XXIX Symposium on Bioinformatics and
Computer-Aided Drug Discovery
18-20 September 2023, on-line

Machine Learning Prediction of Mycobacterial Cell Wall Permeability of Drugs and Drug-like Compounds

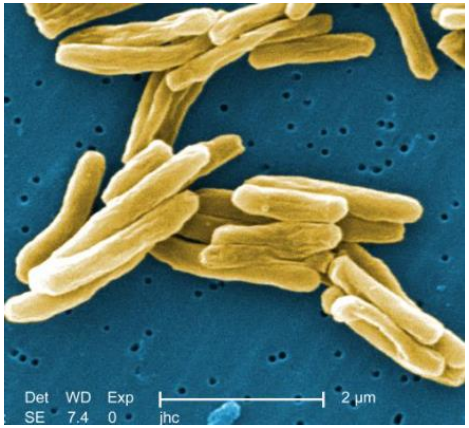
E.V. Radchenko[1,2], *G.V. Antonyan*[1], *S.K. Ignatov*[2], *V.A. Palyulin*[1,2]

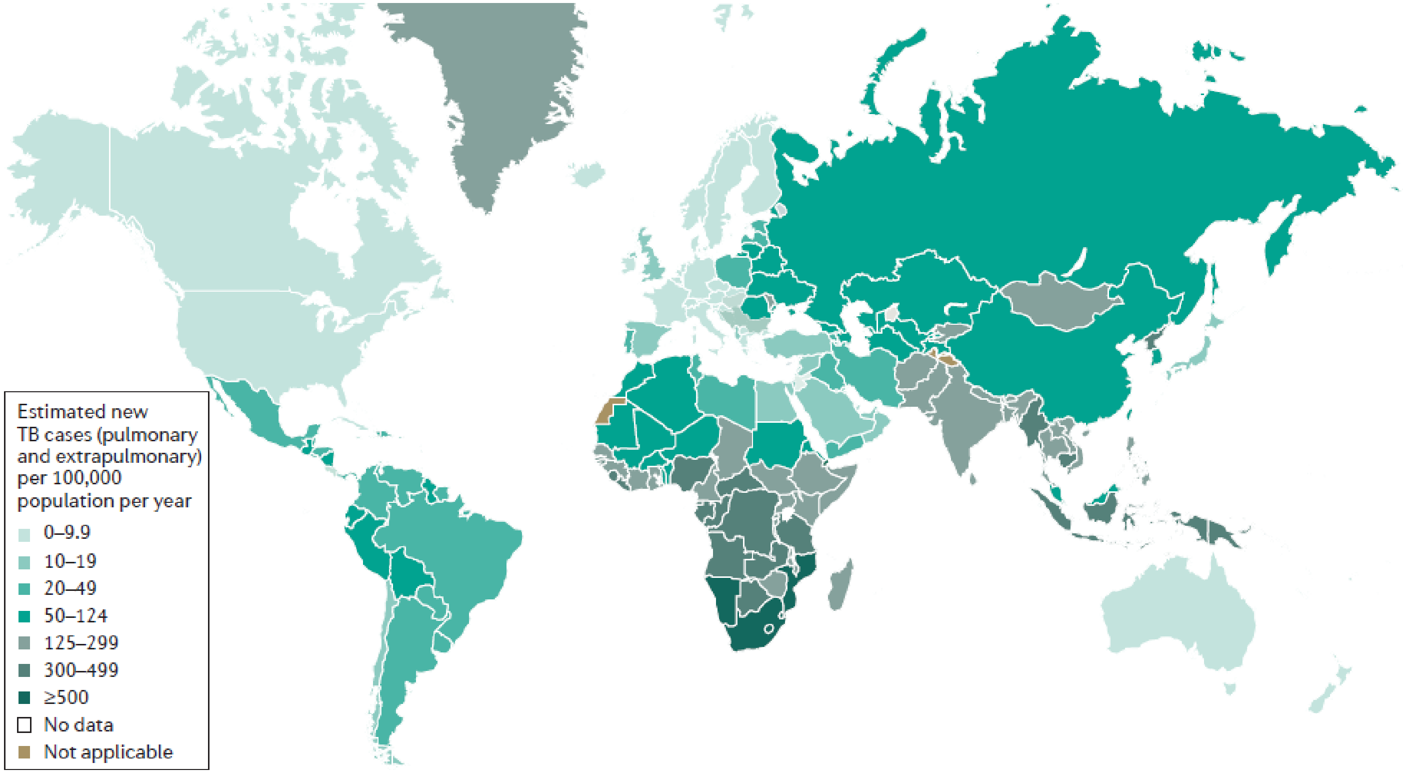
[1] Lomonosov Moscow State University, Russia
[2] Lobachevsky State University of Nizhny Novgorod, Russia

genie@qsar.chem.msu.ru

Drug-resistant tuberculosis (TB), a major global health challenge

- TB is caused by the pathogenic *Mycobacterium tuberculosis (Mtb)*
- One of the most widespread and socially significant infections
- Every year, 1.6 million people die worldwide, making TB the leading cause of death from a single infectious agent
- New emerging strains of mycobacteria: multidrug-resistant (MDR) and extensively drug-resistant (XDR) tuberculosis
- HIV-associated TB




Estimated new TB cases (pulmonary and extrapulmonary) per 100,000 population per year
0–9.9
10–19
20–49
50–124
125–299
300–499
≥500
No data
Not applicable

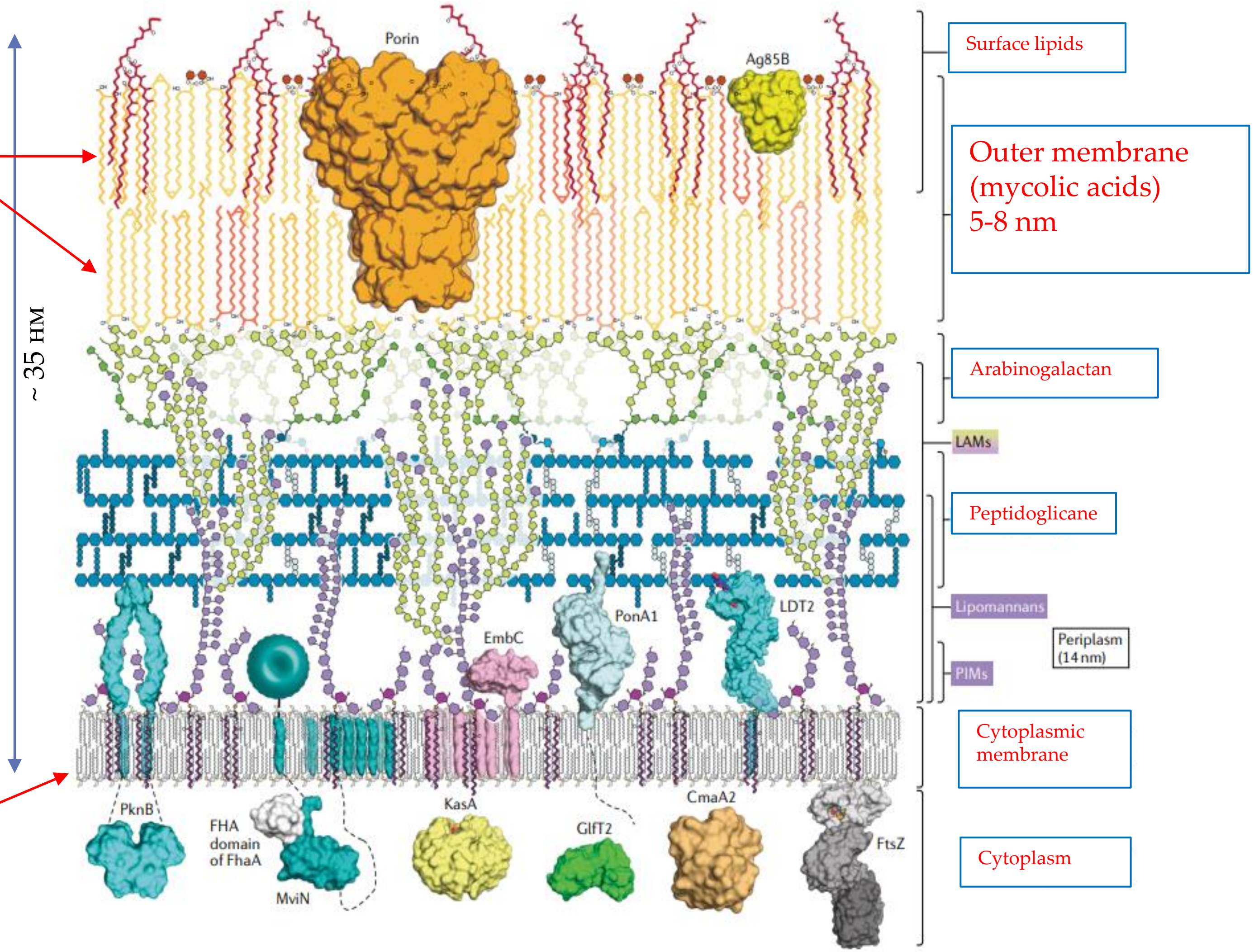
Severely lacking tuberculosis therapy options

- About a dozen antibiotic agents belonging to several drug classes are used clinically
- Often limited efficacy, long and inconvenient regimens, combination therapies
- Often toxicity and other adverse effects
- High risk of preexisting or developing drug resistance

- Massive worldwide efforts to identify novel promising anti-TB drug targets and active compounds
- Target-oriented drug development and optimization often unsuccessful
- High attrition rates
- Many compounds are potent against isolated targets but lack activity in whole-cell or *in vivo* settings
- One of the key causes: low penetration of a drug into *Mtb* cells

Key factor of *Mtb* resilience is its extremely complicated and persistent cell wall

Thick and dense outer membrane of mycolic acids: long molecules with hydrocarbon chains of ~70-90 carbon atoms



Surface lipids

Outer membrane (mycolic acids) 5-8 nm

Arabinogalactan

Peptidoglicane

Cytoplasmic membrane

Cytoplasm

Also contains various porins, efflux pumps, and transporters

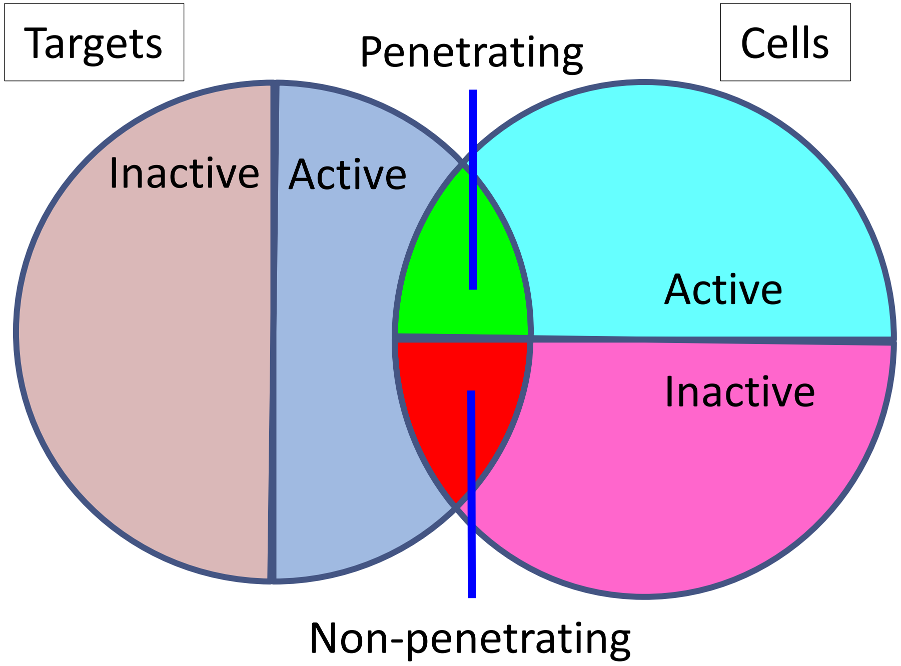
“Normal” lipid membrane

Figure from [Dulberger C.L.et al. *Nat Rev Microbiol*, 2020, **18**, 47]

Prediction and optimization of *Mtb* “pharmacokinetics”

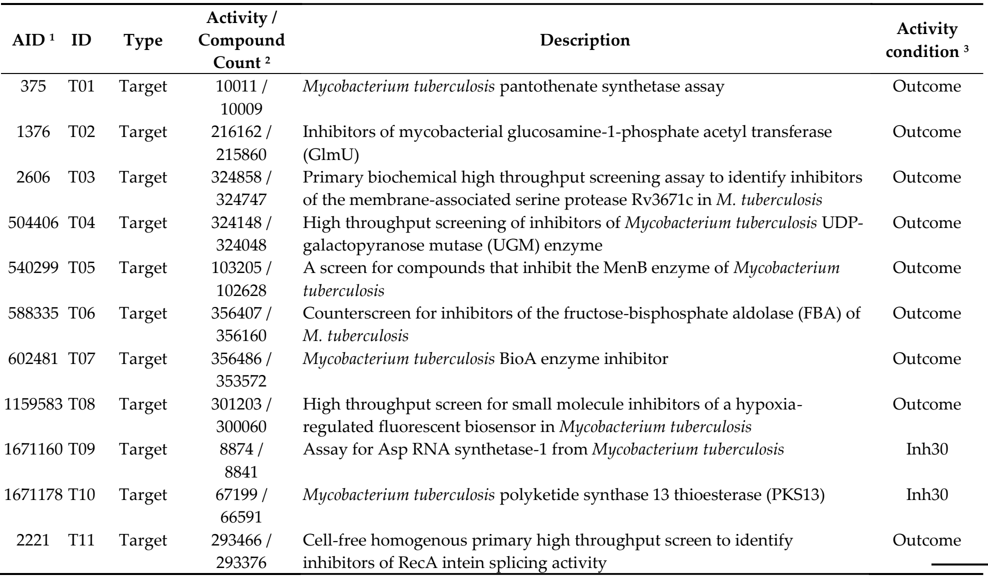
- Explicit modeling of drug permeation promising but complicated
- Effective complementary approach: use general QSAR methodology to derive predictive machine learning models

- Key challenge: lack of direct measurements of permeability
- Solution: indirect estimation from comparison of the target and whole-cell activities [originally proposed for the MycPermCheck model, *Merget et al., Bioinformatics 2013, 29, 62–68*]
- Implicitly captures not only membrane permeation but also active transport/efflux and inactivation



Mtb permeability datasets based on Big Data analysis

- Extensive anti-TB bioassay data are available in PubChem 2022

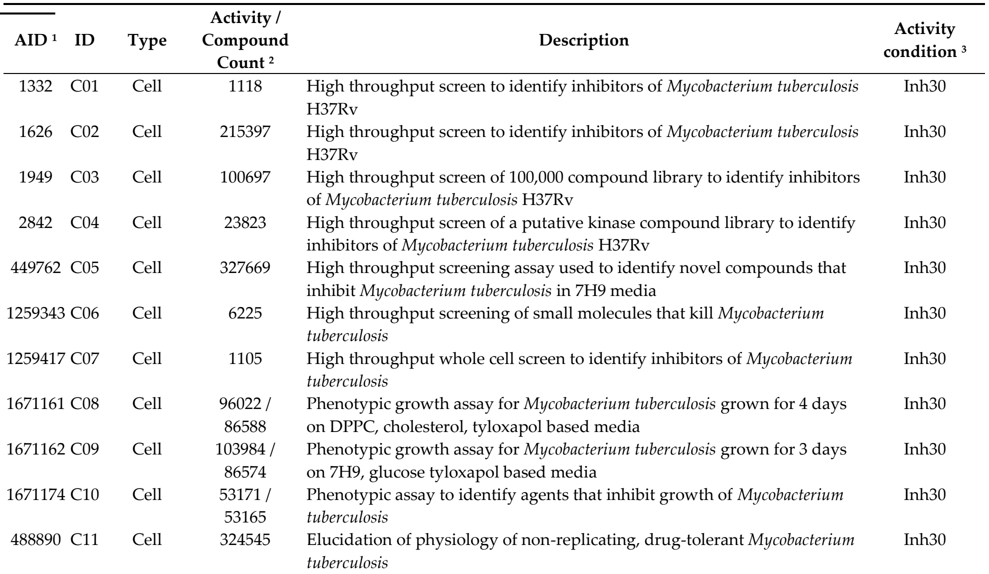


AID [1]	ID	Type	Activity / Compound Count [2]	Description	Activity condition [3]
375	T01	Target	10011 / 10009	*Mycobacterium tuberculosis* pantothenate synthetase assay	Outcome
1376	T02	Target	216162 / 215860	Inhibitors of mycobacterial glucosamine-1-phosphate acetyl transferase (GlmU)	Outcome
2606	T03	Target	324858 / 324747	Primary biochemical high throughput screening assay to identify inhibitors of the membrane-associated serine protease Rv3671c in *M. tuberculosis*	Outcome
504406	T04	Target	324148 / 324048	High throughput screening of inhibitors of *Mycobacterium tuberculosis* UDP-galactopyranose mutase (UGM) enzyme	Outcome
540299	T05	Target	103205 / 102628	A screen for compounds that inhibit the MenB enzyme of *Mycobacterium tuberculosis*	Outcome
588335	T06	Target	356407 / 356160	Counterscreen for inhibitors of the fructose-bisphosphate aldolase (FBA) of *M. tuberculosis*	Outcome
602481	T07	Target	356486 / 353572	*Mycobacterium tuberculosis* BioA enzyme inhibitor	Outcome
1159583	T08	Target	301203 / 300060	High throughput screen for small molecule inhibitors of a hypoxia-regulated fluorescent biosensor in *Mycobacterium tuberculosis*	Outcome
1671160	T09	Target	8874 / 8841	Assay for Asp RNA synthetase-1 from *Mycobacterium tuberculosis*	Inh30
1671178	T10	Target	67199 / 66591	*Mycobacterium tuberculosis* polyketide synthase 13 thioesterase (PKS13)	Inh30
2221	T11	Target	293466 / 293376	Cell-free homogenous primary high throughput screen to identify inhibitors of RecA intein splicing activity	Outcome

Target-based assays

Total 926,660 compounds

9450 compounds active in at least one assay



AID [1]	ID	Type	Activity / Compound Count [2]	Description	Activity condition [3]
1332	C01	Cell	1118	High throughput screen to identify inhibitors of *Mycobacterium tuberculosis* H37Rv	Inh30
1626	C02	Cell	215397	High throughput screen to identify inhibitors of *Mycobacterium tuberculosis* H37Rv	Inh30
1949	C03	Cell	100697	High throughput screen of 100,000 compound library to identify inhibitors of *Mycobacterium tuberculosis* H37Rv	Inh30
2842	C04	Cell	23823	High throughput screen of a putative kinase compound library to identify inhibitors of *Mycobacterium tuberculosis* H37Rv	Inh30
449762	C05	Cell	327669	High throughput screening assay used to identify novel compounds that inhibit *Mycobacterium tuberculosis* in 7H9 media	Inh30
1259343	C06	Cell	6225	High throughput screening of small molecules that kill *Mycobacterium tuberculosis*	Inh30
1259417	C07	Cell	1105	High throughput whole cell screen to identify inhibitors of *Mycobacterium tuberculosis*	Inh30
1671161	C08	Cell	96022 / 86588	Phenotypic growth assay for *Mycobacterium tuberculosis* grown for 4 days on DPPC, cholesterol, tyloxapol based media	Inh30
1671162	C09	Cell	103984 / 86574	Phenotypic growth assay for *Mycobacterium tuberculosis* grown for 3 days on 7H9, glucose tyloxapol based media	Inh30
1671174	C10	Cell	53171 / 53165	Phenotypic assay to identify agents that inhibit growth of *Mycobacterium tuberculosis*	Inh30
488890	C11	Cell	324545	Elucidation of physiology of non-replicating, drug-tolerant *Mycobacterium tuberculosis*	Inh30

Cell-based assays

Total 557,527 compounds

96,040 compounds active in at least one assay

Mtb permeability datasets based on Big Data analysis

- Intersection of target-active and cell-tested compounds: 8242 compounds
- Compounds active in at least one cell-based assay are classified as penetrating (*MtbPen* = 1), otherwise as non-penetrating (*MtbPen* = 0)

Full dataset *MtbPen8242*
8242 compounds
2671 penetrating
5571 non-penetrating
Moderately imbalanced data

Balanced dataset *MtbPen5371ad*
5371 compounds
2671 penetrating
2700 diverse non-penetrating

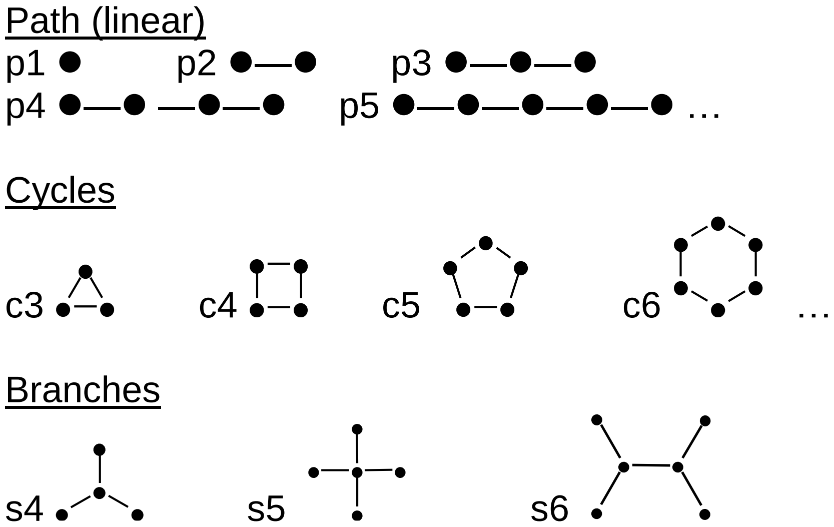
QSAR modeling: fragmental (substructural) molecular descriptors

- Occurrence counts (or presence) of fragments
- Thousands of fragments for real datasets
- “Holographic portrait” of a molecule
- Applicable to diverse series of compounds
- Easy prediction for new compounds
- Simple structural interpretation
- Mutual arrangement of structural features is handled indirectly via larger and/or overlapping fragments
- Acceptable for non-specific properties and/or diverse datasets

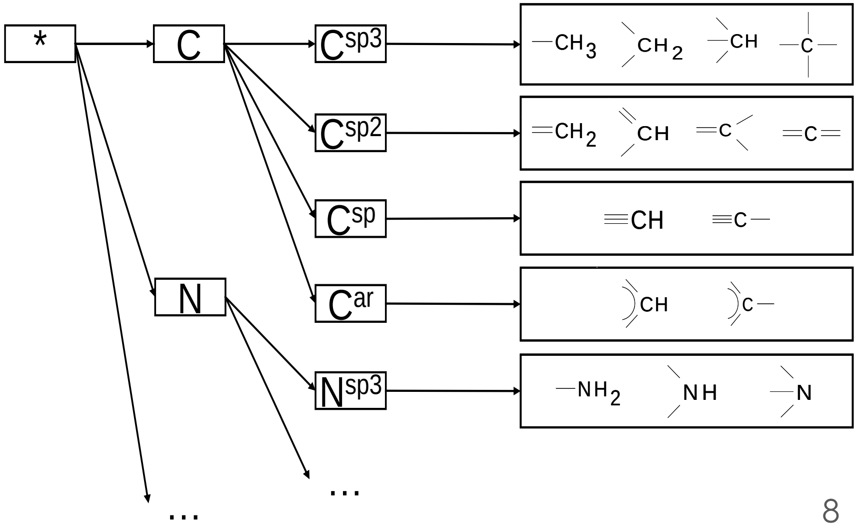
Up to 8 non-hydrogen atoms

Fragments present at least in 100 compounds

Basic subgraphs

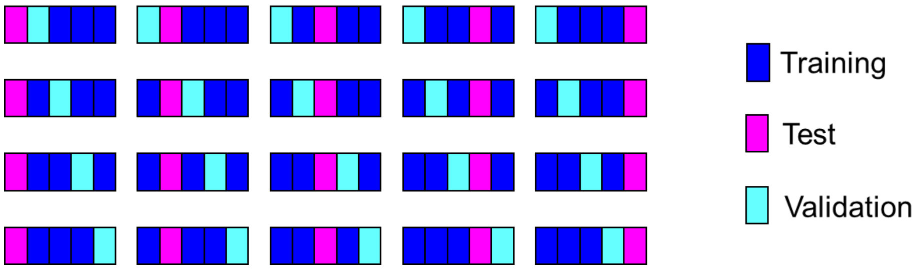


Hierarchical atom type classification



Machine learning modeling approach

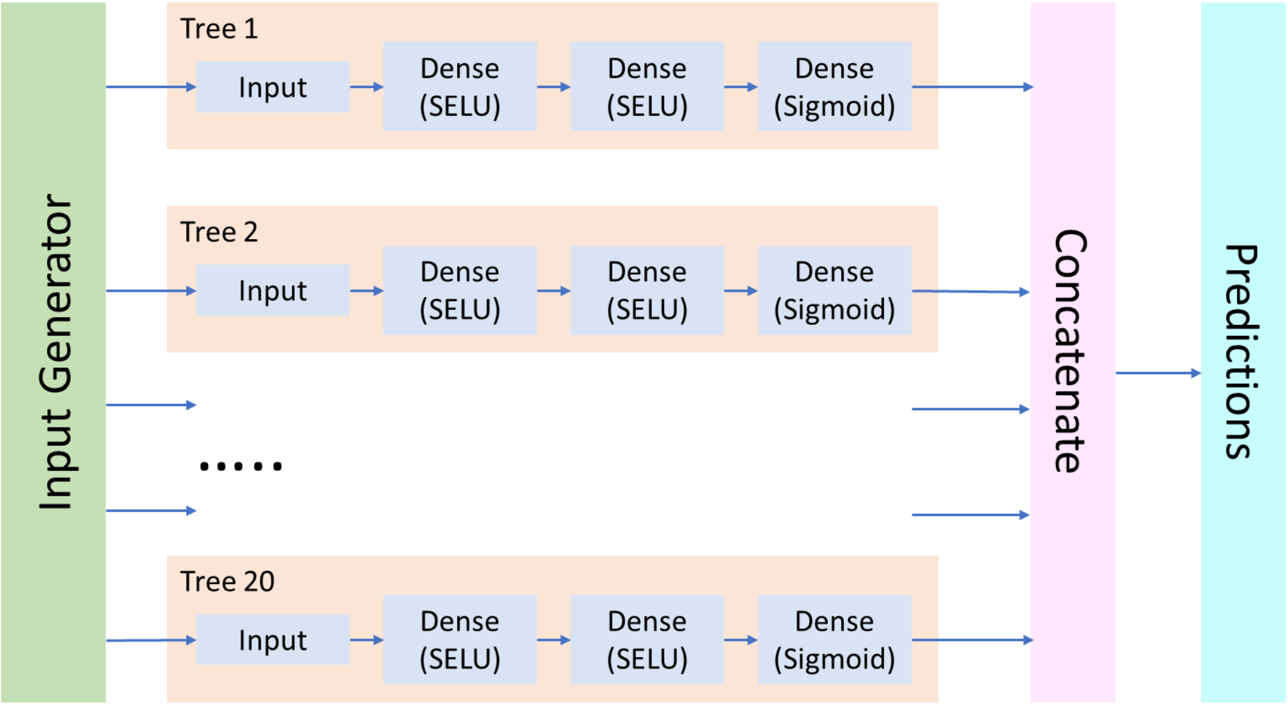
- Similar to ADMET modeling workflow
- Fragmental descriptors
- (Deep) feed-forward back-propagation neural network (BPNN)
- Repeated randomized double cross-validation (5x4 fold) to prevent overfitting and chance correlations
- Ensemble prediction



```
Perform endpoint scaling
Perform descriptor scaling
Perform descriptor selection
Repeat $N_R$ times
    Split dataset into $N_O$ subsets
    For each of $N_O$ subsets
        # Outer loop: use current subset for validation, other subsets for training
        Split outer loop training dataset into $N_I$ subsets
        For each of $N_I$ subsets
            # Inner loop: use current subset for termination, other subsets for training
            Build individual neural network model using other subsets for training and current subset for termination
            Evaluate model on the outer loop validation subset, collect statistics
            Save individual submodel
Consolidate validation errors, compute final statistics
Save complete ensemble model
```

Parallelized double cross-validation

- Neural network "forest" model
- TensorFlow 2.4.1/Keras 2.4.3
- High-performance NVIDIA RTX3080Ti GPU
- Hyperparameter optimization: fragment size, descriptor count, number and sizes of DNN layers, dropout



Predictive *Mtb* permeability models

Full dataset *MtbPen8242*

500 fragmental descriptors up to 6 atoms

2 hidden layers

$Acc_{cv} = 0.752$

$BalAcc_{cv} = 0.683$

$Sens_{cv} = 0.486$

$Spec_{cv} = 0.880$

Low recognition of penetrating compounds, likely due to imbalance in favor of non-penetrating

Balanced dataset *MtbPen5371ad*

900 fragmental descriptors up to 6 atoms

2 hidden layers

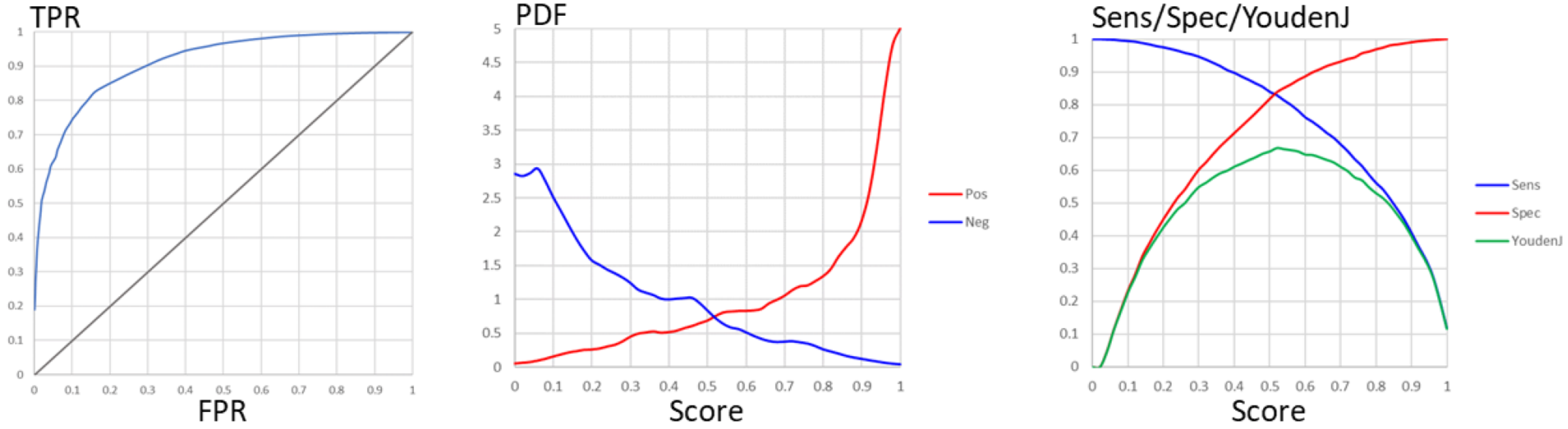
$Acc_{cv} = 0.768$

$BalAcc_{cv} = 0.768$

$Sens_{cv} = 0.768$

$Spec_{cv} = 0.769$

Model can be used to screen or design likely penetrating compounds



Radchenko E.V., Antonyan G.V., Ignatov S.K., Palyulin V.A. *Molecules*, **2023**, *28*, 633

Acknowledgments

Laboratory of chemoinformatics and molecular modeling
Lobachevsky State University of Nizhny Novgorod

Strategic Academic Leadership Program “Priority 2030”