Biological Interactions

A WORKFLOW FOR BUILDING (Q)SAR MODELS WITH IMBALANCED AND HETEROGENEOUS DATA

L. Stolbov, D. Druzhilovsky, A. Rudik, D. Filimonov , V. Poroikov

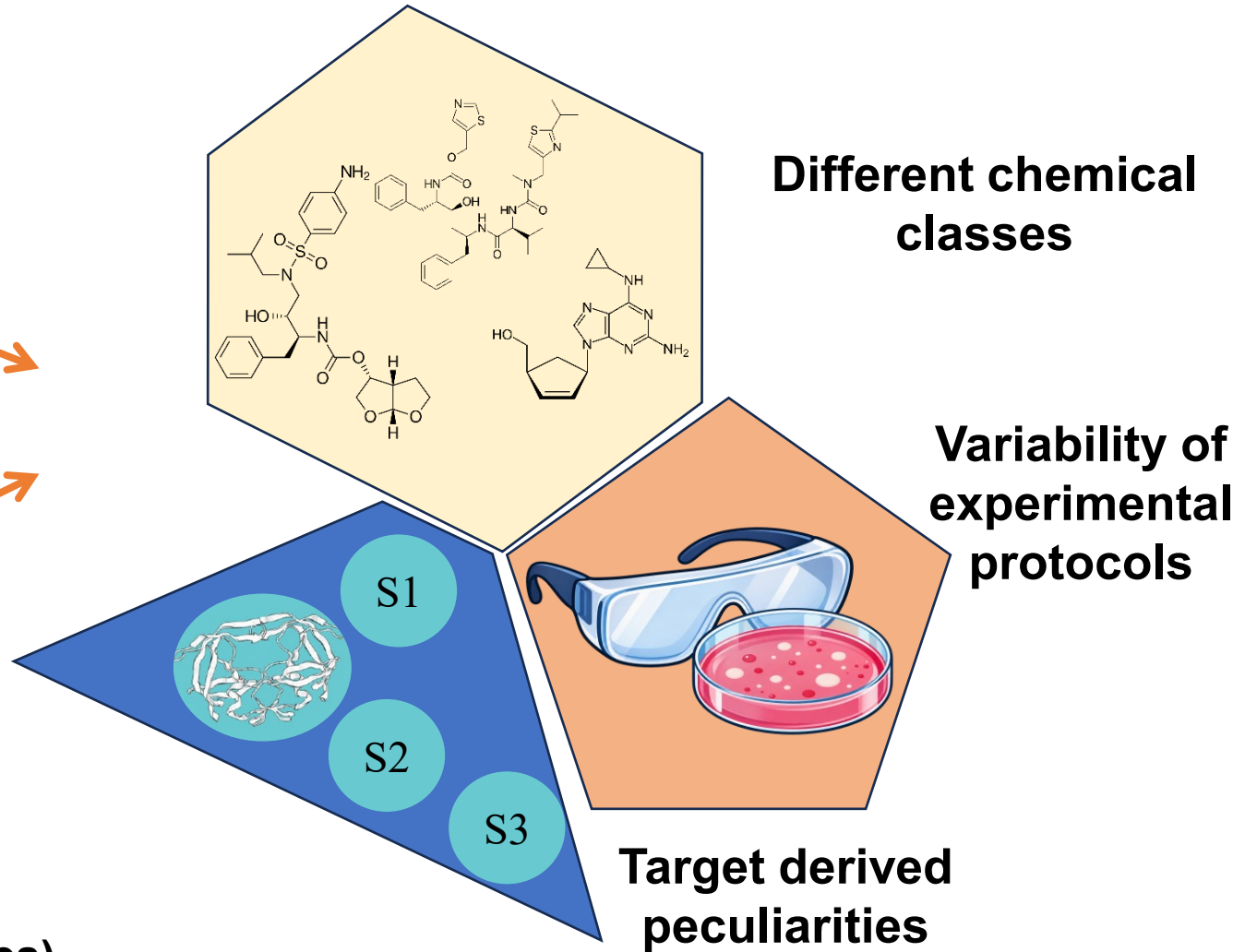


BCADD-2026

Heterogeneous data in cheminformatics/bioinformatics

Extracted literature data/Data from patents

Aggregated data from DBs



Different chemical classes

Variability of experimental protocols

Target derived peculiarities

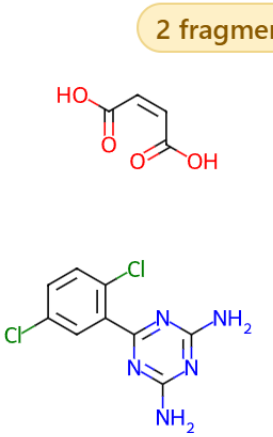
Training set

- Data is heterogeneous (chemical classes, experimental protocols, biological activities)
- Classification of the values produces imbalanced classes

Data Curation

Actions before using the dataset:

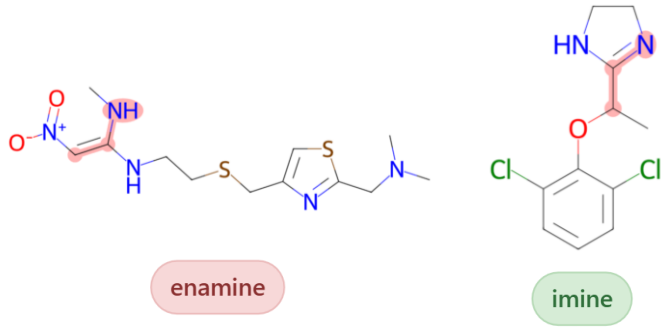
- Kekulization/Aromatization
- Normalization
- Reionization
- Explicit hydrogen handling
- Mixtures and multicomponent entries solution



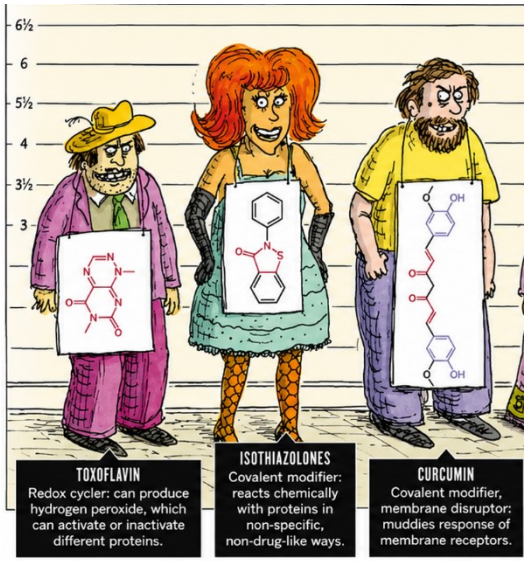
- Checking the isotopes/rare atoms



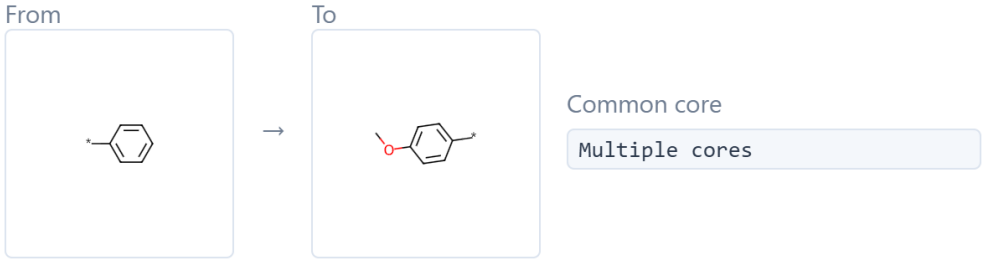
- Tautomer standardization



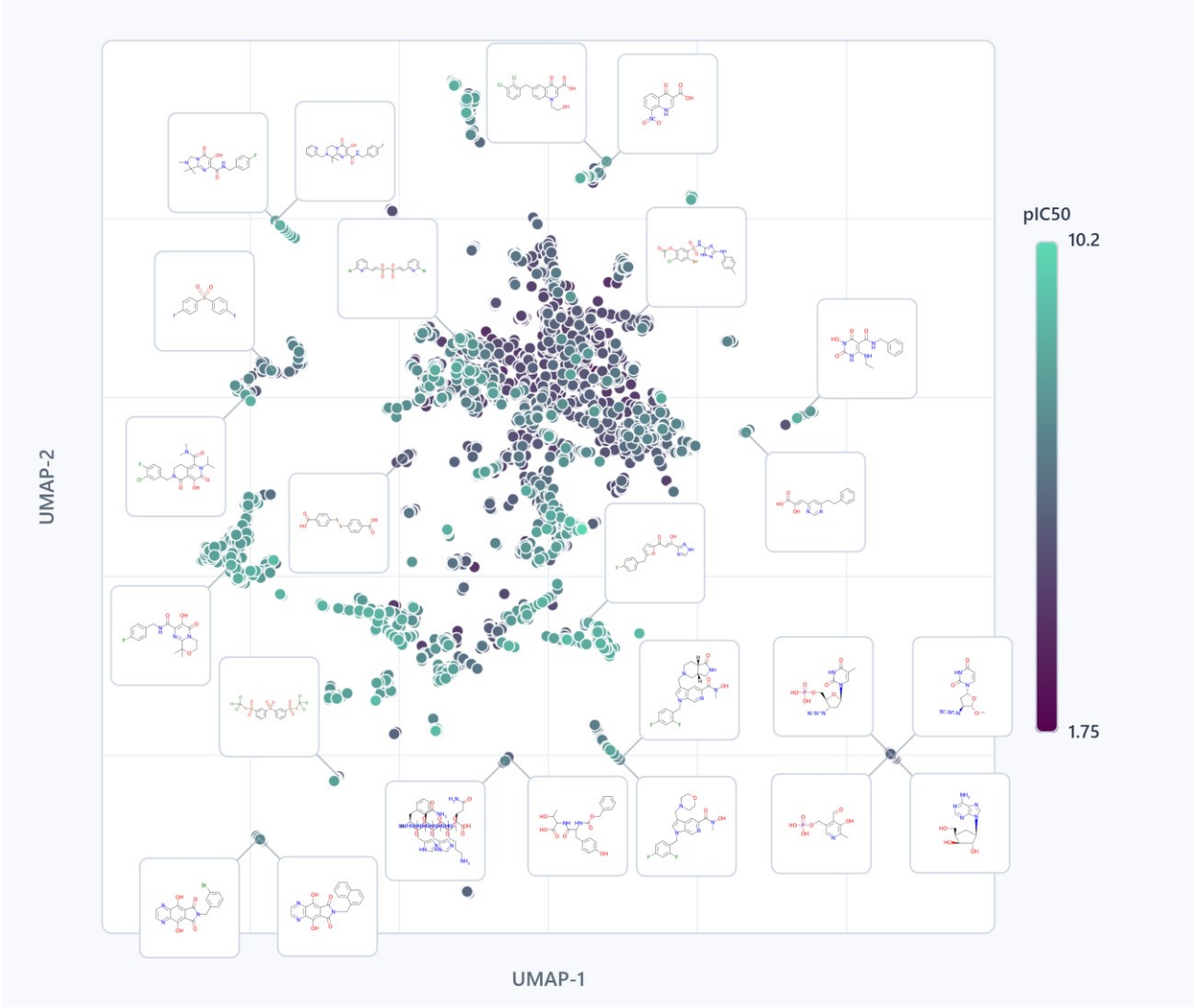
- Detection of “bad” substructures (like PAINS)



Exploratory analysis of chemical structures



Matched Pairs Analysis



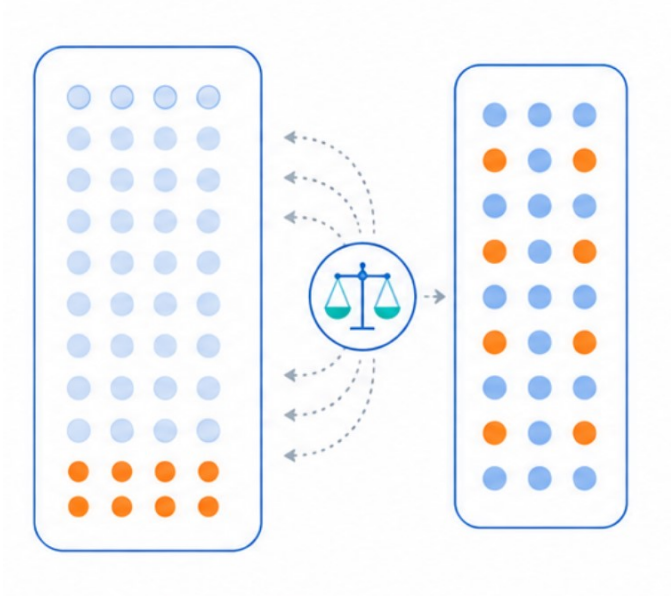
Chemical Space Analysis

Self Consistent Extreme Classifier (SCEC)



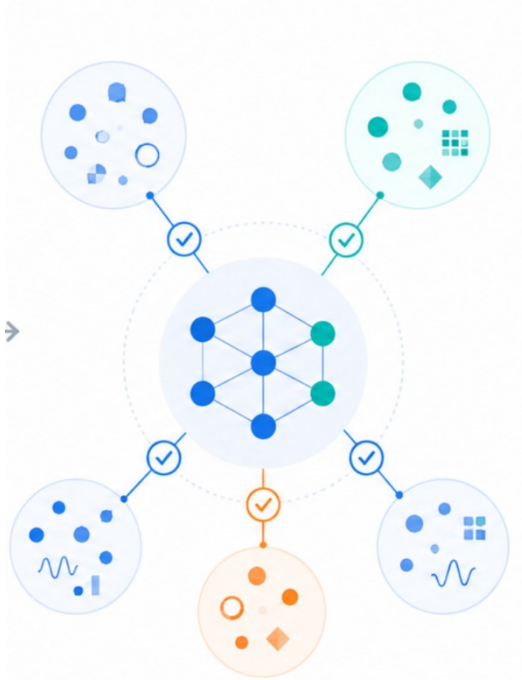
Dimensionality reduction

3x reduction compared to SVM and similar approaches



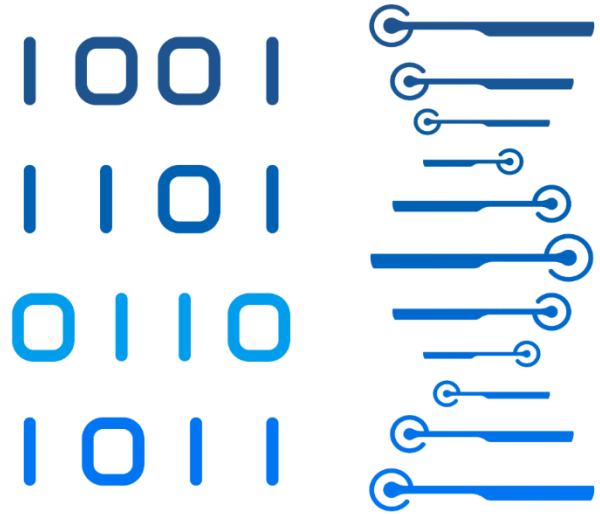
Imbalanced classes handling

The data for a larger class enhances the prediction for smaller class



Generalized predictivity

New structures are predicted more accurately or dropped out of applicability domain

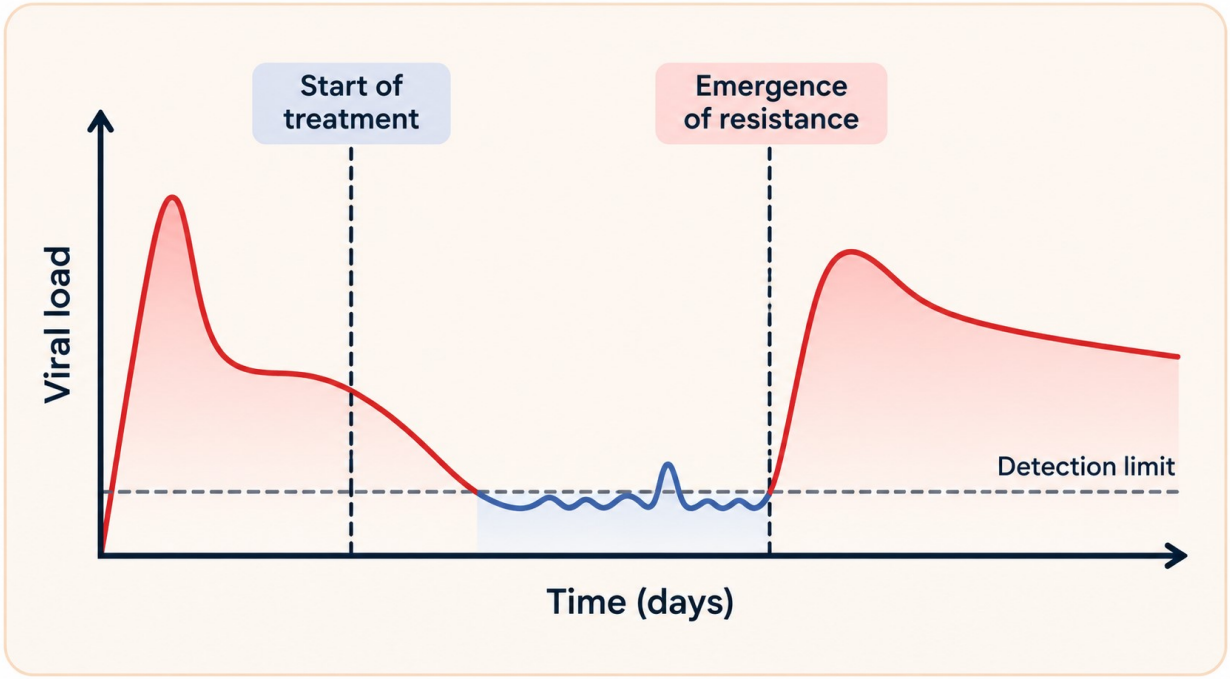


Structure-Activity Relationships Models for HIV-1 enzymes

New HIV-1 inhibitors search

Current treatment:

- Does not lead to full recovery
- Leads to side effects
- Leads to emergence of resistance

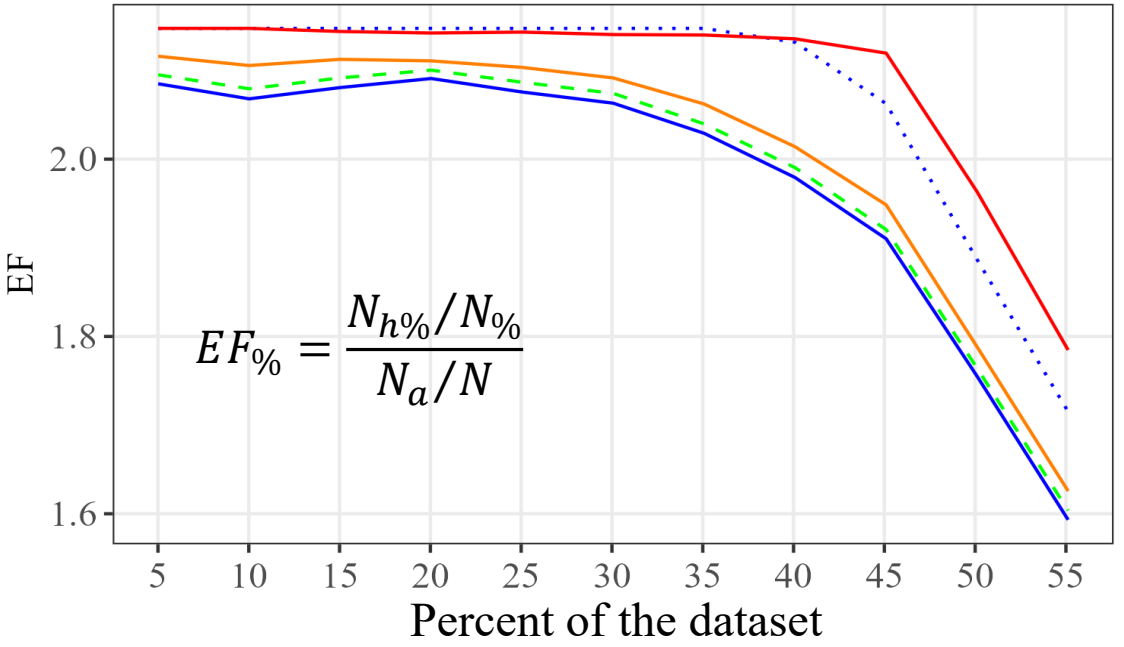


HIV-1 load dynamics

SCEC tested for potential virtual screening

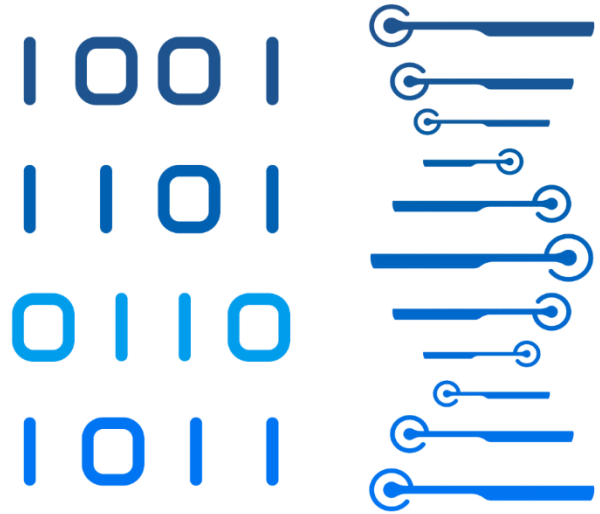
HIV-1 Protease

Approach	Set	N	V	BA	AUC
SCR	PR	6552	470	0.860	0.932
SCEC	PR	6552	130	0.866	0.941
SCEC	PR (+Ki)	8545	163	0.889	0.953
SCEC	PR(+Ki +inactive)	12000	190	**0.891**	**0.954**



Results for Integrase and Reverse Transcriptase showed same trend

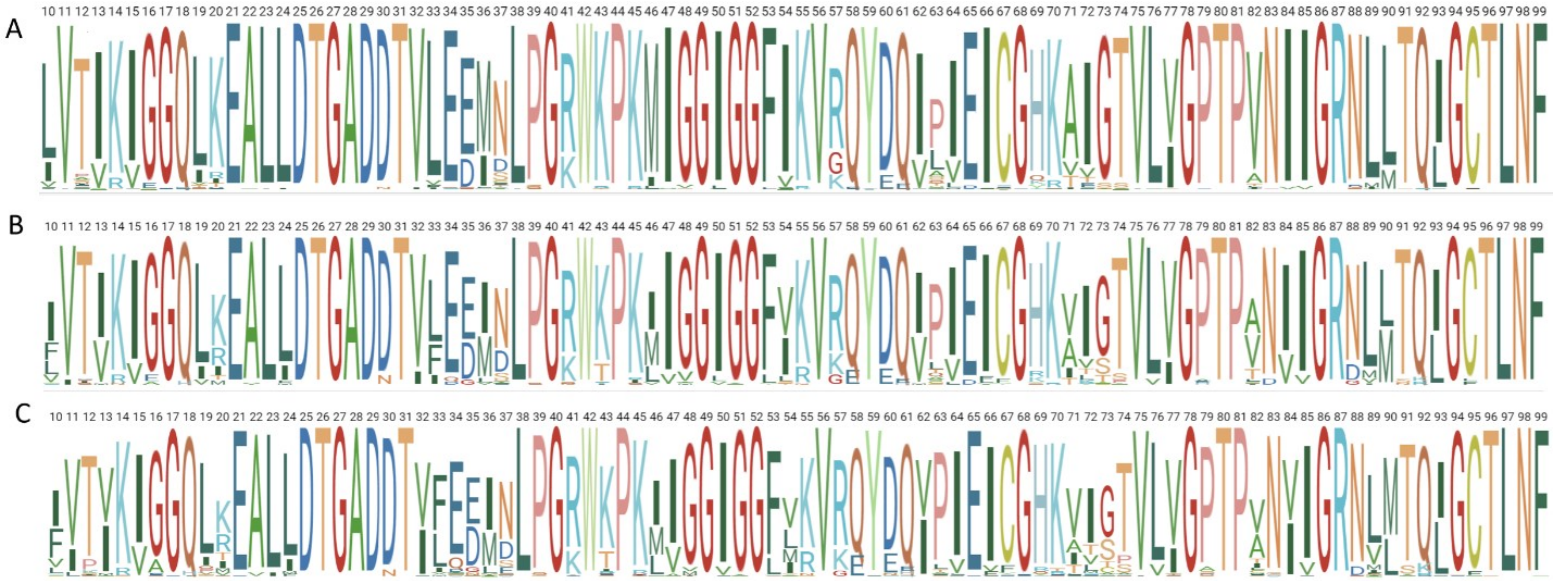
Blue *– comparison*
Blue dotted *– quantitative model*
Green dashed *– SCEC*
Orange *– SCEC (+ Ki)*
Red *– SCEC (+ Ki + inactive)*



Ternary classification of HIV resistance

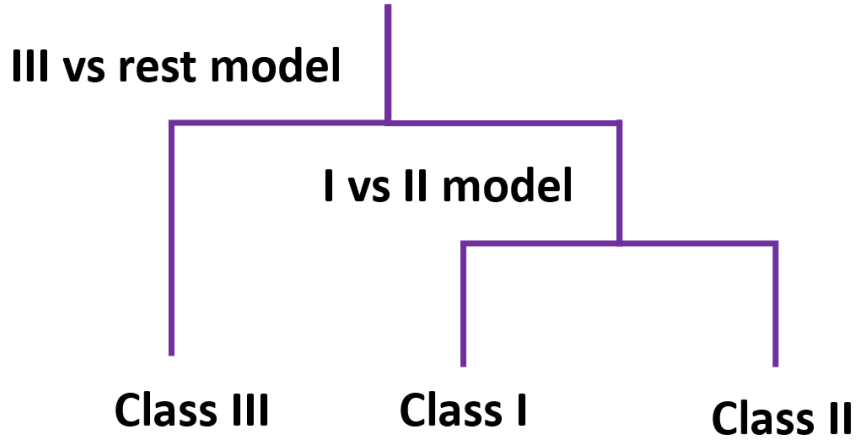
HIV resistance data

- Therapy schemas require the estimation of resistance versus existing drugs for different viral strains
- Not all sequences are present in available databases
- Moderate resistance class is quite narrow



A – No resistance
B – Moderate resistance
C – High resistance

Ternary classification with SCEC

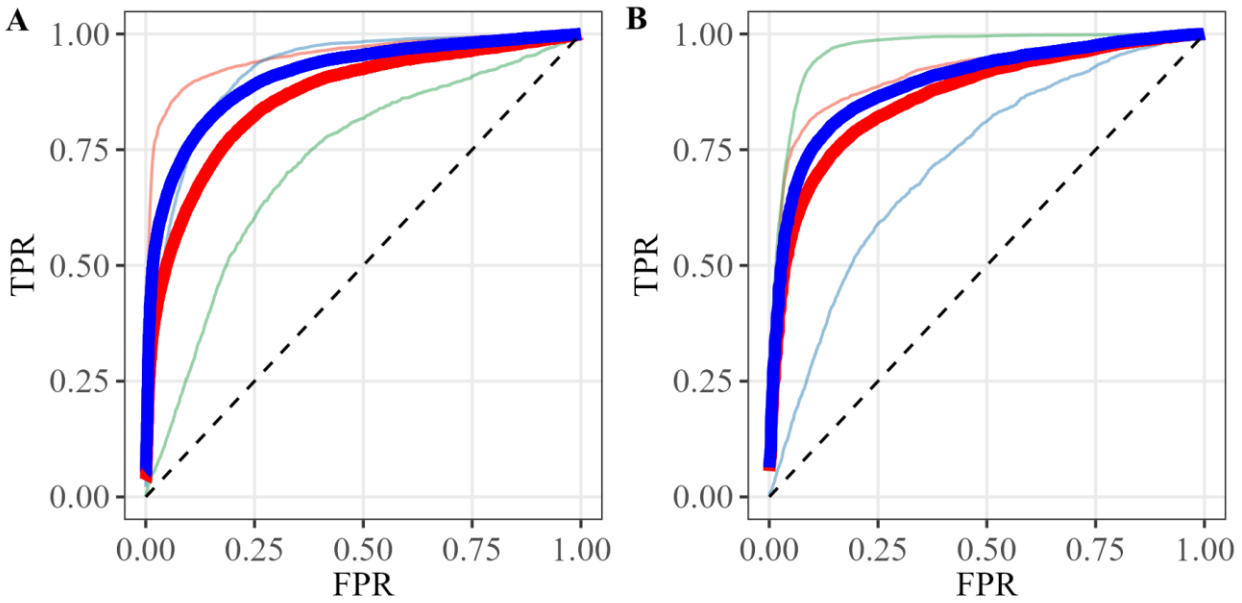


1.Relates to clinical recommendations

2. Applies to continuous nature of resistance

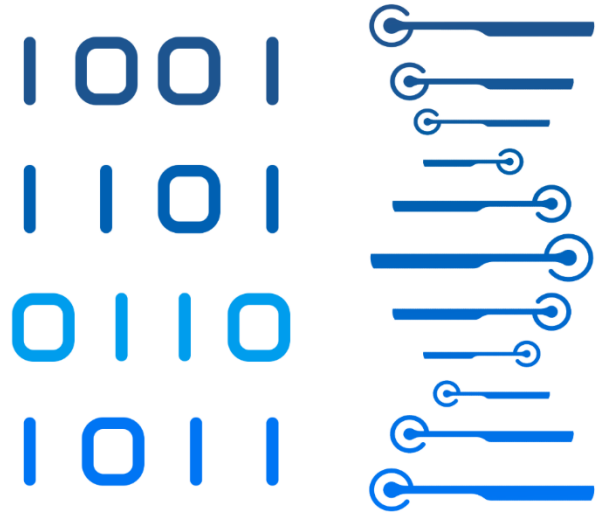
3. Enhances the interpretability of models

4. Explain complex clinical case when the level of resistance is crucial



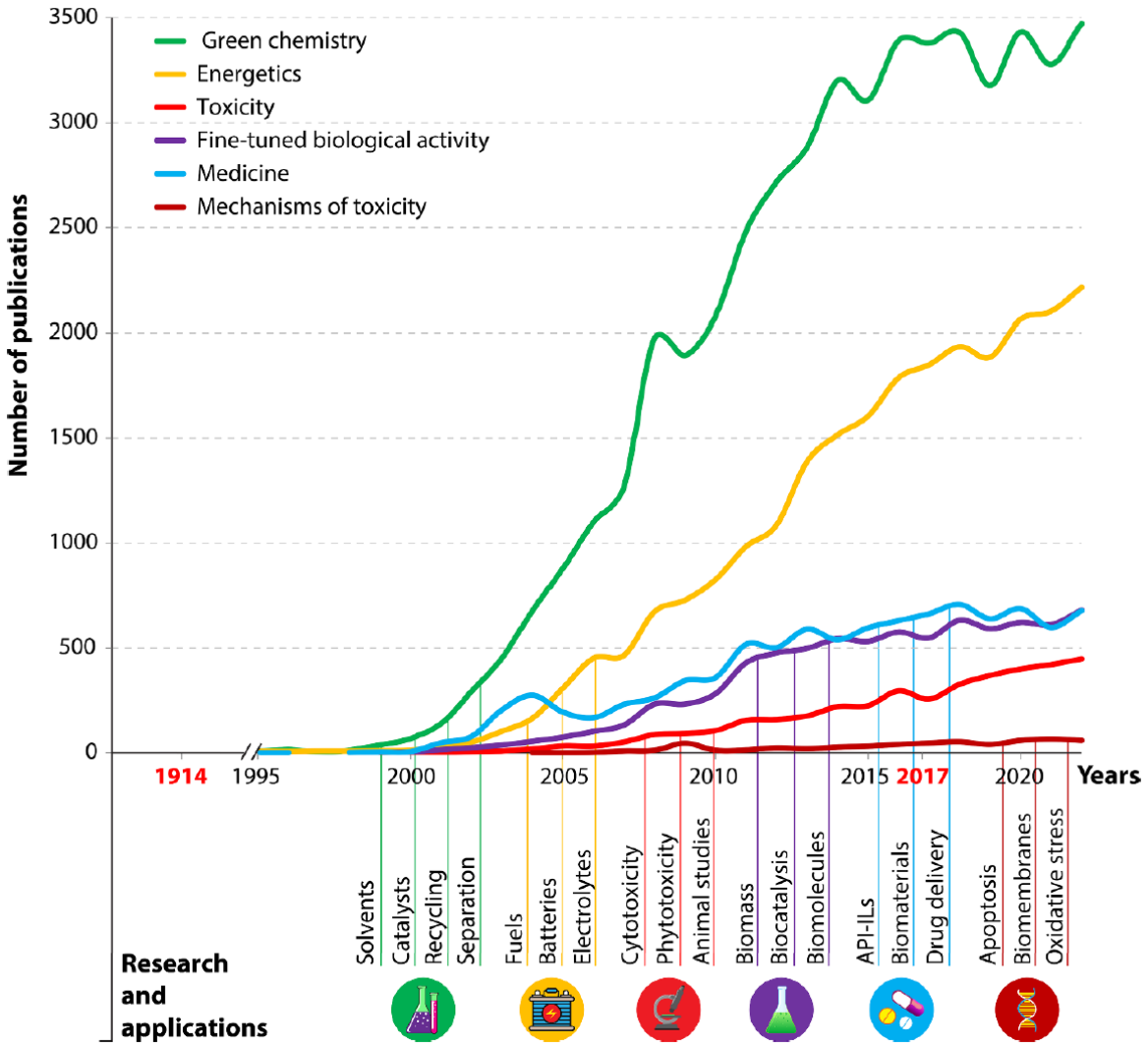
OvR (One versus rest) and OvO (One versus one) validation

External test set validation showed the applicability of the models



QSPR for ion liquids cytotoxicity

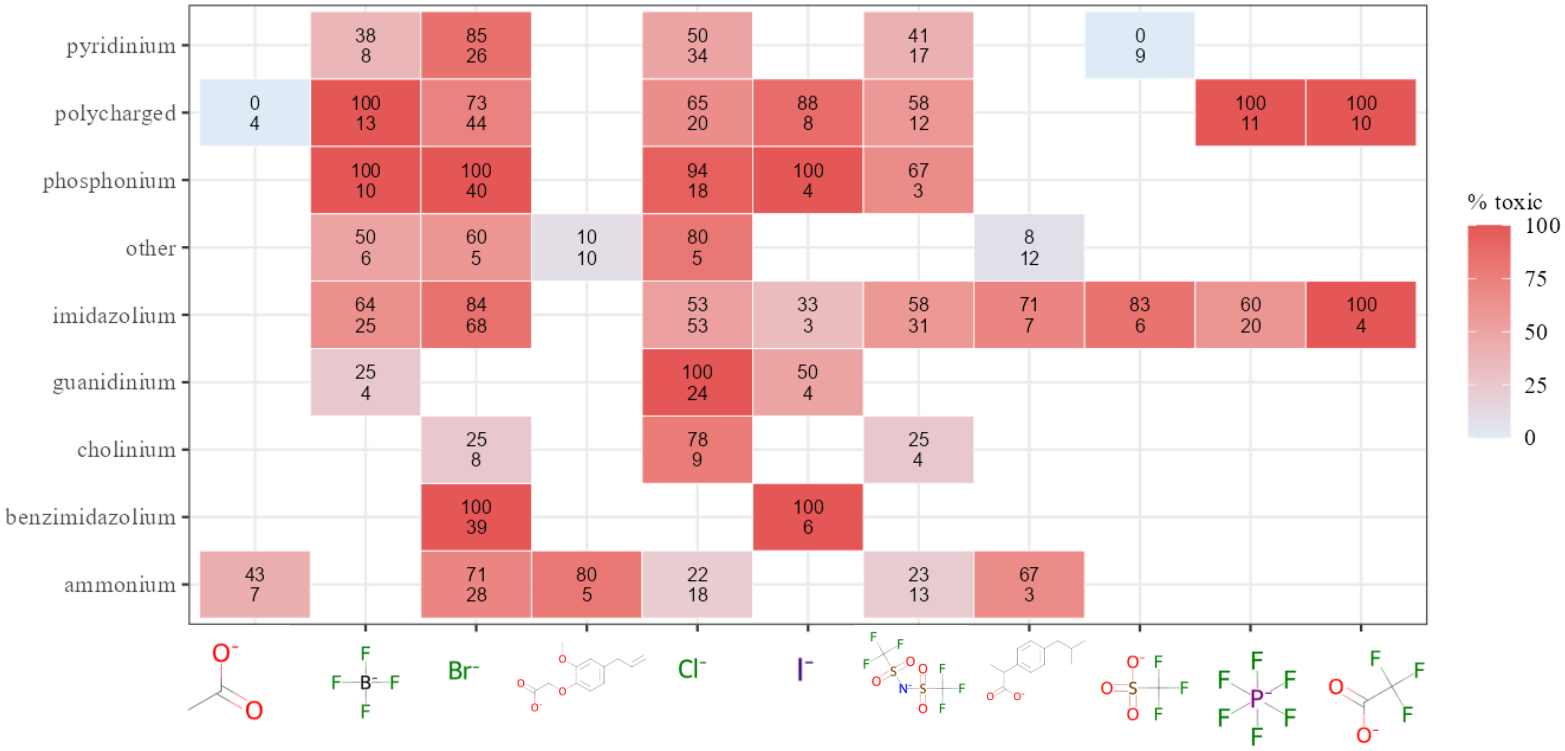
Ion liquids

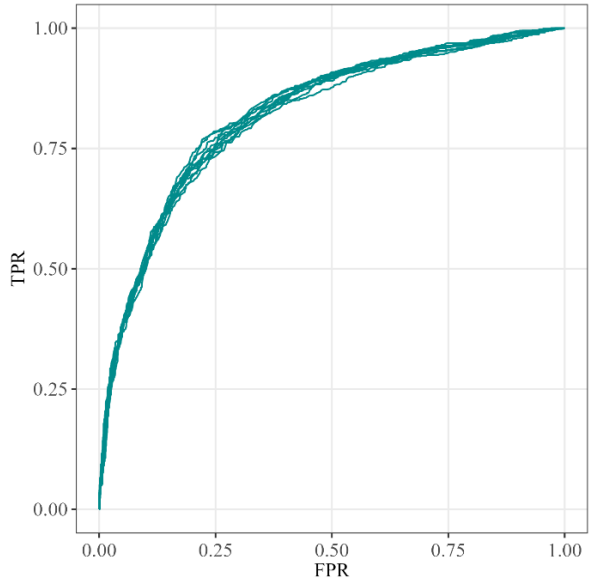


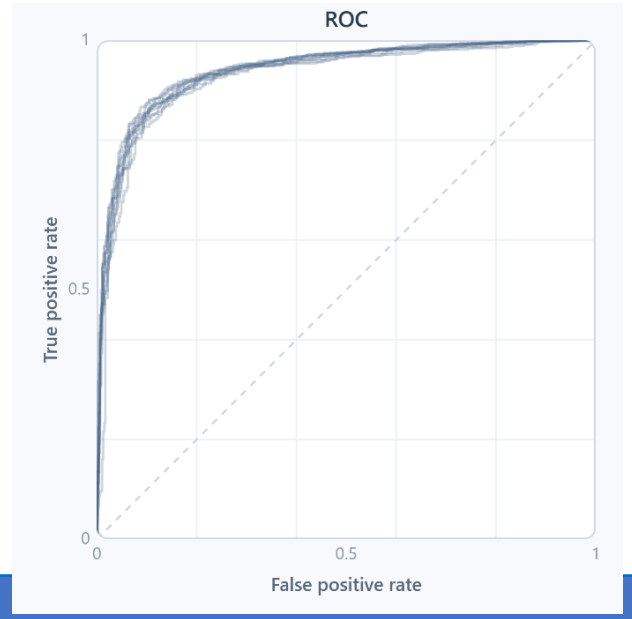
- **Ion liquids – solutions of organic salts** that model some environment for a specific task
- **Key features:** Unique set of physical and chemical properties, low steam pressure, thermic stability, improved solubility for other compounds
- **Issues:** Wide application to biological systems leads to a potential toxicity not quite explored

S. Egorova et al., Mechanisms of Biological Effects of Ionic Liquids: From Single Cells to Multicellular Organisms, Chemical reviews, DOI: 10.1021/acs.chemrev.3c00420

How analysis enhances the model







ROC AUC changed from 0.726 to 0.875 at 5-fold cross validation with 0.981 ROC AUC for external test set

Conclusions

- **High accuracy** for heterogenic and imbalanced data
- **Internal mechanism** for feature selection
- **Multipurpose:** working with different models, data, issue background

All the pipeline is implemented in ***SarMath software*** *available at late 2026 as both web or desktop application*