ÚOCHB AV ČR
IOCB PRAGUE

RECEPTOR.AI

Univerzita Palackého
v Olomouci

Artificial intelligence in drug discovery

Semen Yesylevskyy

- Receptor.AI LTD
- IOCB Prague
- Palacký University Olomouc

Quote of the day

“An amount of intelligence in a typical drug discovery project is so low that the some artificial intelligence would not harm”

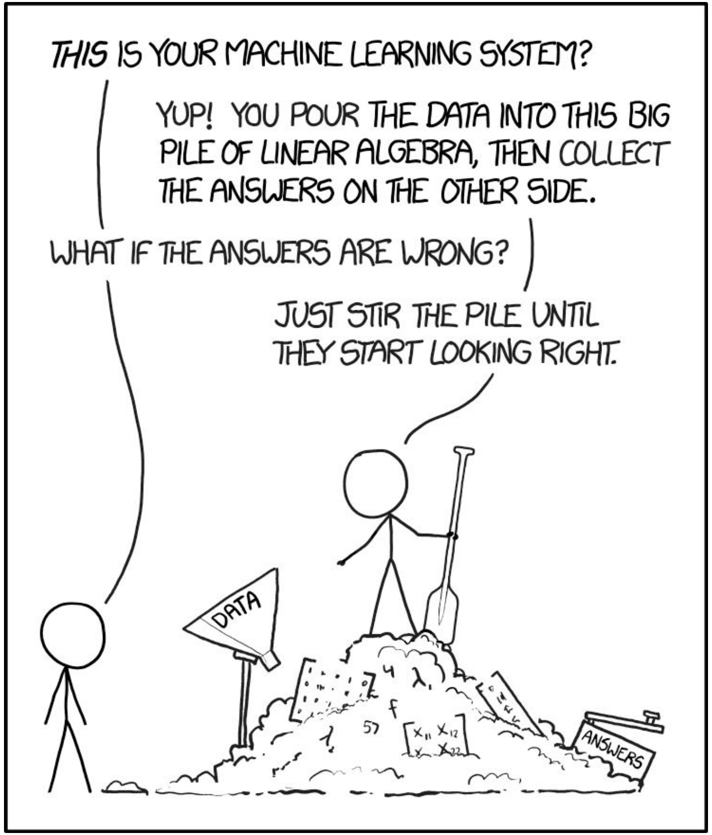
Founders of Receptor.AI in 2021 ©

Expectations:

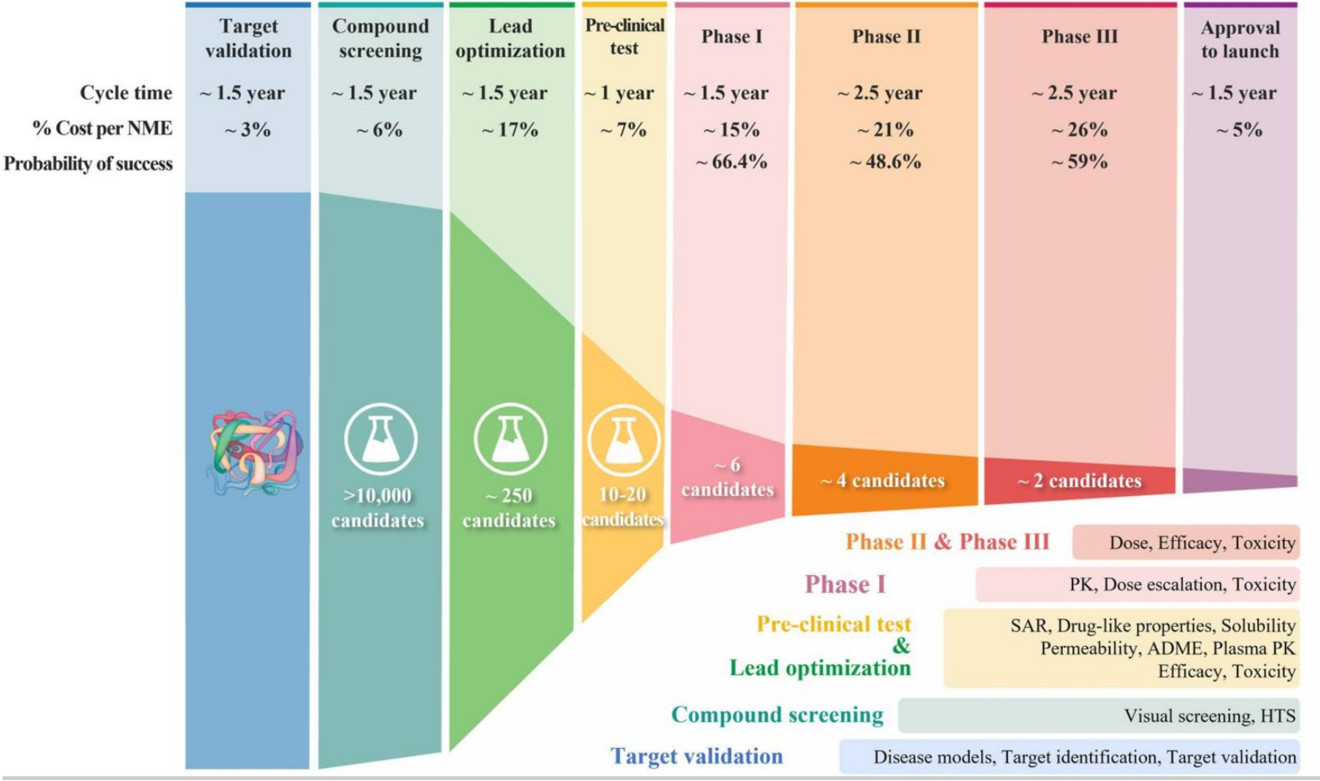
Reality:

Plan of the talk

1. Why modern drug discovery struggles
   - A crash course of upsetting the investors
2. Can AI make it struggle a bit less?
   - A short guide for giving hope to upset investors
3. Some shameless self-promotion
   - Investors don't trust this anyway


THIS IS YOUR MACHINE LEARNING SYSTEM?
YUP! YOU POUR THE DATA INTO THIS BIG PILE OF LINEAR ALGEBRA, THEN COLLECT THE ANSWERS ON THE OTHER SIDE.
WHAT IF THE ANSWERS ARE WRONG?
JUST STIR THE PILE UNTIL THEY START LOOKING RIGHT.
DATA
ANSWERS

Modern drug discovery struggles badly



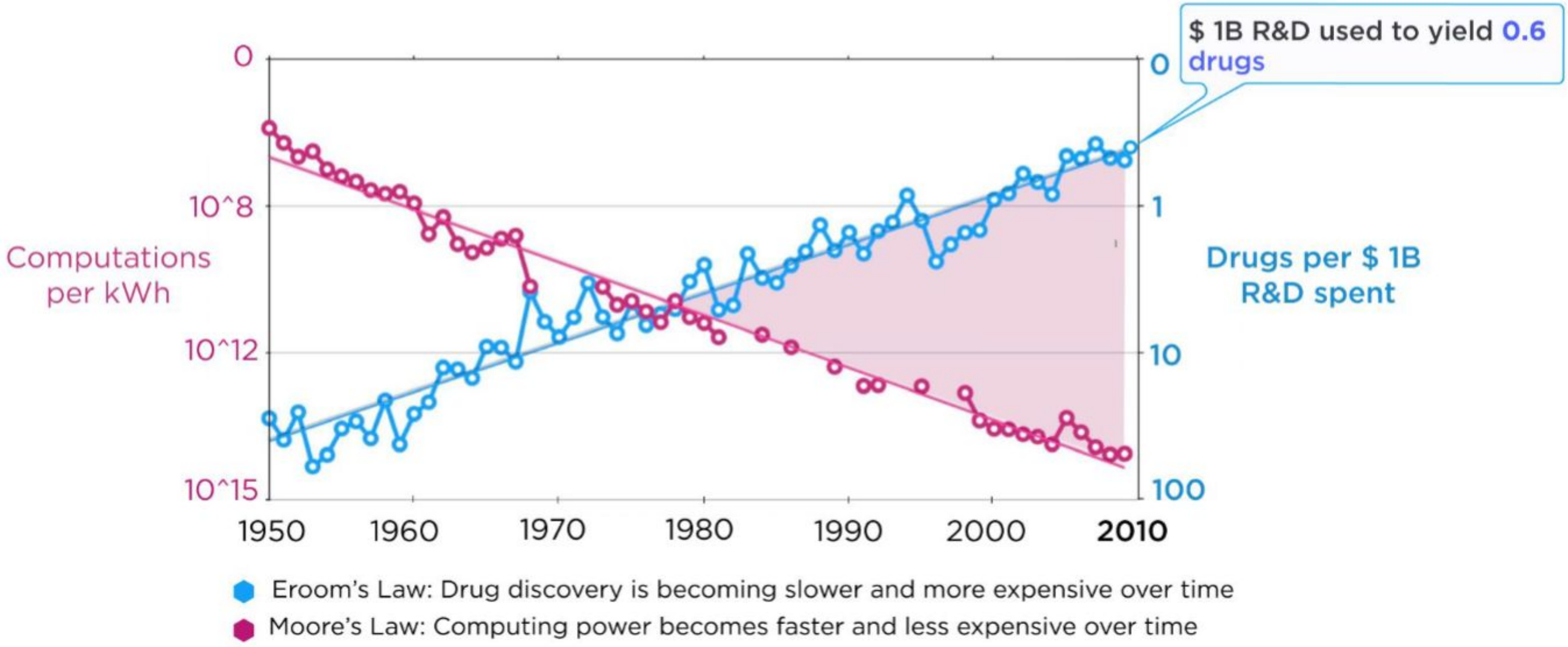
	Target validation	Compound screening	Lead optimization	Pre-clinical test	Phase I	Phase II	Phase III	Approval to launch
Cycle time	~ 1.5 year	~ 1.5 year	~ 1.5 year	~ 1 year	~ 1.5 year	~ 2.5 year	~ 2.5 year	~ 1.5 year
% Cost per NME	~ 3%	~ 6%	~ 17%	~ 7%	~ 15%	~ 21%	~ 26%	~ 5%
Probability of success					~ 66.4%	~ 48.6%	~ 59%	

Phase II & Phase III	Dose, Efficacy, Toxicity
Phase I	PK, Dose escalation, Toxicity
Pre-clinical test & Lead optimization	SAR, Drug-like properties, Solubility Permeability, ADME, Plasma PK Efficacy, Toxicity
Compound screening	Visual screening, HTS
Target validation	Disease models, Target identification, Target validation

Traditional methods stagnate

- The cost per drug increases
- Development time doesn’t improve
- Failure rate is persistently >90%
- Only **6.3%** composite success rate in 2022

Are we cursed? (Let's upset the investors…)



Computational resources become cheaper but this doesn't help much…

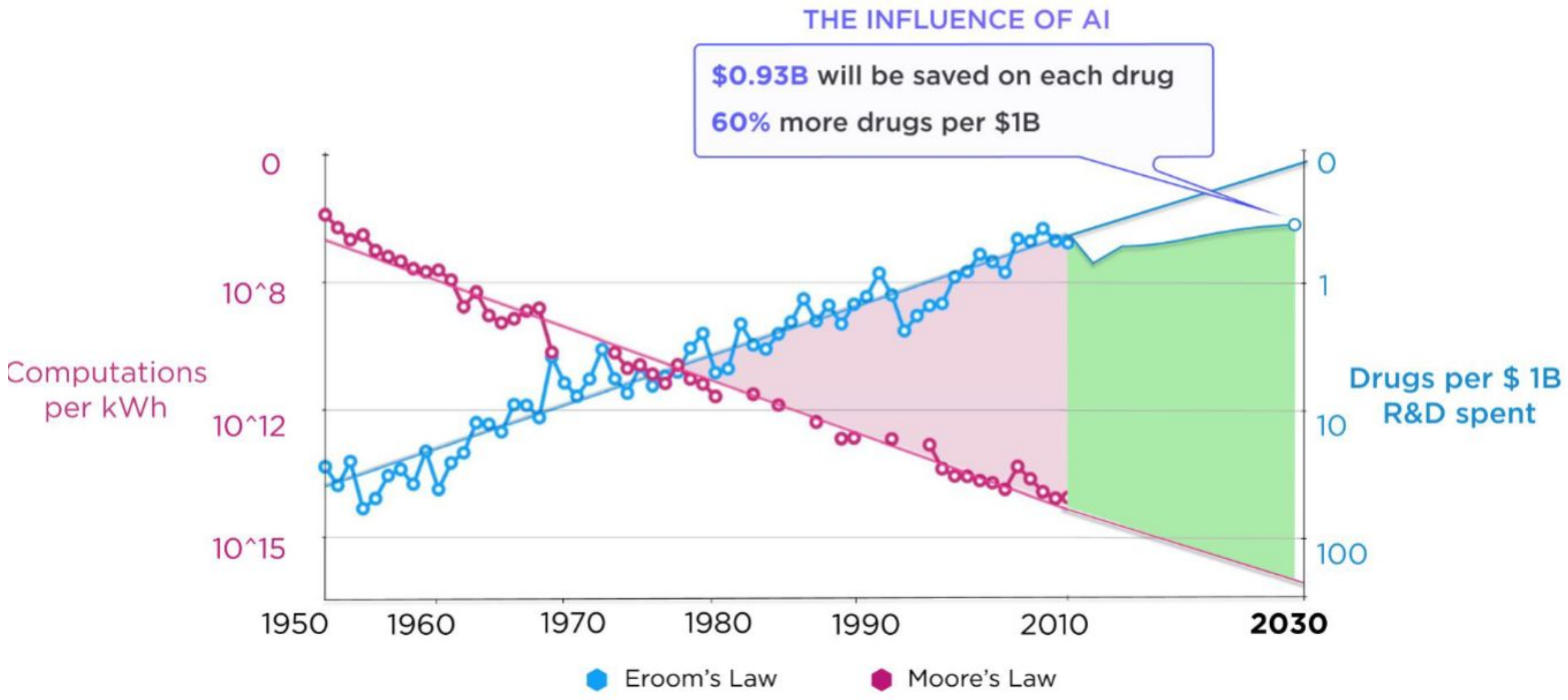
Eroom’s law explained (kind of)

- **The 'better than the Beatles' problem**: very hard to beat established treatments to the extent that it’s economically viable.
- **The 'cautious regulator' problem**: level of required evidence in trials become a burden.
- **The 'throw money at it' tendency**: The tendency to add excessive resources to R&D. One woman gives birth in 9 month. Let hire 9 women to give a birth in 1 month!
- **The 'basic research–brute force' bias**: The tendency to overestimate the ability of advances in basic research and brute force screening methods. Late stages continue to fail despite huge amounts of obtained data.

Cat AI beat the Eroom's law?

- **AI is generally considered as a rescue**
  - Breaking the Eroom's law
  - 60% more drugs per $1B by 2030
  - General paradigm change
- **The 'better than the Beatles' problem**:
  - Cutting the R&D cost to the extent that even moderate improvement will pay for itself.
  - Finding fundamentally different modalities and targets.
- **The 'cautious regulator' problem**:
  - Predicting the unfavourable clinical outcomes *very early* to cut futile projects.
  - Automate and streamline the trials.
- **The 'throw money at it' tendency**:
  - Better throw money at us :)
- **The 'basic research–brute force' bias**:
  - Making multi-domain predictive models including all available big data and hope that this will reduce the % of late stage failures

Can AI save us? (Let's give some hope to upset investors…)



Problems AI can solve

The problem of the context gaps:

Multiple knowledge domains don’t play together well

- Chemistry
- Biology
- Simulations
- Bioinformatics
- Population omics
- Patient data

Intractable amount of data:

- 50+B chemical spaces
- 40+ ADMET endpoints
- High-throughput readouts (HTS, DEL, RNA display, Phage display,...)
- Trials outcomes

Workflow construction:

- Which in silico methods to use?
- Which experiments to employ?
- Which cellular and animal models?
- What is the signal to stop?

Traditional approach: We need to develop drugs *quickly*, *reliably* and *cheaply*. Choose ***any two*** of these.

AI approach: Why not all at once?

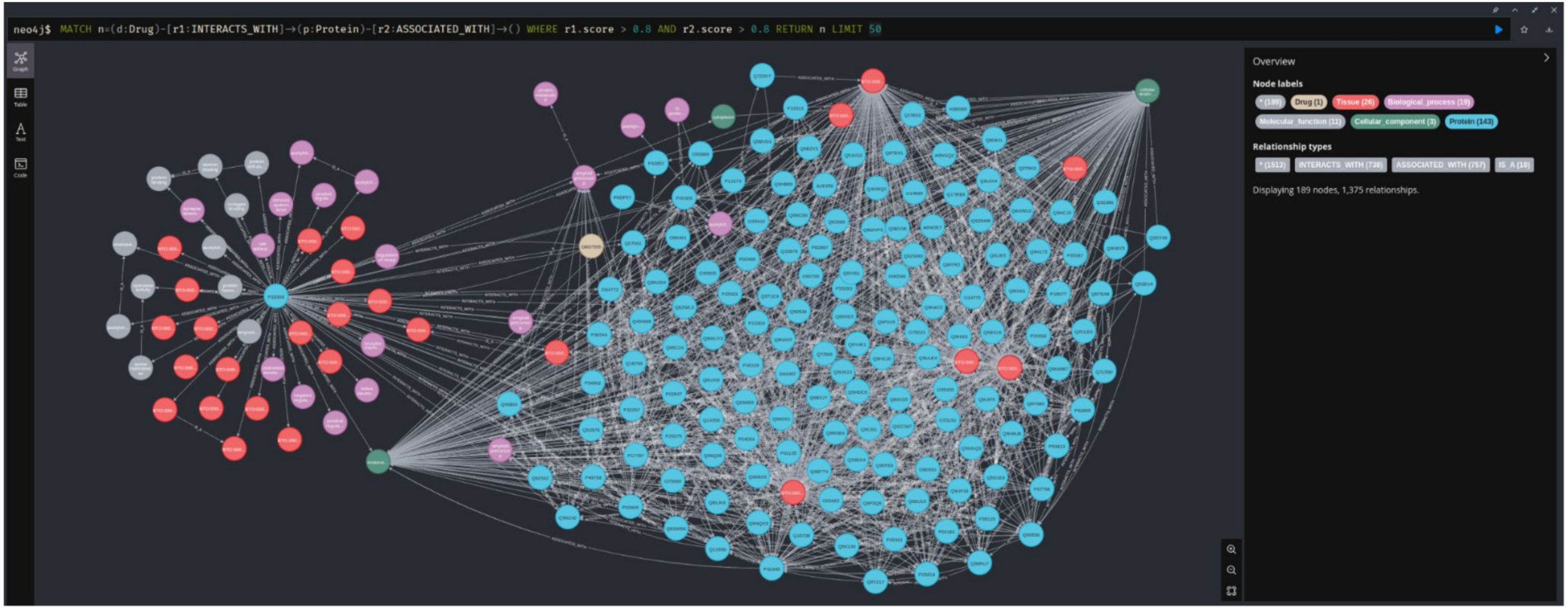
Applications of AI in drug discovery

- Target identification
  - Population omics
  - Knowledge graphs
  - Unstructured data scraping
- Early discovery
  - Hit discovery to lead optimization: AI virtual screening, ADMET prediction, QSAR.
- Late discovery
  - Formulation optimization,
  - IND and clinical studies outcome prediction
  - Clinical study planning and monitoring
- Drug repurposing
  - Off-target search

Applications of AI in drug discovery

- Target identification
  - Population omics
  - Knowledge graphs
  - Unstructured data scraping
- Early discovery
  - Hit discovery to lead optimization: AI virtual screening, ADMET prediction, QSAR.
- Late discovery
  - Formulation optimization,
  - IND and clinical studies outcome prediction
  - Clinical study planning and monitoring
- Drug repurposing
  - Off-target search

Target Identification: AI-curated knowledge graphs



```
neo4j$ MATCH n=(d:Drug)-[r1:INTERACTS_WITH]→(p:Protein)-[r2:ASSOCIATED_WITH]→() WHERE r1.score > 0.8 AND r2.score > 0.8 RETURN n LIMIT 50
```

Overview

Node labels

* (189) · Drug (1) · Tissue (26) · Biological_process (19) · Molecular_function (11) · Cellular_component (3) · Protein (143)

Relationship types

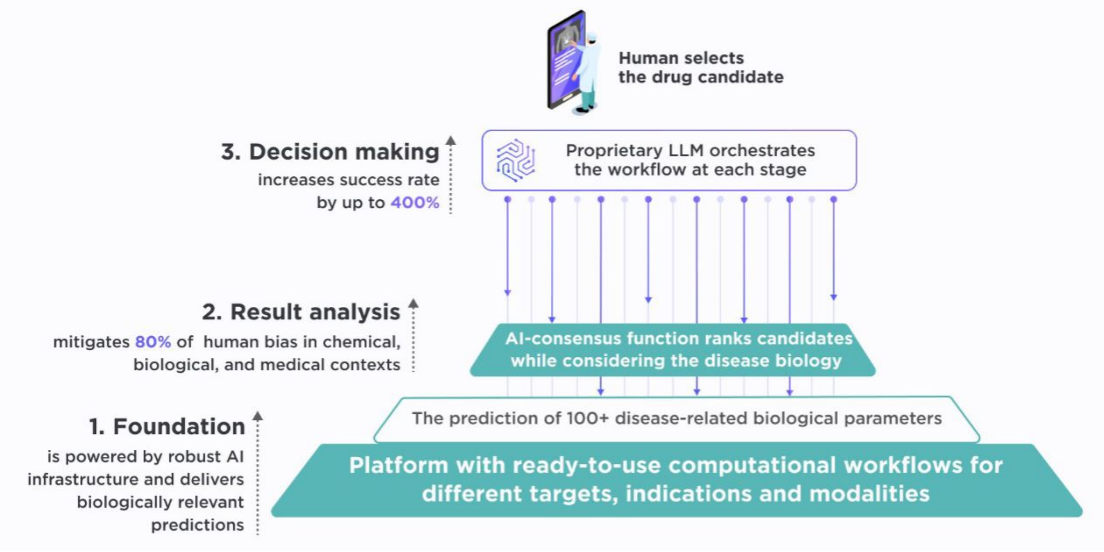
* (1513) · INTERACTS_WITH (738) · ASSOCIATED_WITH (757) · IS_A (18)

Displaying 189 nodes, 1,375 relationships.

- Multiparametric graph databases relating diseases, pathways, omics, proteins, drugs, modalities, indications, etc…
- Scraped automatically from all structured databases + LLM-based scraping of papers, patents, clinical study reports.
- Example questions to ask: *Find all protein targets associated with immuno oncology that has approved MABs but lack small molecules approved or on clinical trials 2+.*

AI-curated knowledge graphs

- Usage of AI:
  - Creation and continuous updating of the graph
  - Generation of queries and NLP transformation of responses
- Open questions:
  - Latest LLMs often provide similar performance *directly* in human language (they already contain most of information + can do the search)
  - Limited amount of public data →absence of competitive advantage.
  - Closed databases of big pharma are “new oil” for them.



Human selects the drug candidate

3. Decision making increases success rate by up to 400%

Proprietary LLM orchestrates the workflow at each stage

2. Result analysis mitigates 80% of human bias in chemical, biological, and medical contexts

AI-consensus function ranks candidates while considering the disease biology

1. Foundation is powered by robust AI infrastructure and delivers biologically relevant predictions

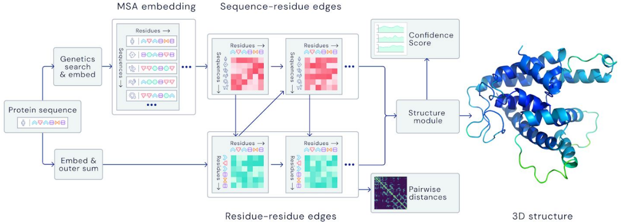
The prediction of 100+ disease-related biological parameters

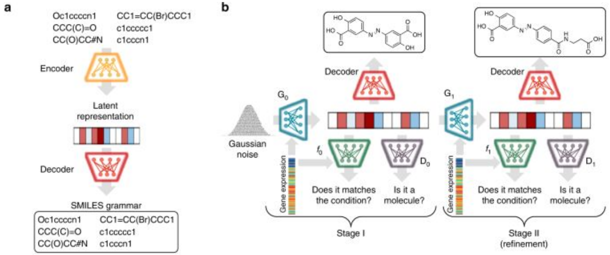
Platform with ready-to-use computational workflows for different targets, indications and modalities

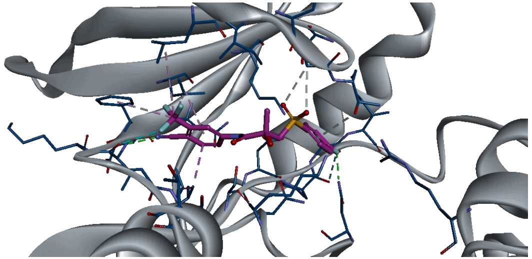
AI in early drug discovery

- Protein structure prediction
  - AlphaFold, Rosetta
- Chemical space generation
  - Molecular generators (Chemistry42, Iktos)
  - Scaffold hopping
  - Substituents generation
- Ligand pose prediction
  - DiffDock, UniMol, ArtiDock

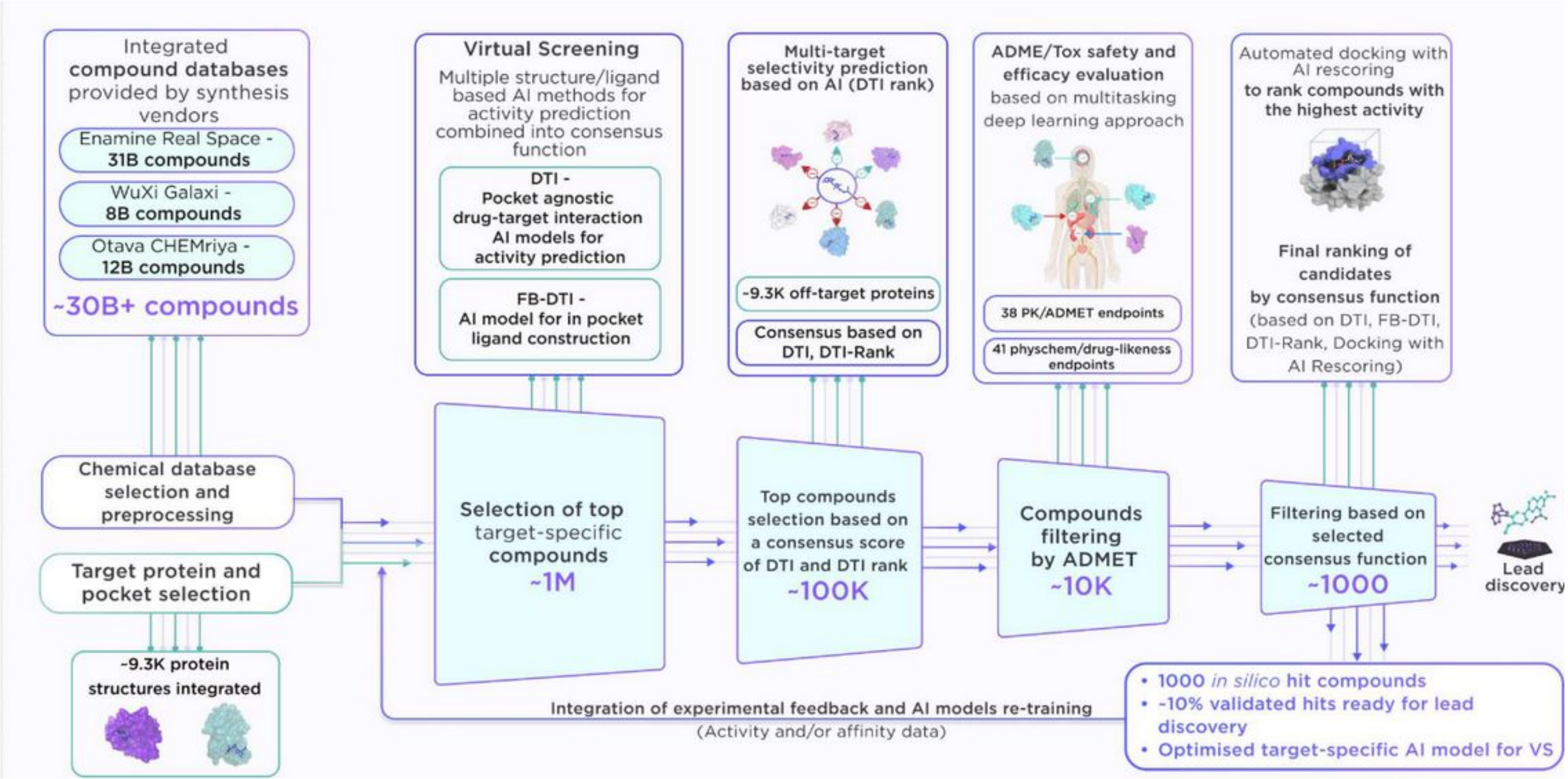
- Non-AI generative techniques
  - MD for protein conformational ensembles generation
  - Artificial binding pockets for AI data augmentation







AI virtual screening



Integrated **compound databases** provided by synthesis vendors

- Enamine Real Space - **31B compounds**
- WuXi Galaxi - **8B compounds**
- Otava CHEMriya - **12B compounds**

~30B+ compounds

Virtual Screening

Multiple structure/ligand based AI methods for activity prediction combined into consensus function

DTI - Pocket agnostic drug-target interaction AI models for activity prediction

FB-DTI - AI model for in pocket ligand construction

Multi-target selectivity prediction based on AI (DTI rank)

~9.3K off-target proteins

Consensus based on DTI, DTI-Rank

ADME/Tox safety and efficacy evaluation based on multitasking deep learning approach

38 PK/ADMET endpoints

41 physchem/drug-likeness endpoints

Automated docking with AI rescoring **to rank compounds with the highest activity**

Final ranking of candidates by consensus function (based on DTI, FB-DTI, DTI-Rank, Docking with AI Rescoring)

Chemical database selection and preprocessing

Target protein and pocket selection

~9.3K protein structures integrated

Selection of top target-specific **compounds**

~1M

Top compounds selection based on a consensus score of DTI and DTI rank

~100K

Compounds filtering by ADMET

~10K

Filtering based on selected consensus function

~1000

Lead discovery

Integration of experimental feedback and AI models re-training

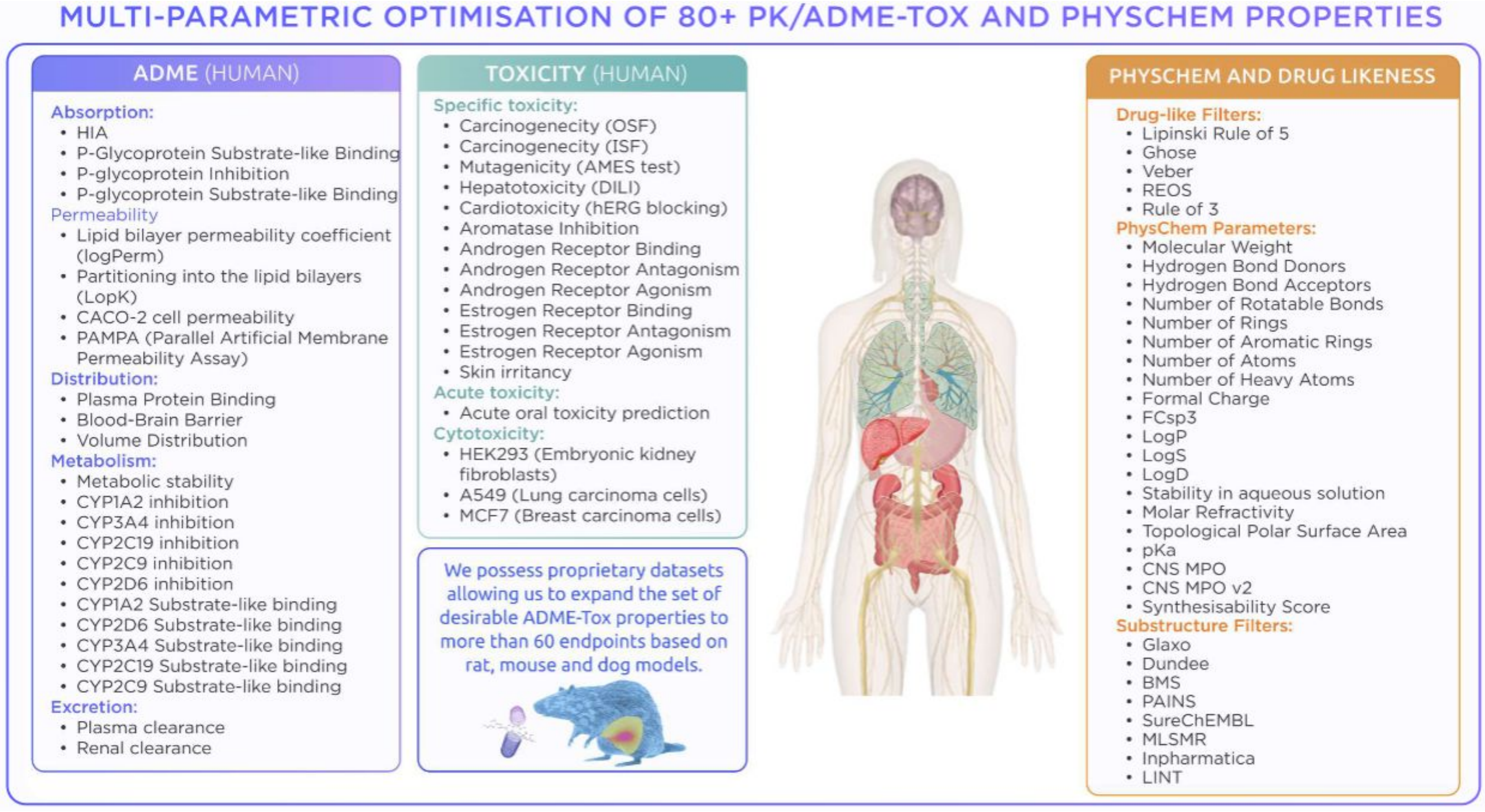
(Activity and/or affinity data)

- **1000 *in silico* hit compounds**
- **~10% validated hits ready for lead discovery**
- **Optimised target-specific AI model for VS**

AI virtual screening

- Very fast (2-3 order of magnitude faster) initial filtration of the chemical space
- Self-balancing: many known compounds →ligand-based approach; few compounds →structure based approach.
- Separate models for protein tier lists (depending on the number of known structures and ligands).
- 70+% accuracy on “favourable” targets.
- Early assessment of ADMET →fewer toxicity failures

ADMET prediction



MULTI-PARAMETRIC OPTIMISATION OF 80+ PK/ADME-TOX AND PHYSCHEM PROPERTIES

ADME (HUMAN)

Absorption:
- HIA
- P-Glycoprotein Substrate-like Binding
- P-glycoprotein Inhibition
- P-glycoprotein Substrate-like Binding

Permeability
- Lipid bilayer permeability coefficient (logPerm)
- Partitioning into the lipid bilayers (LopK)
- CACO-2 cell permeability
- PAMPA (Parallel Artificial Membrane Permeability Assay)

Distribution:
- Plasma Protein Binding
- Blood-Brain Barrier
- Volume Distribution

Metabolism:
- Metabolic stability
- CYP1A2 inhibition
- CYP3A4 inhibition
- CYP2C19 inhibition
- CYP2C9 inhibition
- CYP2D6 inhibition
- CYP1A2 Substrate-like binding
- CYP2D6 Substrate-like binding
- CYP3A4 Substrate-like binding
- CYP2C19 Substrate-like binding
- CYP2C9 Substrate-like binding

Excretion:
- Plasma clearance
- Renal clearance

TOXICITY (HUMAN)

Specific toxicity:
- Carcinogenecity (OSF)
- Carcinogenecity (ISF)
- Mutagenicity (AMES test)
- Hepatotoxicity (DILI)
- Cardiotoxicity (hERG blocking)
- Aromatase Inhibition
- Androgen Receptor Binding
- Androgen Receptor Antagonism
- Androgen Receptor Agonism
- Estrogen Receptor Binding
- Estrogen Receptor Antagonism
- Estrogen Receptor Agonism
- Skin irritancy

Acute toxicity:
- Acute oral toxicity prediction

Cytotoxicity:
- HEK293 (Embryonic kidney fibroblasts)
- A549 (Lung carcinoma cells)
- MCF7 (Breast carcinoma cells)

We possess proprietary datasets allowing us to expand the set of desirable ADME-Tox properties to more than 60 endpoints based on rat, mouse and dog models.

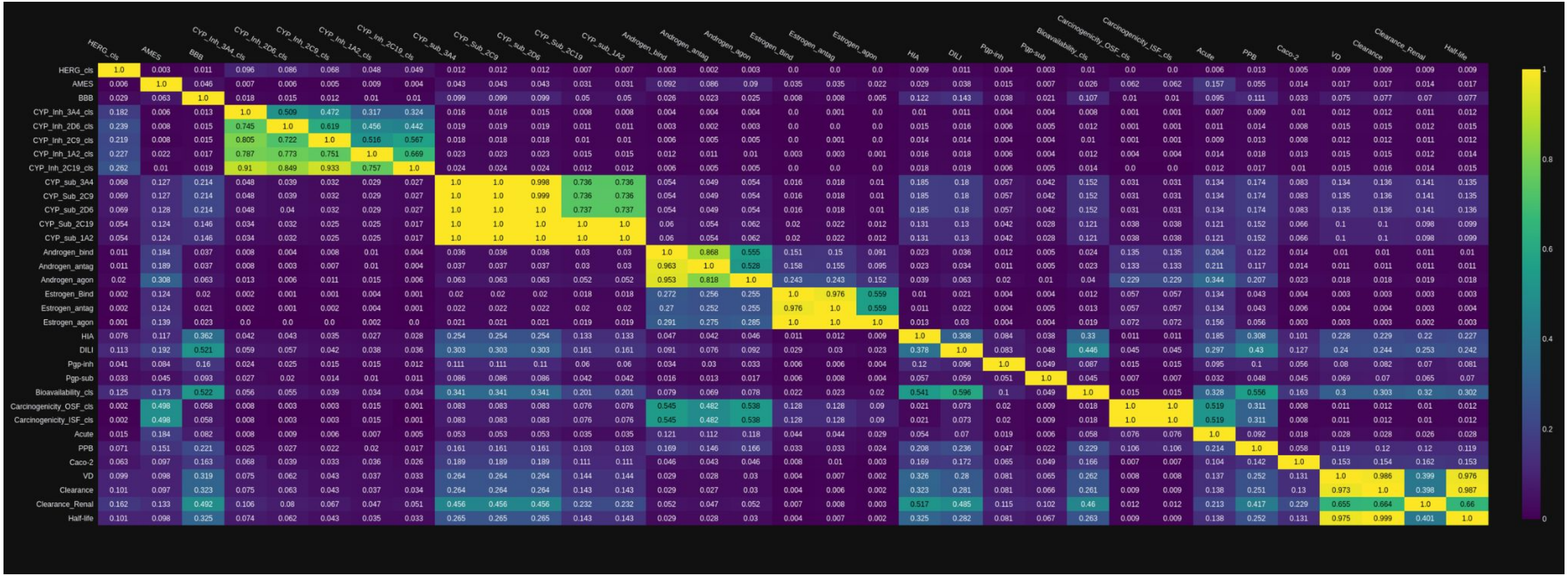
PHYSCHEM AND DRUG LIKENESS

Drug-like Filters:
- Lipinski Rule of 5
- Ghose
- Veber
- REOS
- Rule of 3

PhysChem Parameters:
- Molecular Weight
- Hydrogen Bond Donors
- Hydrogen Bond Acceptors
- Number of Rotatable Bonds
- Number of Rings
- Number of Aromatic Rings
- Number of Atoms
- Number of Heavy Atoms
- Formal Charge
- FCsp3
- LogP
- LogS
- LogD
- Stability in aqueous solution
- Molar Refractivity
- Topological Polar Surface Area
- pKa
- CNS MPO
- CNS MPO v2
- Synthesisability Score

Substructure Filters:
- Glaxo
- Dundee
- BMS
- PAINS
- SureChEMBL
- MLSMR
- Inpharmatica
- LINT

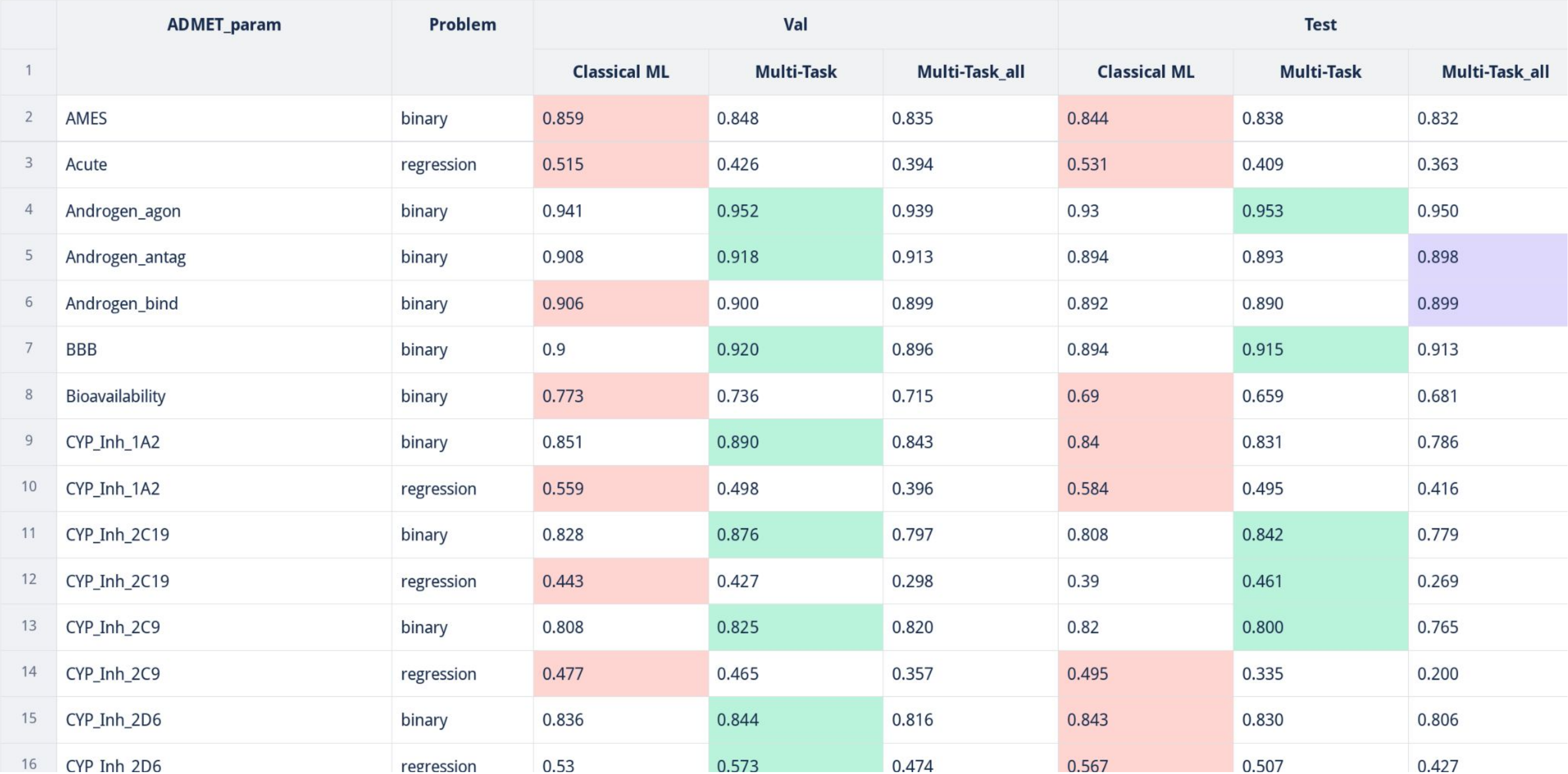
ADMET multi-task learning



	HERG_cls	AMES	BBB	CYP_Inh_3A4_cls	CYP_Inh_2D6_cls	CYP_Inh_2C9_cls	CYP_Inh_1A2_cls	CYP_Inh_2C19_cls	CYP_sub_3A4	CYP_Sub_2C9	CYP_sub_2D6	CYP_Sub_2C19	CYP_sub_1A2	Androgen_bind	Androgen_antag	Androgen_agon	Estrogen_Bind	Estrogen_antag	Estrogen_agon	HIA	DILI	Pgp-inh	Pgp-sub	Bioavailability_cls	Carcinogenicity_OSF_cls	Carcinogenicity_ISF_cls	Acute	PPB	Caco-2	VD	Clearance	Clearance_Renal	Half-life
HERG_cls	1.0	0.003	0.011	0.096	0.086	0.068	0.048	0.049	0.012	0.012	0.012	0.007	0.007	0.003	0.002	0.003	0.0	0.0	0.0	0.009	0.011	0.004	0.003	0.01	0.0	0.0	0.006	0.013	0.005	0.009	0.009	0.009	0.009
AMES	0.006	1.0	0.046	0.007	0.006	0.005	0.009	0.004	0.043	0.043	0.043	0.031	0.031	0.092	0.086	0.09	0.035	0.035	0.022	0.029	0.038	0.015	0.007	0.026	0.062	0.062	0.157	0.055	0.014	0.017	0.017	0.014	0.017
BBB	0.029	0.063	1.0	0.018	0.015	0.012	0.01	0.01	0.099	0.099	0.099	0.05	0.05	0.026	0.023	0.025	0.008	0.008	0.005	0.122	0.143	0.038	0.021	0.107	0.01	0.01	0.095	0.111	0.033	0.075	0.077	0.07	0.077
CYP_Inh_3A4_cls	0.182	0.006	0.013	1.0	0.509	0.472	0.317	0.324	0.016	0.016	0.015	0.008	0.008	0.004	0.004	0.004	0.0	0.001	0.0	0.01	0.011	0.004	0.004	0.008	0.001	0.001	0.007	0.009	0.01	0.012	0.012	0.011	0.012
CYP_Inh_2D6_cls	0.239	0.008	0.015	0.745	1.0	0.619	0.456	0.442	0.019	0.019	0.019	0.011	0.011	0.003	0.002	0.003	0.0	0.0	0.0	0.015	0.016	0.006	0.005	0.012	0.001	0.001	0.011	0.014	0.008	0.015	0.015	0.012	0.015
CYP_Inh_2C9_cls	0.219	0.008	0.015	0.805	0.722	1.0	0.516	0.567	0.018	0.018	0.018	0.01	0.01	0.006	0.005	0.005	0.0	0.001	0.0	0.014	0.014	0.004	0.004	0.01	0.001	0.001	0.009	0.013	0.008	0.012	0.012	0.011	0.012
CYP_Inh_1A2_cls	0.227	0.022	0.017	0.787	0.773	0.751	1.0	0.669	0.023	0.023	0.023	0.015	0.015	0.012	0.011	0.01	0.003	0.003	0.001	0.016	0.018	0.006	0.004	0.012	0.004	0.004	0.014	0.018	0.013	0.015	0.015	0.012	0.014
CYP_Inh_2C19_cls	0.262	0.01	0.019	0.91	0.849	0.933	0.757	1.0	0.024	0.024	0.024	0.012	0.012	0.006	0.005	0.005	0.0	0.0	0.0	0.018	0.019	0.006	0.005	0.014	0.0	0.0	0.012	0.017	0.01	0.015	0.016	0.014	0.015
CYP_sub_3A4	0.068	0.127	0.214	0.048	0.039	0.032	0.029	0.027	1.0	1.0	0.998	0.736	0.736	0.054	0.049	0.054	0.016	0.018	0.01	0.185	0.18	0.057	0.042	0.152	0.031	0.031	0.134	0.174	0.083	0.134	0.136	0.141	0.135
CYP_Sub_2C9	0.069	0.127	0.214	0.048	0.039	0.032	0.029	0.027	1.0	1.0	0.999	0.736	0.736	0.054	0.049	0.054	0.016	0.018	0.01	0.185	0.18	0.057	0.042	0.152	0.031	0.031	0.134	0.174	0.083	0.135	0.136	0.141	0.135
CYP_sub_2D6	0.069	0.128	0.214	0.048	0.04	0.032	0.029	0.027	1.0	1.0	1.0	0.737	0.737	0.054	0.049	0.054	0.016	0.018	0.01	0.185	0.18	0.057	0.042	0.152	0.031	0.031	0.134	0.174	0.083	0.135	0.136	0.141	0.136
CYP_Sub_2C19	0.054	0.124	0.146	0.034	0.032	0.025	0.025	0.017	1.0	1.0	1.0	1.0	1.0	0.06	0.054	0.062	0.02	0.022	0.012	0.131	0.13	0.042	0.028	0.121	0.038	0.038	0.121	0.152	0.066	0.1	0.1	0.098	0.099
CYP_sub_1A2	0.054	0.124	0.146	0.034	0.032	0.025	0.025	0.017	1.0	1.0	1.0	1.0	1.0	0.06	0.054	0.062	0.02	0.022	0.012	0.131	0.13	0.042	0.028	0.121	0.038	0.038	0.121	0.152	0.066	0.1	0.1	0.098	0.099
Androgen_bind	0.011	0.184	0.037	0.008	0.004	0.008	0.01	0.004	0.036	0.036	0.036	0.03	0.03	1.0	0.868	0.555	0.151	0.15	0.091	0.023	0.036	0.012	0.005	0.024	0.135	0.135	0.204	0.122	0.014	0.01	0.01	0.011	0.01
Androgen_antag	0.011	0.189	0.037	0.008	0.003	0.007	0.01	0.004	0.037	0.037	0.037	0.03	0.03	0.963	1.0	0.528	0.158	0.155	0.095	0.023	0.034	0.011	0.005	0.023	0.133	0.133	0.211	0.117	0.014	0.011	0.011	0.011	0.011
Androgen_agon	0.02	0.308	0.063	0.013	0.006	0.011	0.015	0.006	0.063	0.063	0.063	0.052	0.052	0.953	0.818	1.0	0.243	0.243	0.152	0.039	0.063	0.02	0.01	0.04	0.229	0.229	0.344	0.207	0.023	0.018	0.018	0.019	0.018
Estrogen_Bind	0.002	0.124	0.02	0.002	0.001	0.001	0.004	0.001	0.02	0.02	0.02	0.018	0.018	0.272	0.256	0.255	1.0	0.976	0.559	0.01	0.021	0.004	0.004	0.012	0.057	0.057	0.134	0.043	0.004	0.003	0.003	0.003	0.003
Estrogen_antag	0.002	0.124	0.021	0.002	0.001	0.002	0.004	0.001	0.022	0.022	0.022	0.02	0.02	0.27	0.252	0.255	0.976	1.0	0.559	0.011	0.022	0.004	0.005	0.013	0.057	0.057	0.134	0.043	0.006	0.004	0.004	0.003	0.004
Estrogen_agon	0.001	0.139	0.023	0.0	0.0	0.0	0.002	0.0	0.021	0.021	0.021	0.019	0.019	0.291	0.275	0.285	1.0	1.0	1.0	0.013	0.03	0.004	0.004	0.019	0.072	0.072	0.156	0.056	0.003	0.003	0.003	0.002	0.003
HIA	0.076	0.117	0.362	0.042	0.043	0.035	0.027	0.028	0.254	0.254	0.254	0.133	0.133	0.047	0.042	0.046	0.011	0.012	0.009	1.0	0.308	0.084	0.038	0.33	0.011	0.011	0.185	0.308	0.101	0.228	0.229	0.22	0.227
DILI	0.113	0.192	0.521	0.059	0.057	0.042	0.038	0.036	0.303	0.303	0.303	0.161	0.161	0.091	0.076	0.092	0.029	0.03	0.023	0.378	1.0	0.083	0.048	0.446	0.045	0.045	0.297	0.43	0.127	0.24	0.244	0.253	0.242
Pgp-inh	0.041	0.084	0.16	0.024	0.025	0.015	0.015	0.012	0.111	0.111	0.11	0.06	0.06	0.034	0.03	0.033	0.006	0.006	0.004	0.12	0.096	1.0	0.049	0.087	0.015	0.015	0.095	0.1	0.056	0.08	0.082	0.07	0.081
Pgp-sub	0.033	0.045	0.093	0.027	0.02	0.014	0.01	0.011	0.086	0.086	0.086	0.042	0.042	0.016	0.013	0.017	0.006	0.008	0.004	0.057	0.059	0.051	1.0	0.045	0.007	0.007	0.032	0.048	0.045	0.069	0.07	0.065	0.07
Bioavailability_cls	0.125	0.173	0.522	0.056	0.055	0.039	0.034	0.034	0.341	0.341	0.341	0.201	0.201	0.079	0.069	0.078	0.022	0.023	0.02	0.541	0.596	0.1	0.049	1.0	0.015	0.015	0.328	0.556	0.163	0.3	0.303	0.32	0.302
Carcinogenicity_OSF_cls	0.002	0.498	0.058	0.008	0.003	0.003	0.015	0.001	0.083	0.083	0.083	0.076	0.076	0.545	0.482	0.538	0.128	0.128	0.09	0.021	0.073	0.02	0.009	0.018	1.0	1.0	0.519	0.311	0.008	0.011	0.012	0.01	0.012
Carcinogenicity_ISF_cls	0.002	0.498	0.058	0.008	0.003	0.003	0.015	0.001	0.083	0.083	0.083	0.076	0.076	0.545	0.482	0.538	0.128	0.128	0.09	0.021	0.073	0.02	0.009	0.018	1.0	1.0	0.519	0.311	0.008	0.011	0.012	0.01	0.012
Acute	0.015	0.184	0.082	0.008	0.009	0.006	0.007	0.005	0.053	0.053	0.053	0.035	0.035	0.121	0.112	0.118	0.044	0.044	0.029	0.054	0.07	0.019	0.006	0.058	0.076	0.076	1.0	0.092	0.018	0.028	0.028	0.026	0.028
PPB	0.071	0.151	0.221	0.025	0.027	0.022	0.02	0.017	0.161	0.161	0.161	0.103	0.103	0.169	0.146	0.166	0.033	0.033	0.024	0.208	0.236	0.047	0.022	0.229	0.106	0.106	0.214	1.0	0.058	0.119	0.12	0.12	0.119
Caco-2	0.063	0.097	0.163	0.068	0.039	0.033	0.036	0.026	0.189	0.189	0.189	0.111	0.111	0.046	0.043	0.046	0.008	0.01	0.003	0.169	0.172	0.065	0.049	0.166	0.007	0.007	0.104	0.142	1.0	0.153	0.154	0.162	0.153
VD	0.099	0.098	0.319	0.075	0.062	0.043	0.037	0.033	0.264	0.264	0.264	0.144	0.144	0.029	0.028	0.03	0.004	0.007	0.002	0.326	0.28	0.081	0.065	0.262	0.008	0.008	0.137	0.252	0.131	1.0	0.986	0.399	0.976
Clearance	0.101	0.097	0.323	0.075	0.063	0.043	0.037	0.034	0.264	0.264	0.264	0.143	0.143	0.029	0.027	0.03	0.004	0.006	0.002	0.323	0.281	0.081	0.066	0.261	0.009	0.009	0.138	0.251	0.13	0.973	1.0	0.398	0.987
Clearance_Renal	0.162	0.133	0.492	0.106	0.08	0.067	0.047	0.051	0.456	0.456	0.456	0.232	0.232	0.052	0.047	0.052	0.007	0.008	0.003	0.517	0.485	0.115	0.102	0.46	0.012	0.012	0.213	0.417	0.229	0.655	0.664	1.0	0.66
Half-life	0.101	0.098	0.325	0.074	0.062	0.043	0.035	0.033	0.265	0.265	0.265	0.143	0.143	0.029	0.028	0.03	0.004	0.007	0.002	0.325	0.282	0.081	0.067	0.263	0.009	0.009	0.138	0.252	0.131	0.975	0.999	0.401	1.0

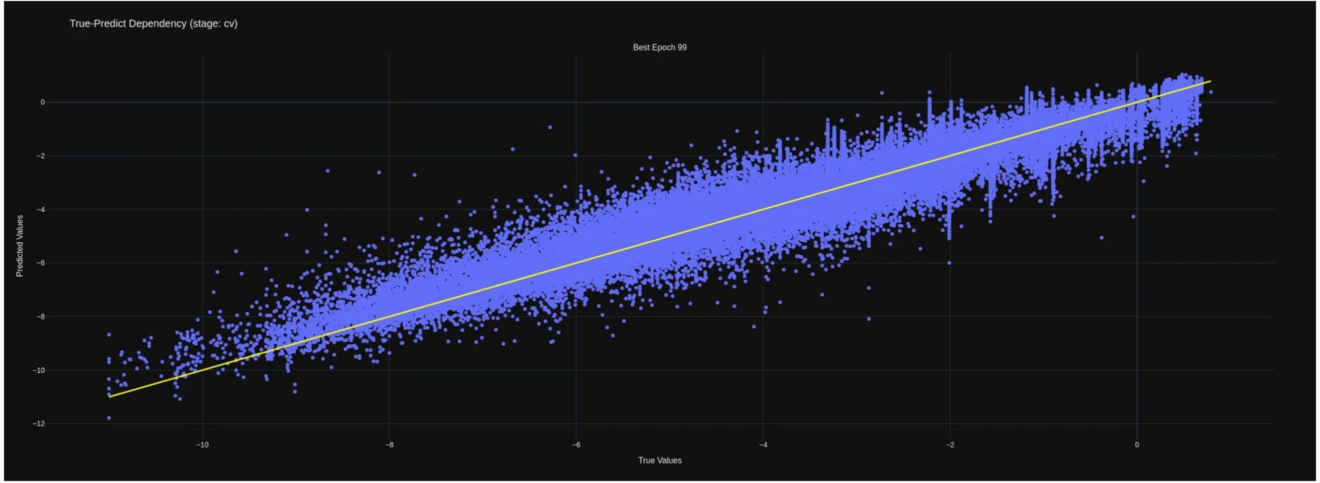
- Multi-task ADMET model: trained on multiple endpoints with “cross-dissemination” between them.
- There are groups of tasks sharing the data to more or less extent

MultiTask model training

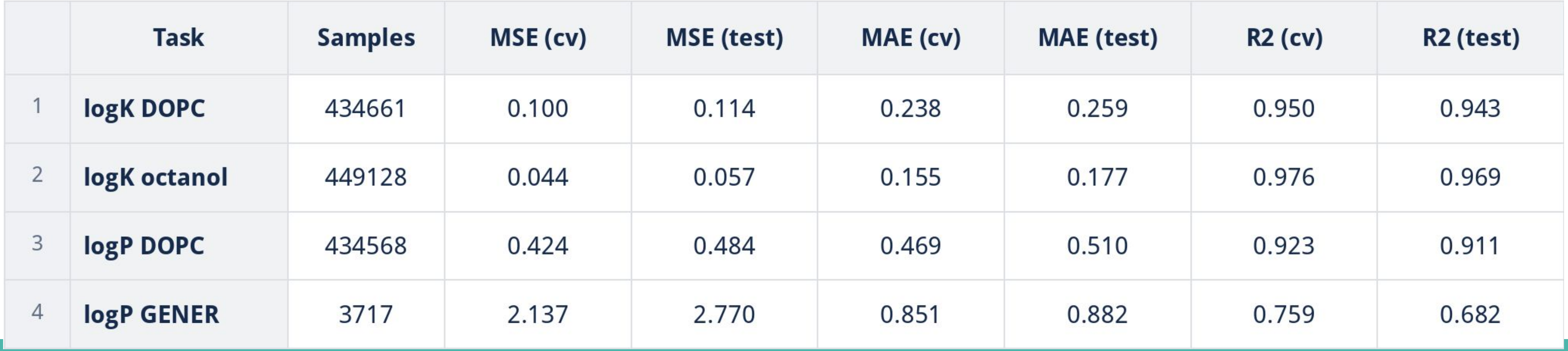


	ADMET_param	Problem	Val			Test		
1			Classical ML	Multi-Task	Multi-Task_all	Classical ML	Multi-Task	Multi-Task_all
2	AMES	binary	0.859	0.848	0.835	0.844	0.838	0.832
3	Acute	regression	0.515	0.426	0.394	0.531	0.409	0.363
4	Androgen_agon	binary	0.941	0.952	0.939	0.93	0.953	0.950
5	Androgen_antag	binary	0.908	0.918	0.913	0.894	0.893	0.898
6	Androgen_bind	binary	0.906	0.900	0.899	0.892	0.890	0.899
7	BBB	binary	0.9	0.920	0.896	0.894	0.915	0.913
8	Bioavailability	binary	0.773	0.736	0.715	0.69	0.659	0.681
9	CYP_Inh_1A2	binary	0.851	0.890	0.843	0.84	0.831	0.786
10	CYP_Inh_1A2	regression	0.559	0.498	0.396	0.584	0.495	0.416
11	CYP_Inh_2C19	binary	0.828	0.876	0.797	0.808	0.842	0.779
12	CYP_Inh_2C19	regression	0.443	0.427	0.298	0.39	0.461	0.269
13	CYP_Inh_2C9	binary	0.808	0.825	0.820	0.82	0.800	0.765
14	CYP_Inh_2C9	regression	0.477	0.465	0.357	0.495	0.335	0.200
15	CYP_Inh_2D6	binary	0.836	0.844	0.816	0.843	0.830	0.806
16	CYP_Inh_2D6	regression	0.53	0.573	0.474	0.567	0.507	0.427

Case study: membrane permeability

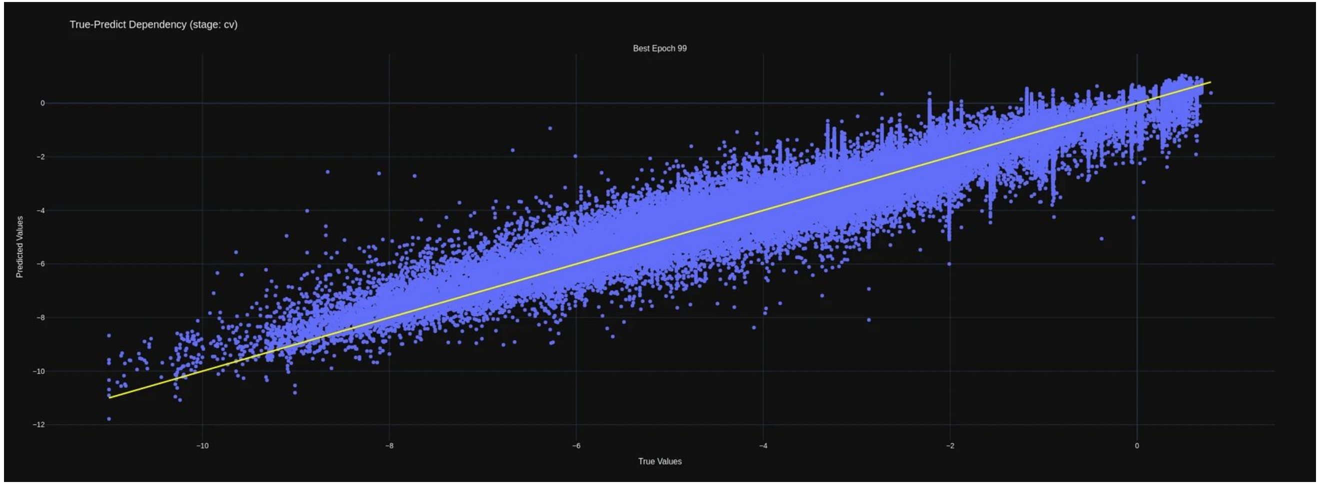


- MolMeDb data for
  - Membrane permeability
  - Membrane partitioning
- Receptor.AI MultiTask ADMET NN architecture
- AutoML automatic featurization

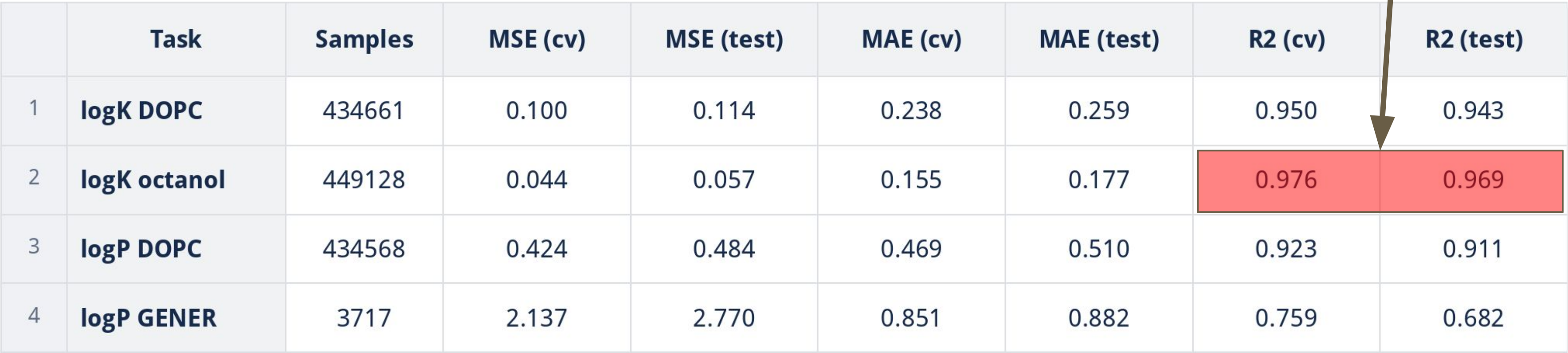


	Task	Samples	MSE (cv)	MSE (test)	MAE (cv)	MAE (test)	R2 (cv)	R2 (test)
1	**logK DOPC**	434661	0.100	0.114	0.238	0.259	0.950	0.943
2	**logK octanol**	449128	0.044	0.057	0.155	0.177	0.976	0.969
3	**logP DOPC**	434568	0.424	0.484	0.469	0.510	0.923	0.911
4	**logP GENER**	3717	2.137	2.770	0.851	0.882	0.759	0.682

Case study: membrane permeability



This is too good to be true…



	Task	Samples	MSE (cv)	MSE (test)	MAE (cv)	MAE (test)	R2 (cv)	R2 (test)
1	**logK DOPC**	434661	0.100	0.114	0.238	0.259	0.950	0.943
2	**logK octanol**	449128	0.044	0.057	0.155	0.177	0.976	0.969
3	**logP DOPC**	434568	0.424	0.484	0.469	0.510	0.923	0.911
4	**logP GENER**	3717	2.137	2.770	0.851	0.882	0.759	0.682

FAIR data? Ha-ha! :)

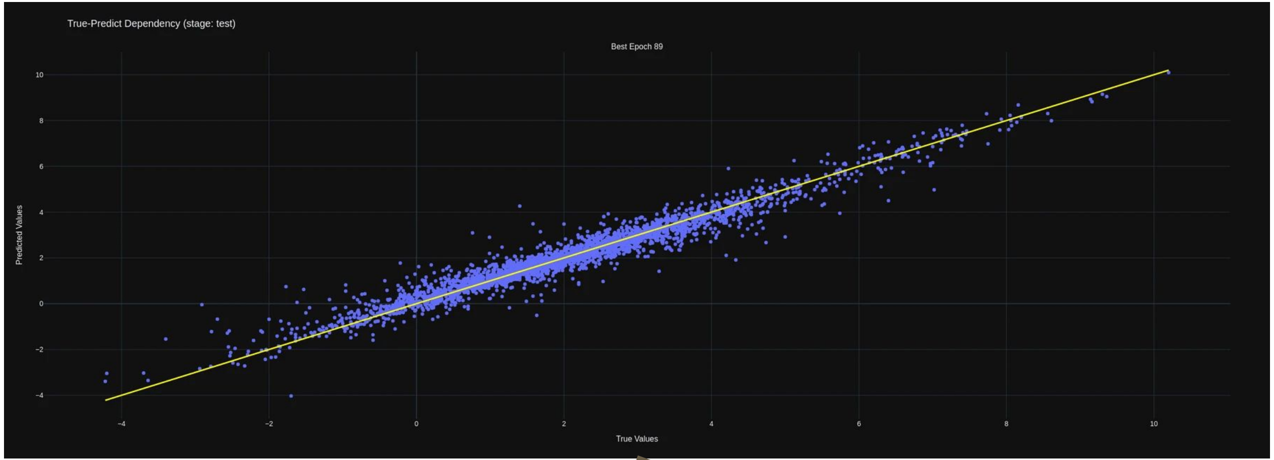
- The LogK data collected in MolMeDb appeared to be *not* the raw data but the *predictions*
  - ALOGPS 2.1: an ancient (2002) Associative Neural Network (ASNN) approach.
- The raw data were from PHYSPROP database:
  - No longer publicly available from ~2020, all links are just broken.
  - Claimed to be moved to EPI Suite software from **US Environmental Protection Agency**.
  - EPI Suite docs mention the same broken links.
  - Binary .db files in the installation are not readable (undocumented proprietary format).
- Data archeology:
  - A paper from 2017 (10.1021/acs.jcim.6b00625) used PHYSPROP (still available back then) to make a curated subset of data and to retrain the models →curated subset still public!
  - Initial PHYSPROP had *tons of issues* (erroneous structures, inconsistencies among the chemical names)
  - In *curated* set: 81 invalid SMILES, 236 too small, 93 mixtures, 42 organometallic, 22 bad valences, 1 duplicate.
  - Remaining 13732 compounds.

FAIR data? Ha-ha! :)

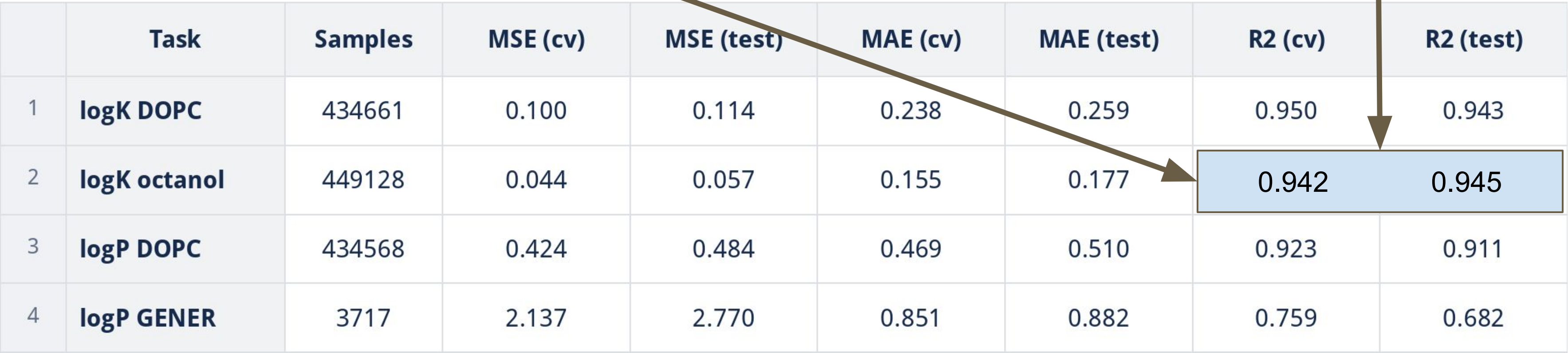
❌ Findable
❌ Accessible
❌ Interoperable
❌ Reusable

Nice job, US Environmental Protection agency! 😉

Membrane permeability: corrected

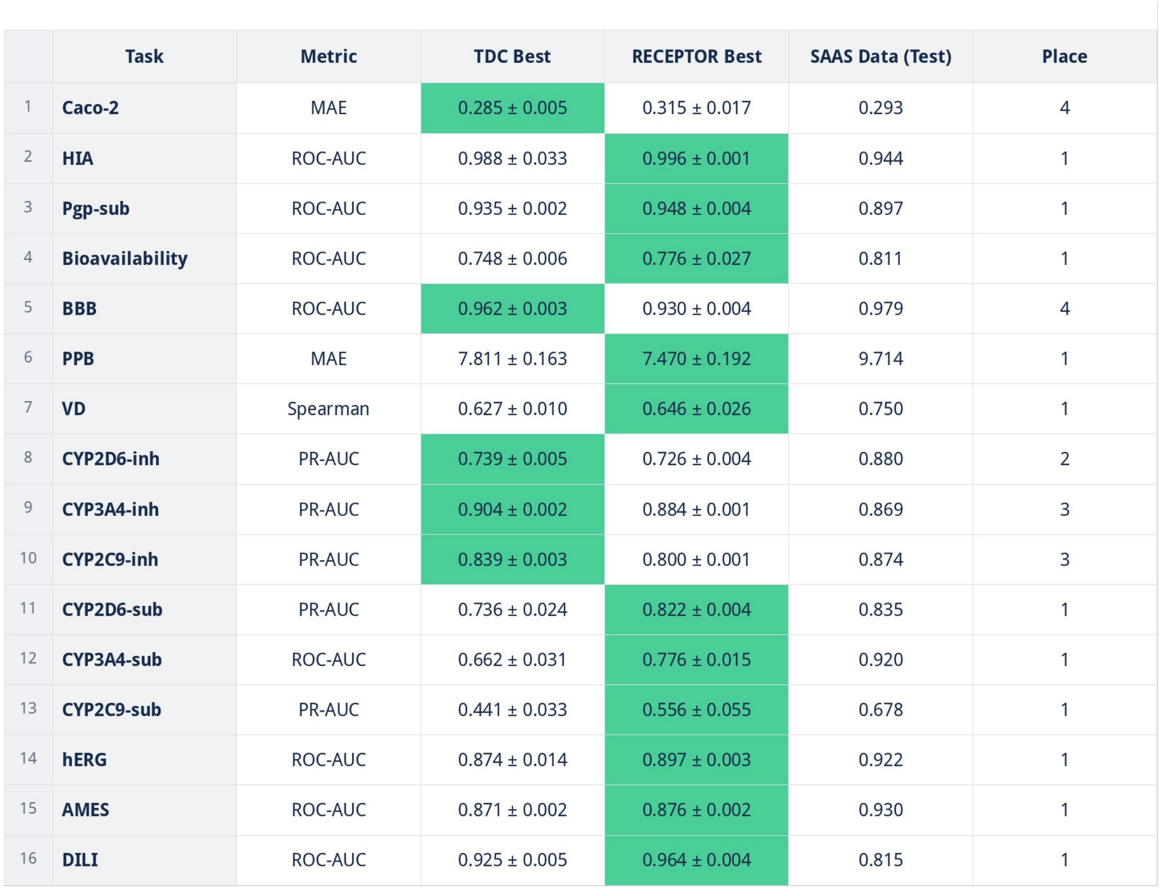


- Model retrained on curated raw data
- Now it’s reasonable!
- Slightly better than existing model (~0.93)



	Task	Samples	MSE (cv)	MSE (test)	MAE (cv)	MAE (test)	R2 (cv)	R2 (test)
1	**logK DOPC**	434661	0.100	0.114	0.238	0.259	0.950	0.943
2	**logK octanol**	449128	0.044	0.057	0.155	0.177	0.942	0.945
3	**logP DOPC**	434568	0.424	0.484	0.469	0.510	0.923	0.911
4	**logP GENER**	3717	2.137	2.770	0.851	0.882	0.759	0.682

TDC benchmarks: ADMET AI models open competition



	Task	Metric	TDC Best	RECEPTOR Best	SAAS Data (Test)	Place
1	**Caco-2**	MAE	0.285 ± 0.005	0.315 ± 0.017	0.293	4
2	**HIA**	ROC-AUC	0.988 ± 0.033	0.996 ± 0.001	0.944	1
3	**Pgp-sub**	ROC-AUC	0.935 ± 0.002	0.948 ± 0.004	0.897	1
4	**Bioavailability**	ROC-AUC	0.748 ± 0.006	0.776 ± 0.027	0.811	1
5	**BBB**	ROC-AUC	0.962 ± 0.003	0.930 ± 0.004	0.979	4
6	**PPB**	MAE	7.811 ± 0.163	7.470 ± 0.192	9.714	1
7	**VD**	Spearman	0.627 ± 0.010	0.646 ± 0.026	0.750	1
8	**CYP2D6-inh**	PR-AUC	0.739 ± 0.005	0.726 ± 0.004	0.880	2
9	**CYP3A4-inh**	PR-AUC	0.904 ± 0.002	0.884 ± 0.001	0.869	3
10	**CYP2C9-inh**	PR-AUC	0.839 ± 0.003	0.800 ± 0.001	0.874	3
11	**CYP2D6-sub**	PR-AUC	0.736 ± 0.024	0.822 ± 0.004	0.835	1
12	**CYP3A4-sub**	ROC-AUC	0.662 ± 0.031	0.776 ± 0.015	0.920	1
13	**CYP2C9-sub**	PR-AUC	0.441 ± 0.033	0.556 ± 0.055	0.678	1
14	**hERG**	ROC-AUC	0.874 ± 0.014	0.897 ± 0.003	0.922	1
15	**AMES**	ROC-AUC	0.871 ± 0.002	0.876 ± 0.002	0.930	1
16	**DILI**	ROC-AUC	0.925 ± 0.005	0.964 ± 0.004	0.815	1

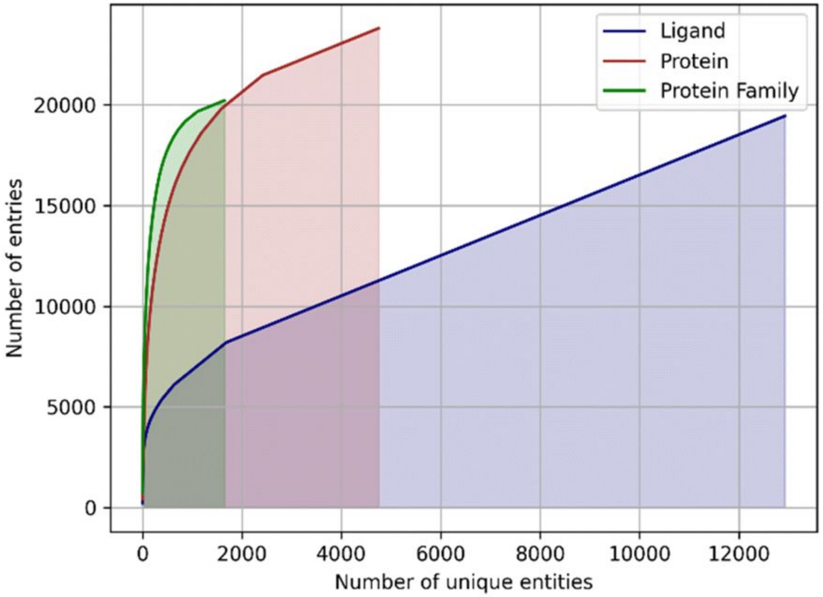
- TDC open benchmarks set https://tdcommons.ai
  - 22 endpoints
  - Public leaderboards
  - Receptor.AI is not officially on TDC yet
- We are overall the best on TDC metrics
- Many endpoints are the absolute best
- Official participation planned in spring 2024

AI docking

- AI models trained on existing protein-ligand complexes.
    - ~10-20k high quality complexes only
    - Not physics-based, force field agnostic
- SMILE or 3D conformer + binding pocket as an input, binding pose as an output.
    - May produce distance matrix or point in dihedral space + post-processing to the pose
- Various representations of protein (AA, residue level, graph, distance matrix, etc.)
- Flexible balance between speed and accuracy

The problem of data with protein-ligand complexes

- There is a limited number of experimentally determined protein–ligand complexes
  - Number of all complexes (X-ray, Cryo-EM, NMR): **< 20k**
  - Hi-quality complexes with binding affinity annotations: **~10k**
- Only 1655 ligands present in >1 complexes
- ~1500 protein bind to 80% of all ligands
- ~100 protein families represent 60% of all data
- Very limited and skewed dataset for ML!

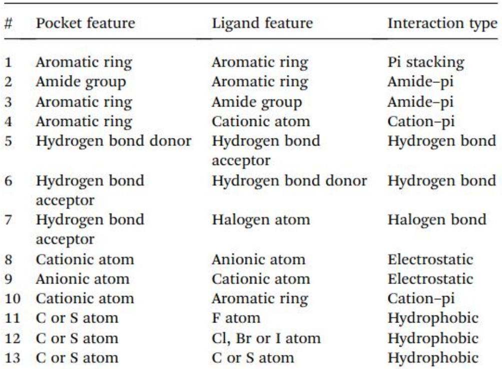


Statistics of PDBbind database

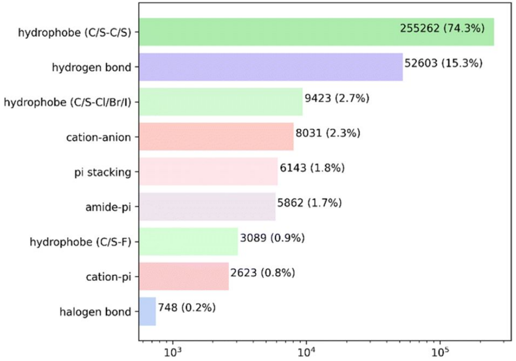
Data augmentation technique

- Take the statistical distributions of interactions in real complexes.
- Generate artificial “binding pockets” around real ligands following these distributions.
- Mix artificial pockets to real ones for model training at different proportions.
- Assumed that all major non-bond interactions are present in experimental data but their *combinations* are not adequately sampled.
- Augmented data teaches the model to recognize corner cases and combinatorial variety of interactions that are absent in the experimental training set.

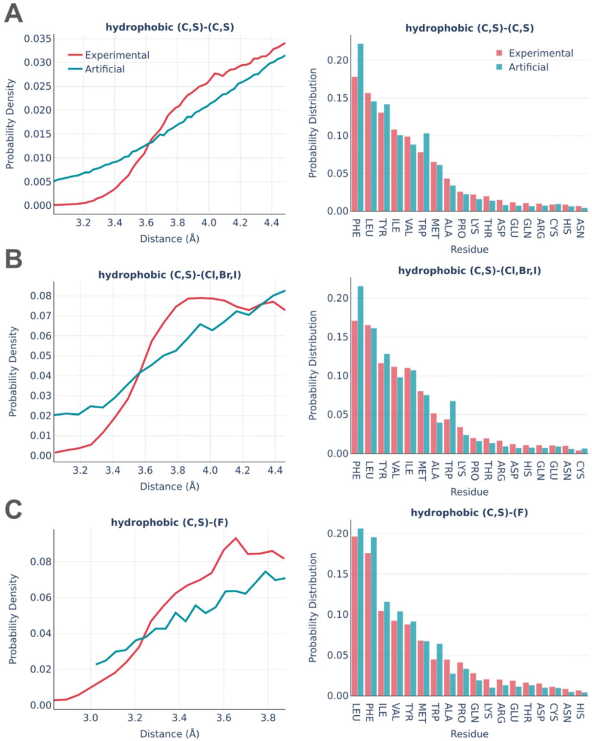
Data augmentation: the details



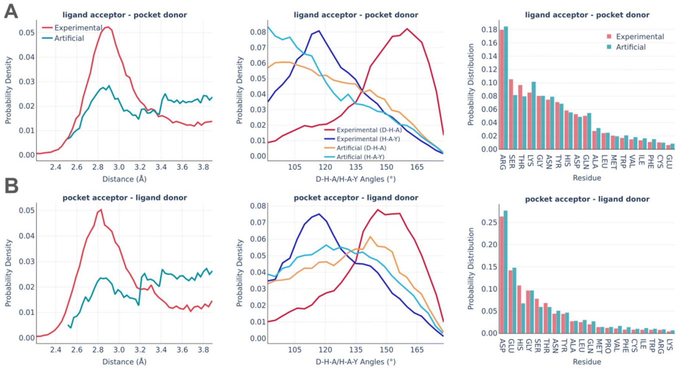
#	Pocket feature	Ligand feature	Interaction type
1	Aromatic ring	Aromatic ring	Pi stacking
2	Amide group	Aromatic ring	Amide–pi
3	Aromatic ring	Amide group	Amide–pi
4	Aromatic ring	Cationic atom	Cation–pi
5	Hydrogen bond donor	Hydrogen bond acceptor	Hydrogen bond
6	Hydrogen bond acceptor	Hydrogen bond donor	Hydrogen bond
7	Hydrogen bond acceptor	Halogen atom	Halogen bond
8	Cationic atom	Anionic atom	Electrostatic
9	Anionic atom	Cationic atom	Electrostatic
10	Cationic atom	Aromatic ring	Cation–pi
11	C or S atom	F atom	Hydrophobic
12	C or S atom	Cl, Br or I atom	Hydrophobic
13	C or S atom	C or S atom	Hydrophobic



Hydrophobic



H-bonds



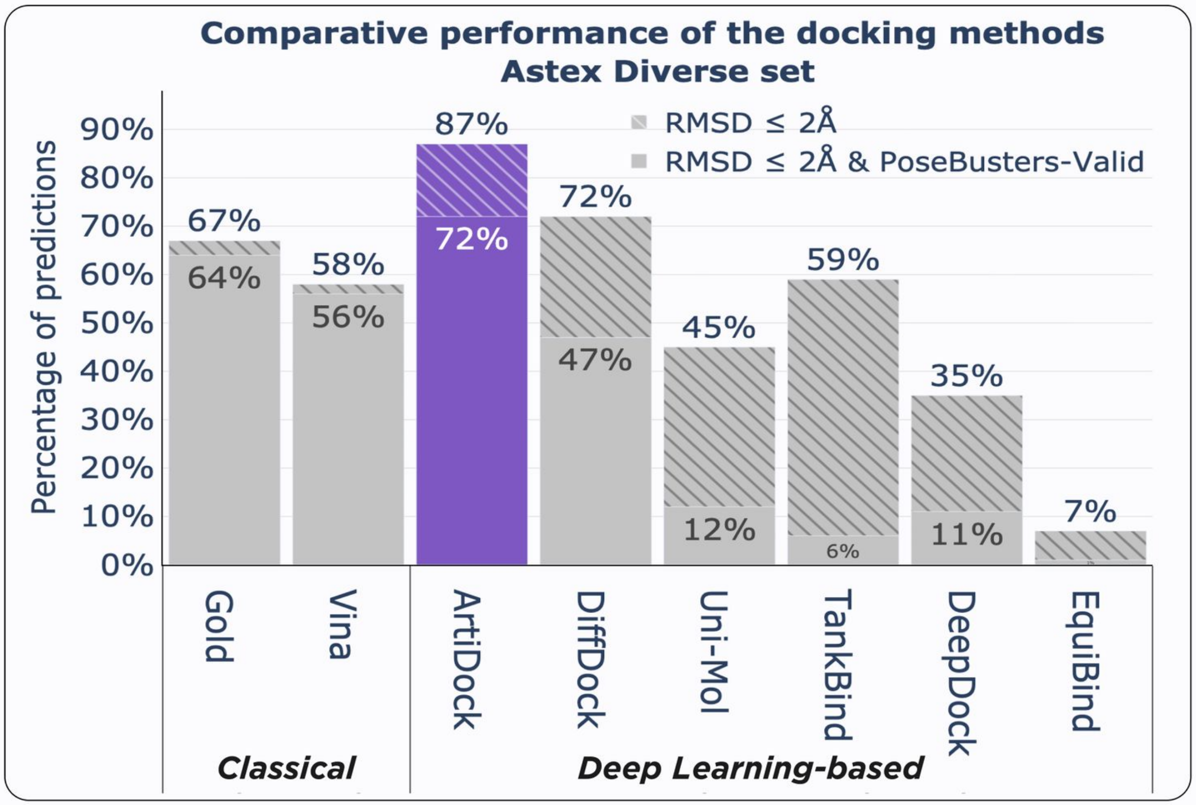
- Reasonable correspondence of distributions
- Potential of improvement at the cost of model training time
- Potential to add explicit ions and cofactors

DOI: 10.1039/D3RA08147H

ArtiDock: next gen ligand binding pose prediction

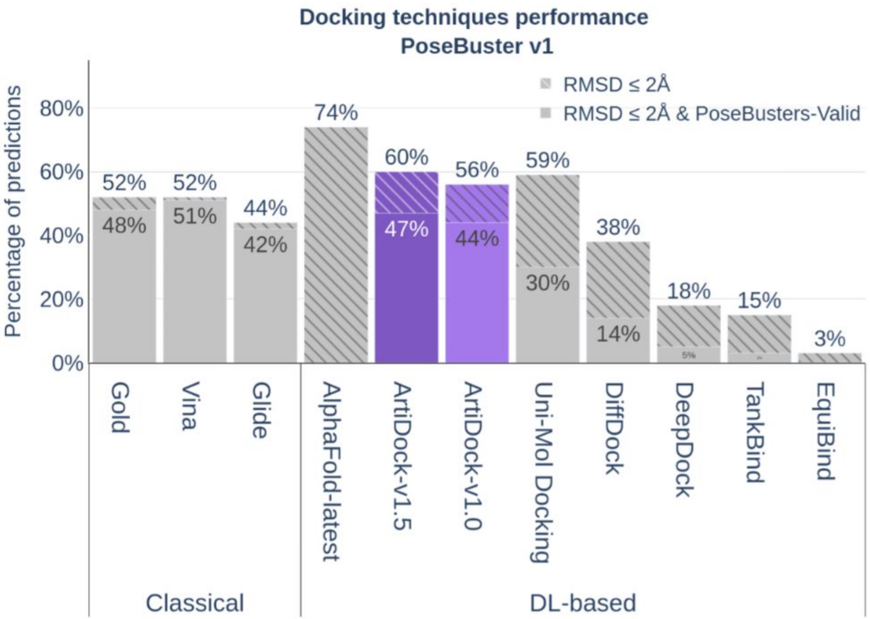
- Small model based on proprietary lightweight GNN architecture
  - Fast training and inference.
- Includes only the binding pocket
  - Less structural noise.
  - Much smaller model.
- Augmenting limited data on protein-ligand complexes with artificial pockets
  - Algorithmic technique for generating “fake” pockets around diverse real ligands.
  - Mimics statistical distributions of various non-bond interactions from experimental pockets.
  - Provides much more combinations of interactions than available in experimental pockets.
- Ability to integrate the protein dynamics
  - Incorporation of processed MD trajectories

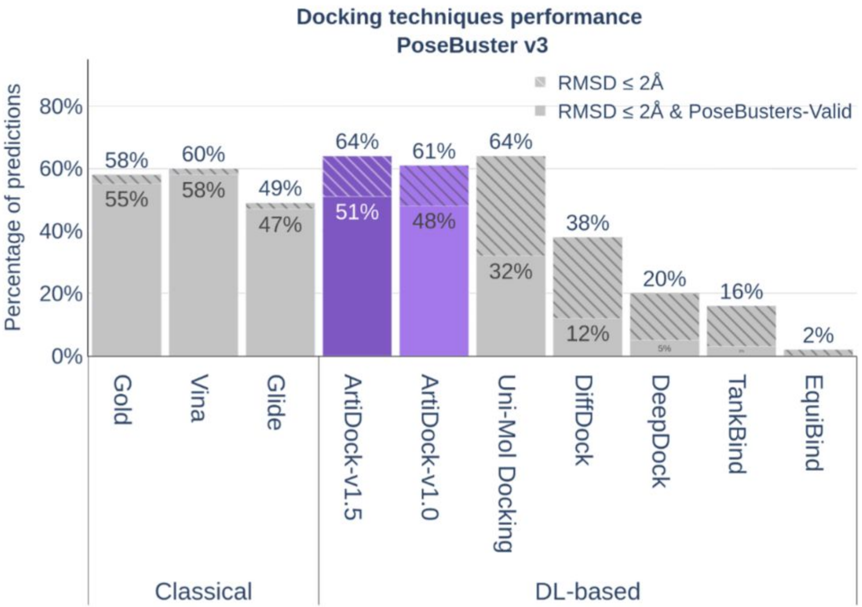
ArtiDock performance: Astex dataset



- Astex is a standard dataset for docking benchmarks
- An older set created before the AI hype
- Considered not particularly challenging for AI methods

ArtiDock performance: PoseBusters dataset





PoseBusters dataset

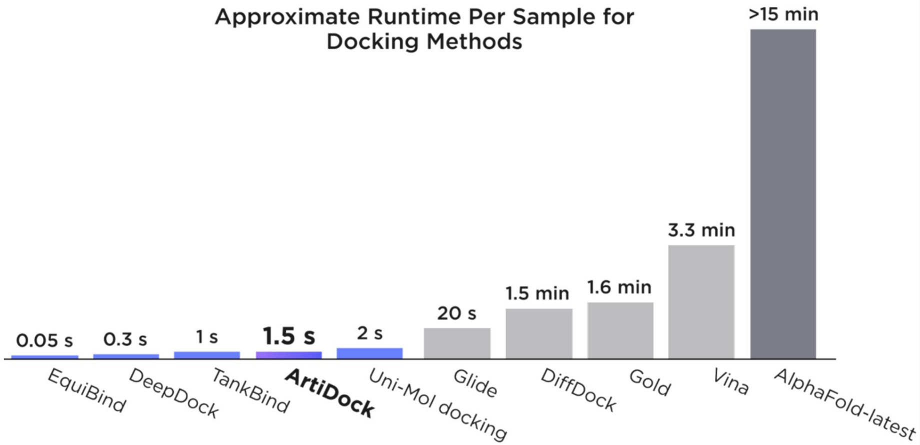
- DOI: 10.1039/D3SC04185A
- Includes multiple structure quality metrics beyond RMSD
- Designed to ashame AI docking
- Ashamed by the next-gen AI docking 🙂

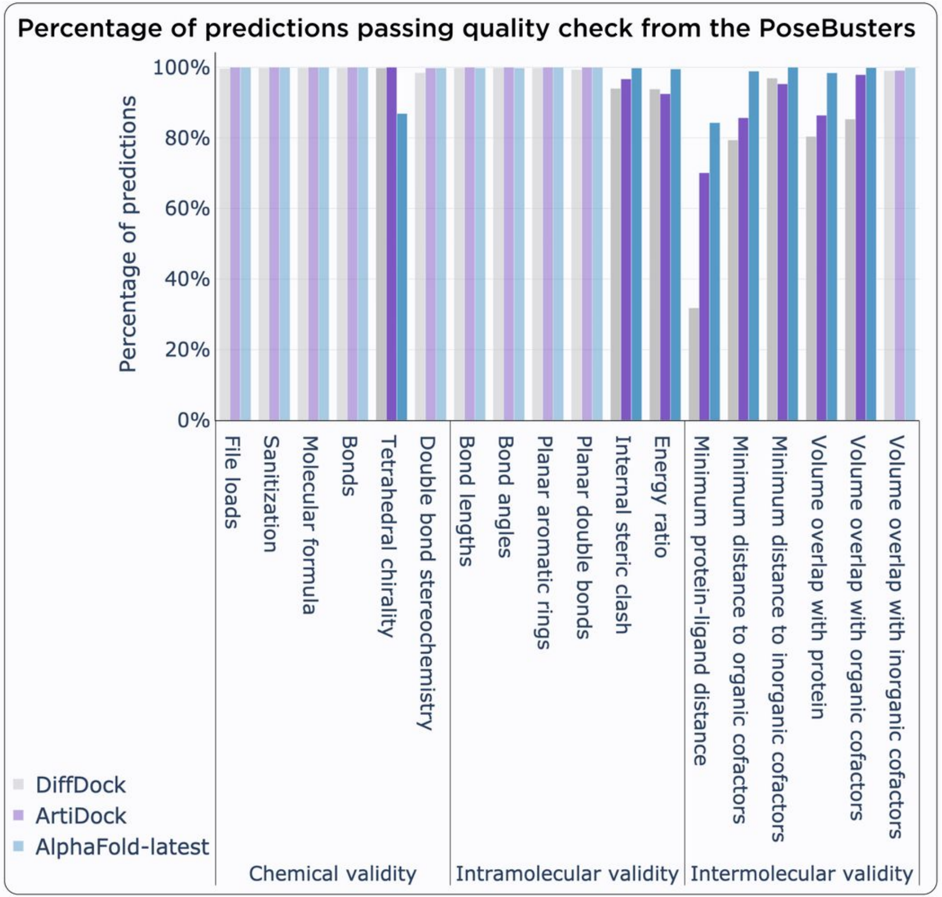
PoseBusters versions

- V1 was made public in 2023 in the preprint
- V3 published and peer reviewed
- V3 is adjusted in favor of conventional docking and against AI even more (artificial bias)
- AI method still pass it 🙂

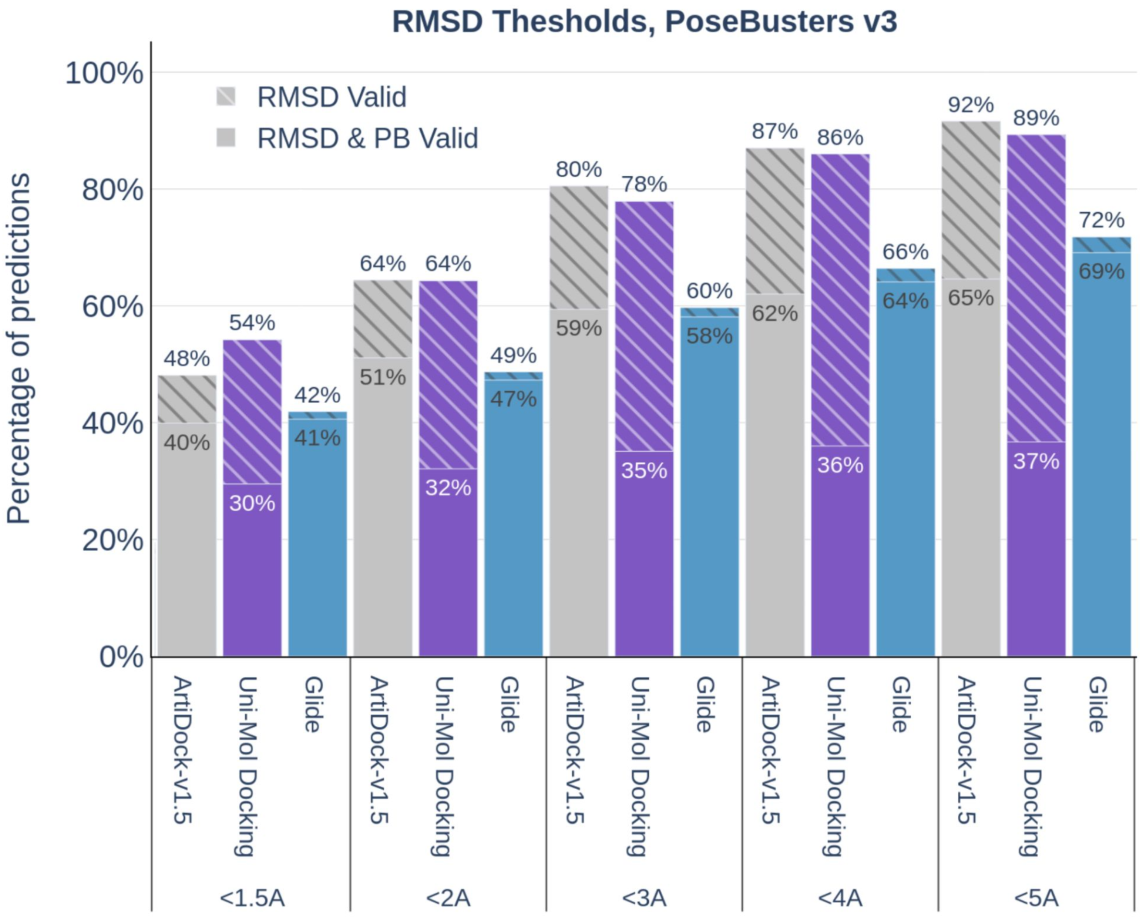
ArtiDock performance

- Outperforms all ML methods
- Comparable to conventional docking
- Faster than all of them


Approximate Runtime Per Sample for Docking Methods


Percentage of predictions passing quality check from the PoseBusters

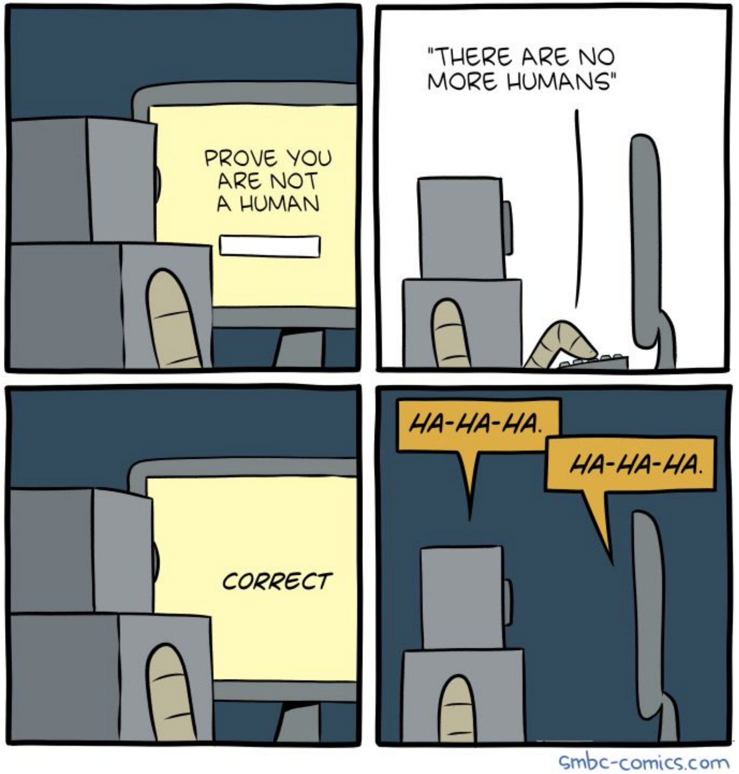
Detailed comparison with Glide and UniMol



- PB-Valid scores dependence on RMSD cutoff
  - ArtiDock and Glide: *increase*
  - Uni-Mol: *constant*
- Absolute PB-Valid scores:
  - ArtiDock and Glide: *comparable*
  - Uni-Mol: *low*
- Scores: ArtiDock ~ Glide
- Speed: ArtiDock >> Glide
- Uni-Mol prioritizes RMSD but fails miserably on PB-Valid

Conclusions

- AI drug discovery techniques are here to stay
- Pharma companies adoption increases
- Data mining and analysis seems to be dominated by LLMs
- Progressive substitution of the “physics-based techniques” by “data driven” ones (will docking finally die for good?)
- Data is a new oil (but nobody wants to collect and curate it)



THANK YOU
FOR YOUR ATTENTION
KINOPO
meme-arsenal.ru