

---

# Artificial intelligence in drug discovery

**Semen Yesylevskyy**

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

# The uncomfortable truth about drug discovery

When you decide to go into the drug discovery...

Expectations:



# The uncomfortable truth about drug discovery

When you decide to go into the drug discovery...

Expectations:

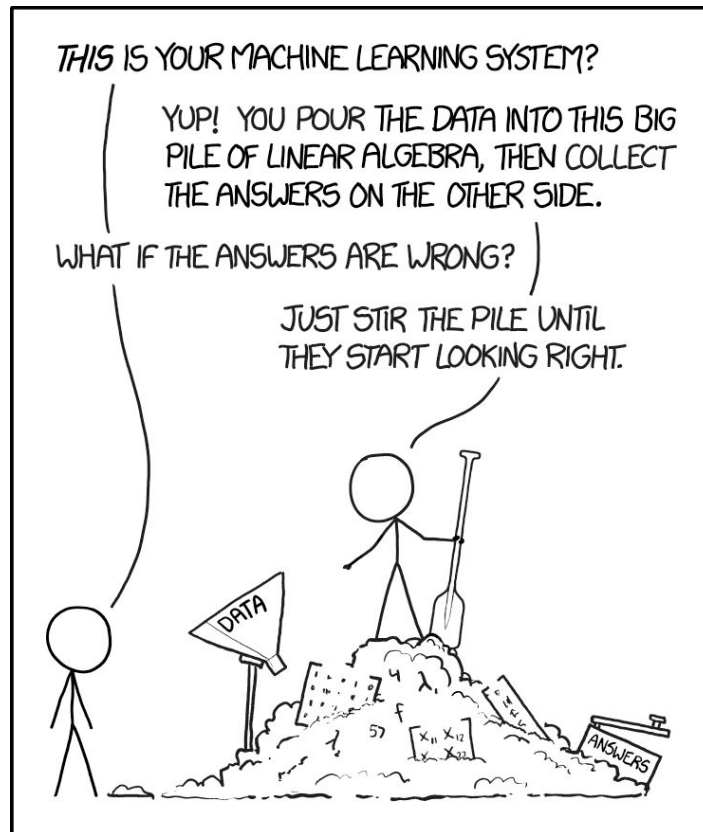


Reality:

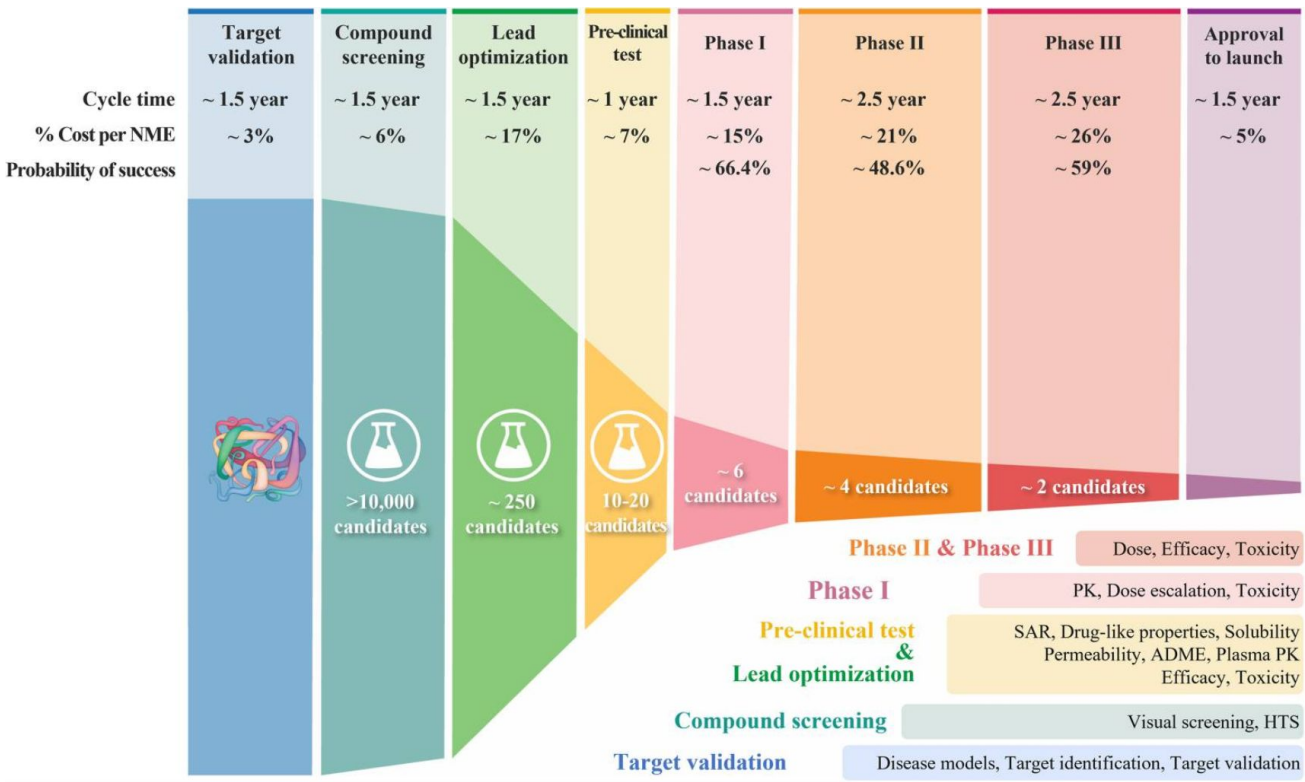


# Plan of the talk

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



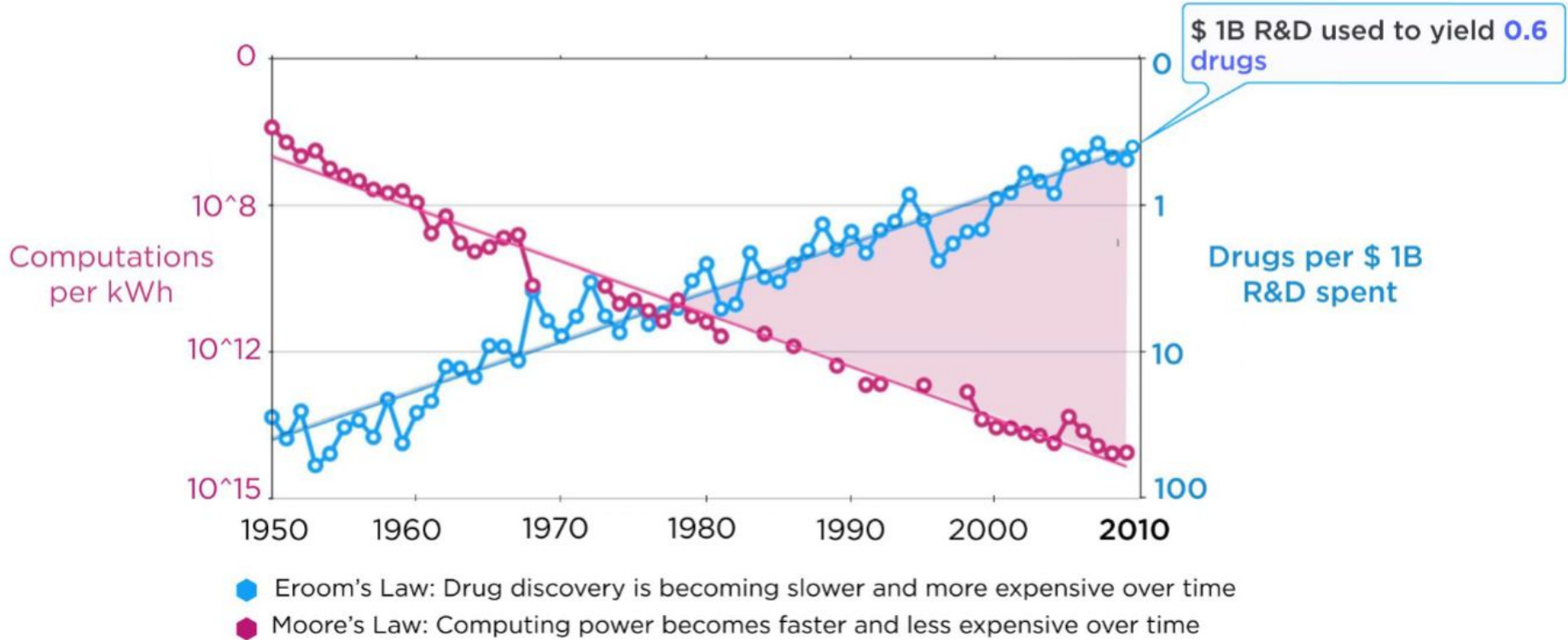
# Modern drug discovery struggles badly



## Reasons of stagnation

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

# Eroom's law: are we cursed?



Computational resources become cheaper but this doesn't help at all so far...



# Eroom's law explained (kind of)

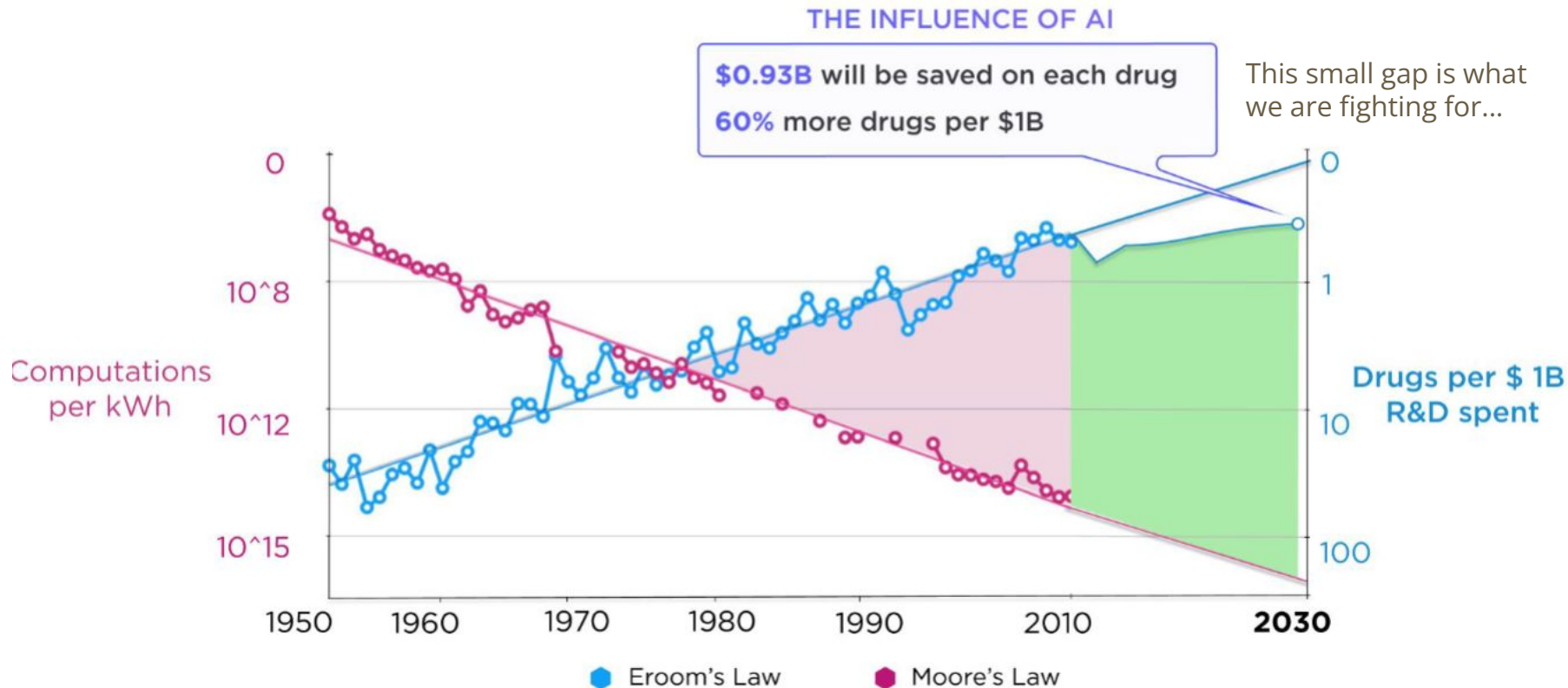
- **The 'better than the Beatles' problem:** very hard to beat established treatments to the extent that it's economically viable.
- **The 'cautious regulator' problem:** level of required evidence in trials become a burden.
- **The 'throw money at it' tendency:** The tendency to add excessive resources 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. 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**
  - General paradigm change.
  - Estimated 60% more drugs per \$1B by 2030.
- **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?



# 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?
- How many iterations to perform?
- What data should be generated?
- 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

- Multi-omics (genomics, transcriptomics, proteomics, interactomics, metabolomics)
- Knowledge graphs
- Unstructured data scraping (papers, patents)

## Early discovery

- De novo molecular generation
- AI virtual screening
- ADMET prediction
- Automatic QSAR comprehension
- Drug repurposing

## Late discovery

- Formulation optimization
- IND and clinical studies outcome prediction
- Data mining for patent clearance
- Simulated *in vivo* testing

## Clinical studies

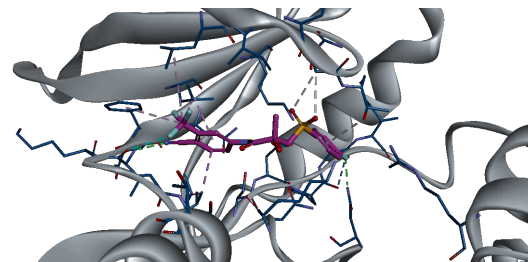
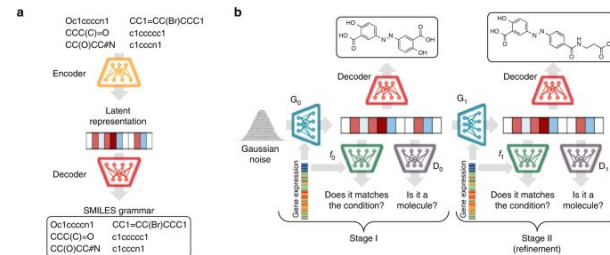
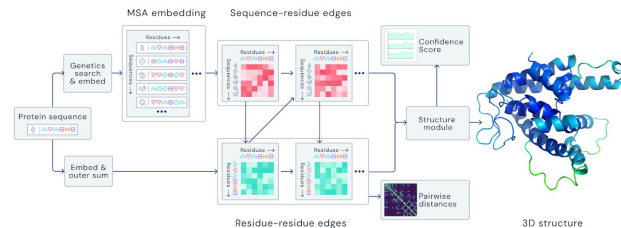
- Clinical study planning and monitoring
- Risk factors prediction
- Automated patient recruitment and triage
- On-the fly adaptive data analysis

## Data management

- Automatic data mining and integration
- Data quality assessment
- Data generation plans
- Explainable data

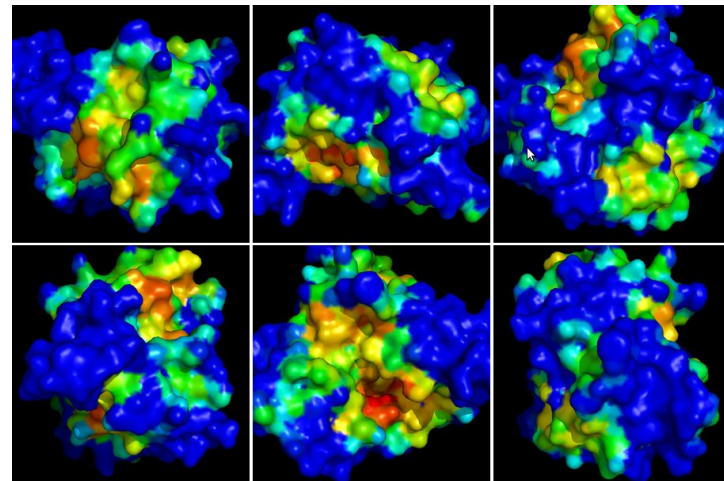
# AI in early drug discovery

- Protein structure prediction
  - AlphaFold, RoseTTAFold
- Binding pocket prediction and prioritization
- Chemical space generation
  - Molecular generators (Chemistry42, Iktos)
  - Scaffold hopping
  - Substituents generation
- Ligand pose prediction (AI docking)
  - DiffDock, UniMol, ArtiDock
- Predicting dynamic properties
  - Protein ensembles (AI conformation generation / AI-enhanced MD)
  - Transient / cryptic binding pockets prediction



# Case study: LLMs in binding pocket prioritization

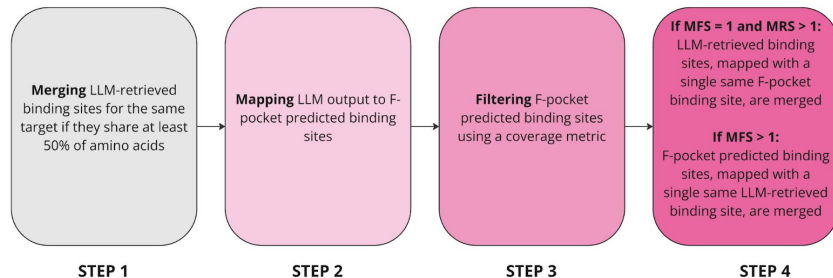
- There are a lot of algorithmic techniques to find “pocket like” cavities on the protein surface.
  - Fpocket is one of the most used.
- Predicts much more pockets than biologically relevant or somehow validated
  - Tedious manual filtration by searching the literature for residues that are confirmed to be involved in the ligand binding.
- Can we automate it by using LLMs?



# Experimental setup

- Test set of proteins:
  - DNA polymerase alpha catalytic subunit
  - Tyrosine-protein kinase ABL1
  - 5-hydroxytryptamine receptor 2A
  - Muscarinic acetylcholine receptor M2/3
  - Sodium channel types 4, 7
  - Programmed cell death 1 ligand 1
  - Gamma-aminobutyric acid receptor
  - KRas kinase
  - Dihydroorotate dehydrogenase
  - Mixed lineage kinase domain-like protein
- For each protein 4-7 peer-reviewed research articles (45 in total) + 3D structures from PDB.
- Baseline defined as pockets identified by several human experts using the same literature.

- Articles complexity tears:
  - One binding pocket for single target.
  - Multiple binding pockets for single target.
  - Multiple binding pockets for target and other proteins.
  - No pocket description (negative control).

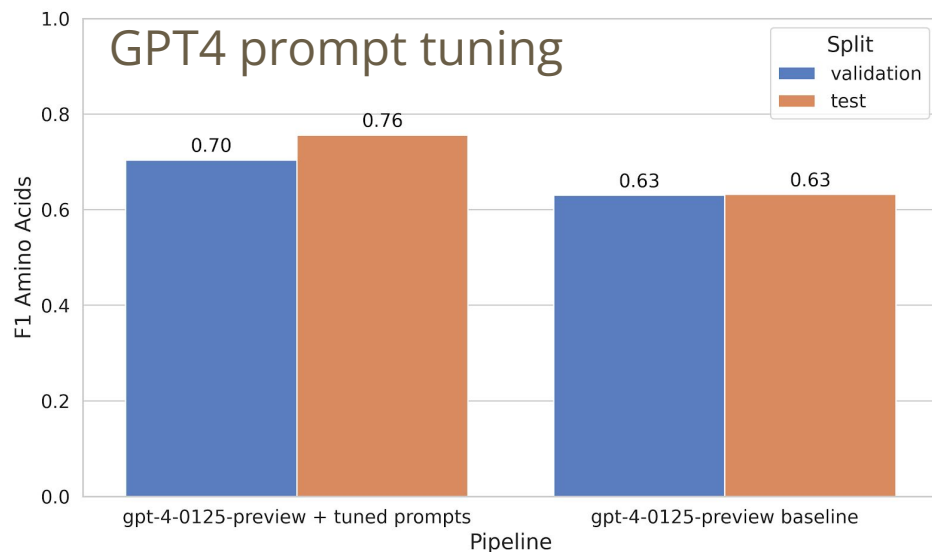
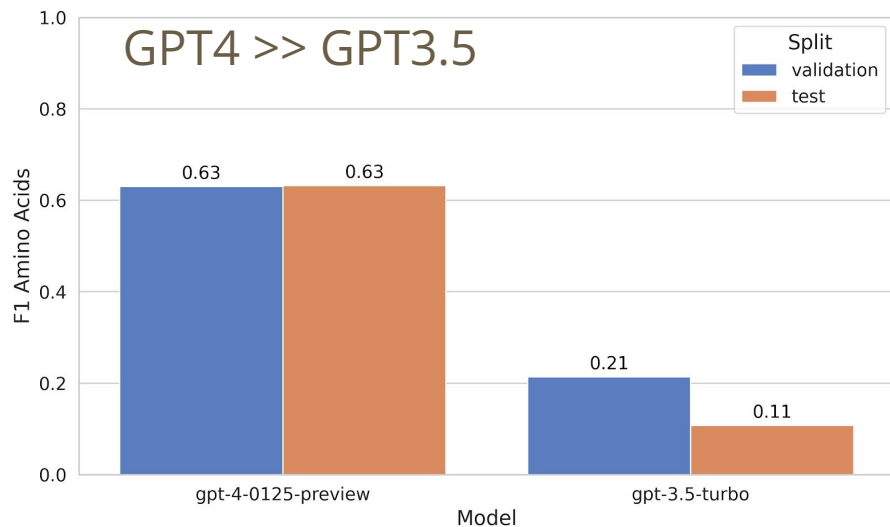


# Results improves as LLMs progress

- LLM prompt (simplified):

You are a Senior Medicinal Chemist exploring binding pockets for {target\_protein}".

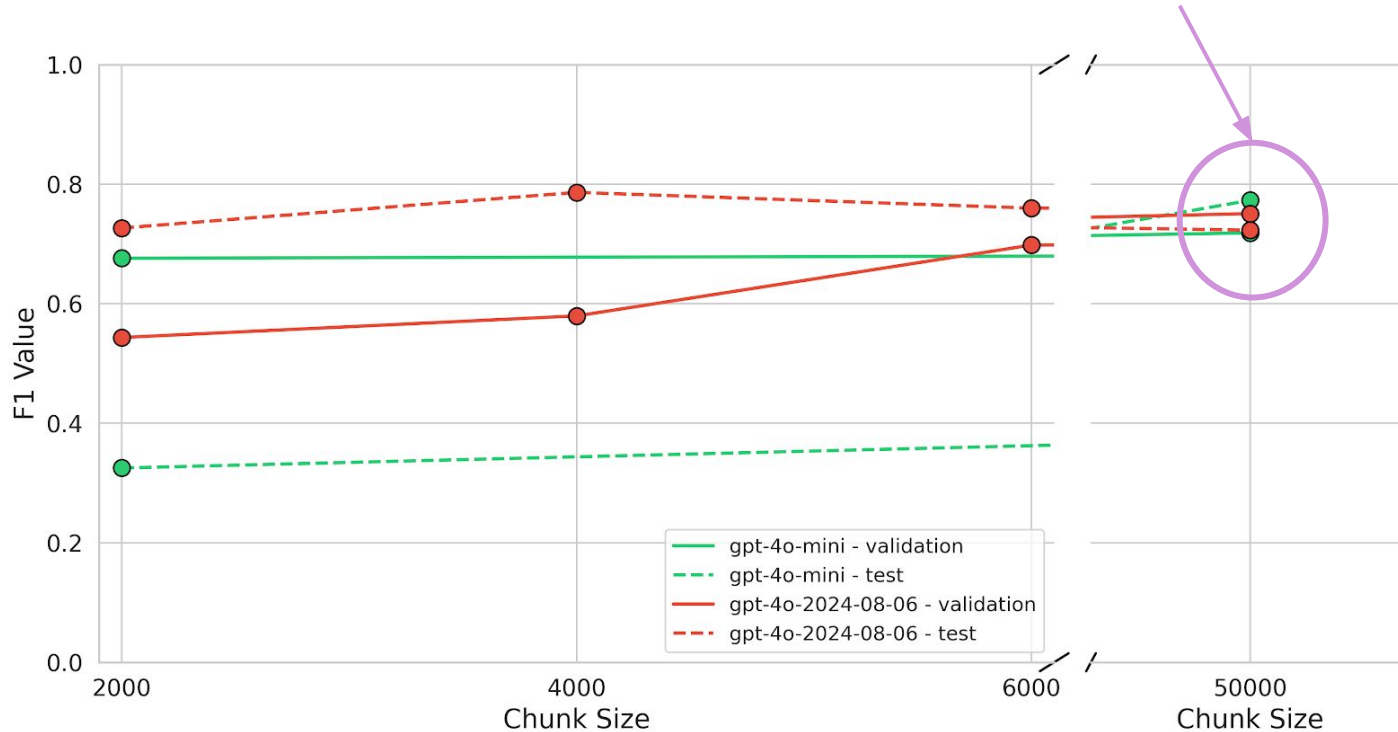
- determine the number of the unique binding pockets in {target\_protein} described in the text;
- make a short, very specific and discriminative characteristic for each of the binding pockets;
- output the list of amino acids forming each of the binding pockets.





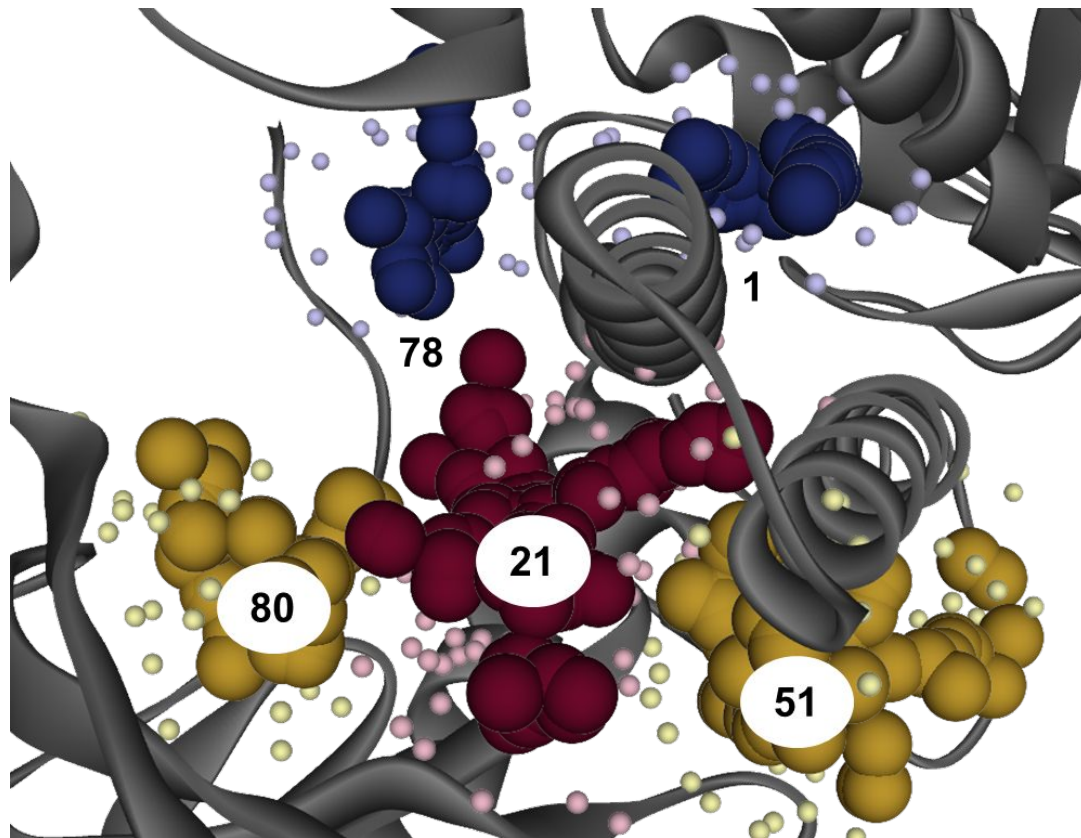
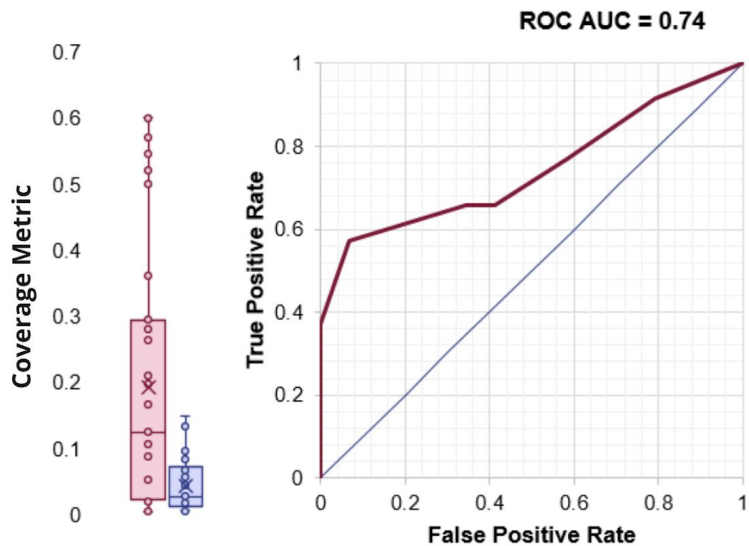
# Larger model is not necessarily better

- GPT4o is even better than GPT4
- GPT4o-mini is as good as the “large” (and more expensive) 4o!



# Merging and filtering the pockets

- LLM is good in filtering out “bad” pockets
- However, post-processing is required to merge overlapping “good” pockets and to tidy them up.



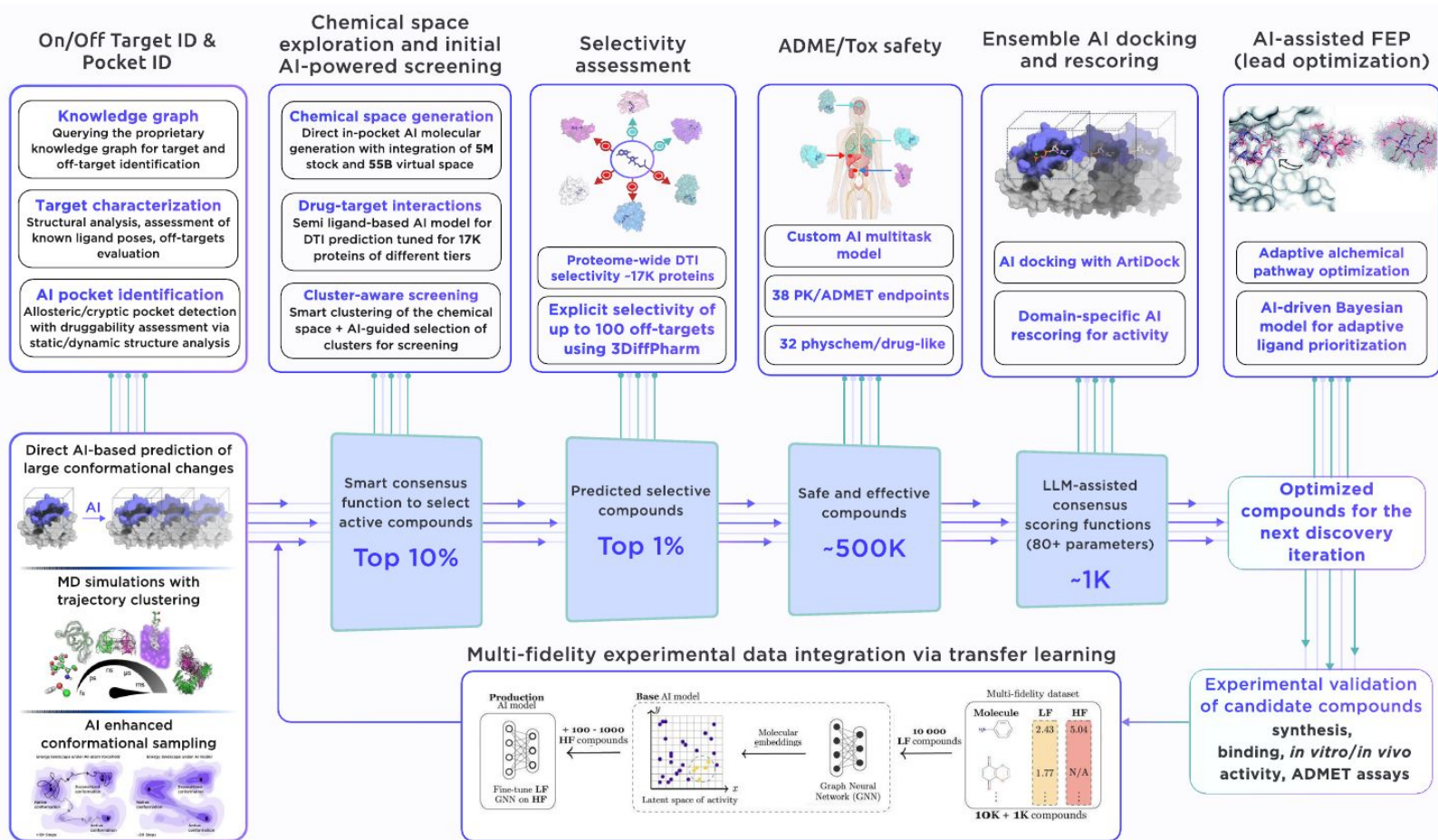
# LLMs in binding pocket prioritization: conclusions

- Accuracy of LLM pocket prioritization with GPT4o is overall decent but not great.
- Various tricks and post-processing are still needed to get usable results.
- LLMs are progressing fast in understanding biological context and reasoning, but...
  - Price and inference time also increase
  - Scientific papers remain hard to parse (tables, figures, etc)
- Claims that “LLMs will soon replace human researchers” look unjustified so far.
  - LLMs are likely to be our assistants rather than our replacement.
  - OpenAI promises that o3 will be a game changer. Will see.



\* Satiric CGI from Boston dynamics, no animals were harmed.

# AI virtual screening pipeline



# 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:

- Carcinogenicity (OSF)
- Carcinogenicity (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

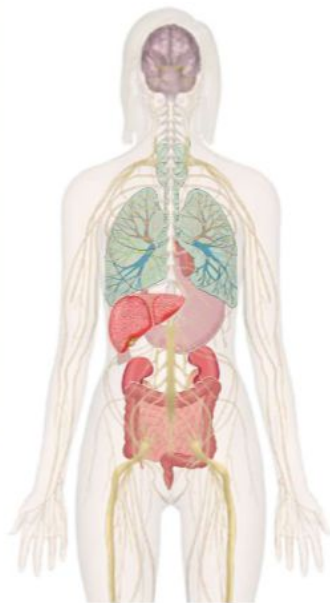
#### Acute toxicity:

- 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

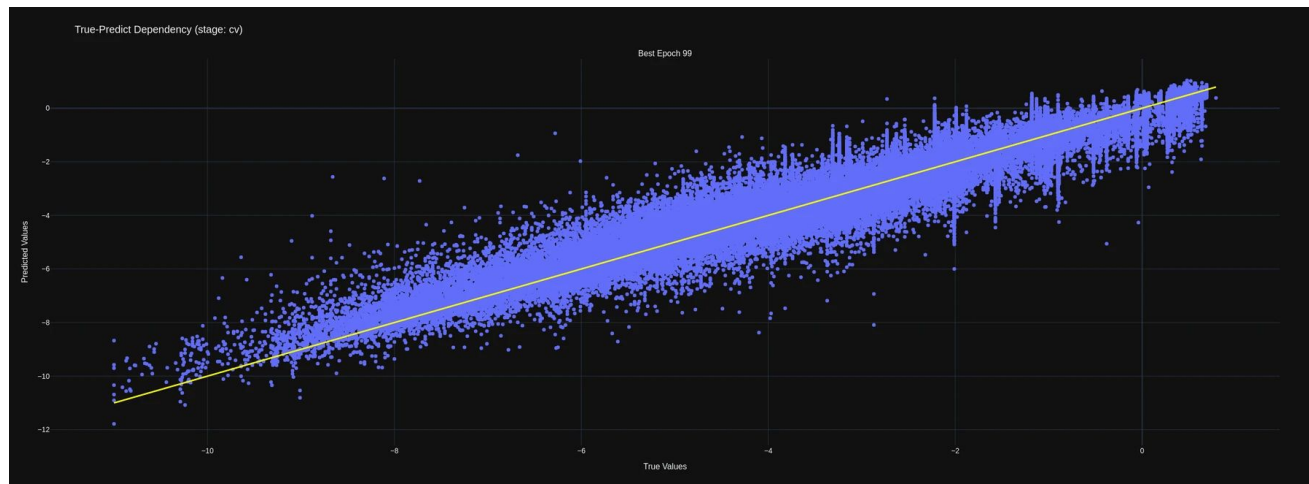
#### PhysChem Parameters:

- 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

# Case study: membrane permeability

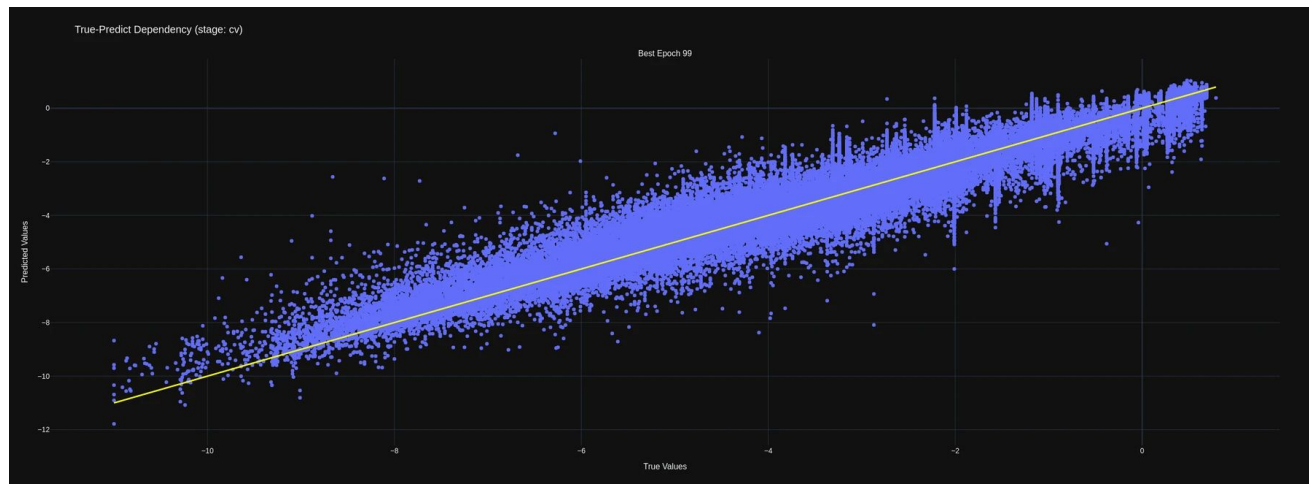


- MolMeDb data for
  - Membrane permeability
  - Membrane partitioning
- Receptor.AI MultiTask ADMET NN architecture
- AutoML automatic featurization

	Task	Samples	MSE (cv)	MSE (test)	MAE (cv)	MAE (test)	R2 (cv)	R2 (test)
1	<b>logK DOPC</b>	434661	0.100	0.114	0.238	0.259	0.950	0.943
2	<b>logK octanol</b>	449128	0.044	0.057	0.155	0.177	0.976	0.969
3	<b>logP DOPC</b>	434568	0.424	0.484	0.469	0.510	0.923	0.911
4	<b>logP GENER</b>	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 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](https://doi.org/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.
  - Remained 13732 compounds.


# FAIR data? Ha-ha! :)

- ~~X~~ Findable
- ~~X~~ Accessible
- ~~X~~ Interoperable
- ~~X~~ Reusable



Nice job, US Environmental Protection agency! 😏

 + Machine Learning = 



Data

 + Artificial Intelligence = 

Data

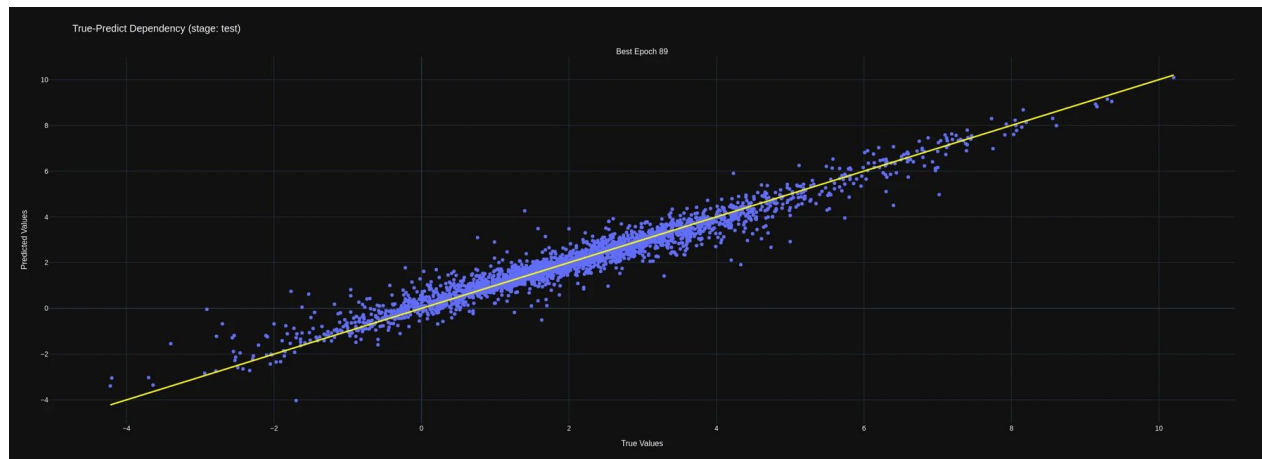
 + Generative AI = 

Data

 + Agentic AI = 

Data

# 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

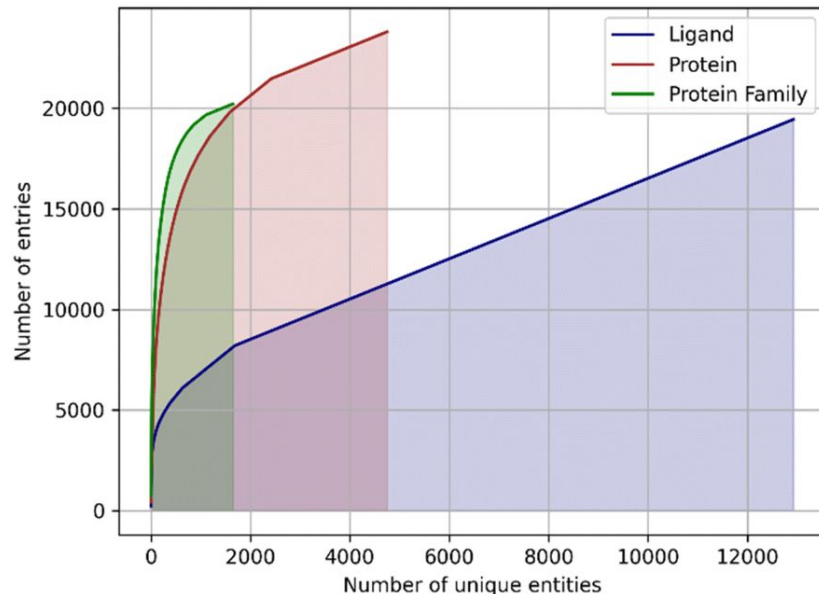
- TDC open benchmarks set <https://tdcommons.ai>
  - 22 endpoints
  - Public leaderboards
- We are overall the best on TDC metrics
- Many endpoints are the absolute best
- Official participation planned in ~~2024~~ 2025

# 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
  - Total number of complexes: **~55k**
  - Number of all complexes with measured affinities (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

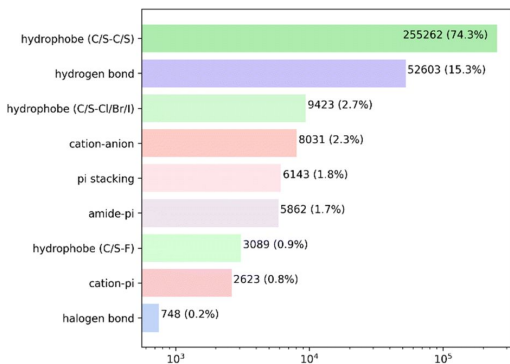


# Data augmentation technique

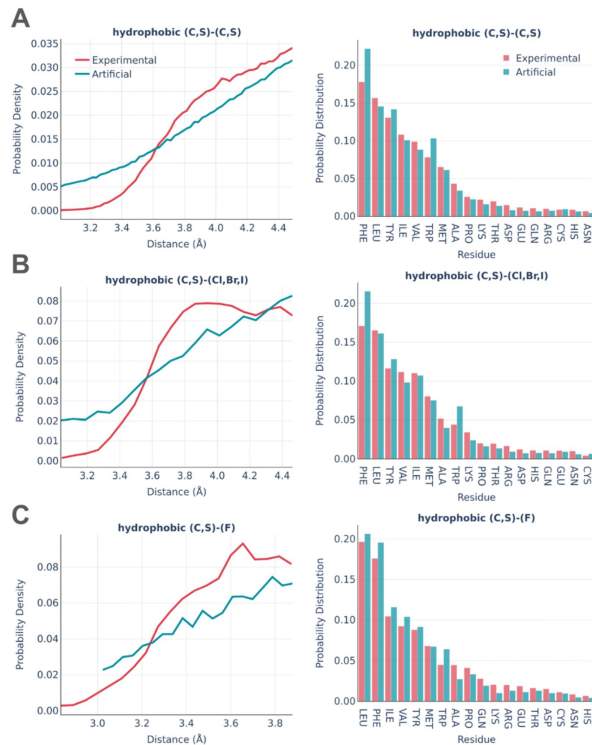
- Take the statistical distributions of interactions in real complexes.
- Generate artificial “binding pockets” around diverse ligands following these distributions.
- Mix artificial pockets to real ones for model training at different proportions.
- Assumed that all major non-bond interactions are present 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

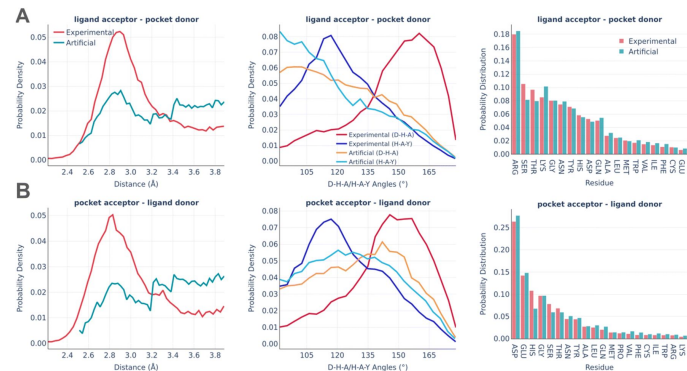
#	Pocket feature	Ligand feature	Interaction type
1	Aromatic ring	Aromatic ring	Pi stacking
2	Amide group	Aromatic ring	Amide- $\pi$
3	Aromatic ring	Amide group	Amide- $\pi$
4	Aromatic ring	Cationic atom	Cation- $\pi$
5	Hydrogen bond donor	Hydrogen bond acceptor	Hydrogen bond
6	Hydrogen bond acceptor	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

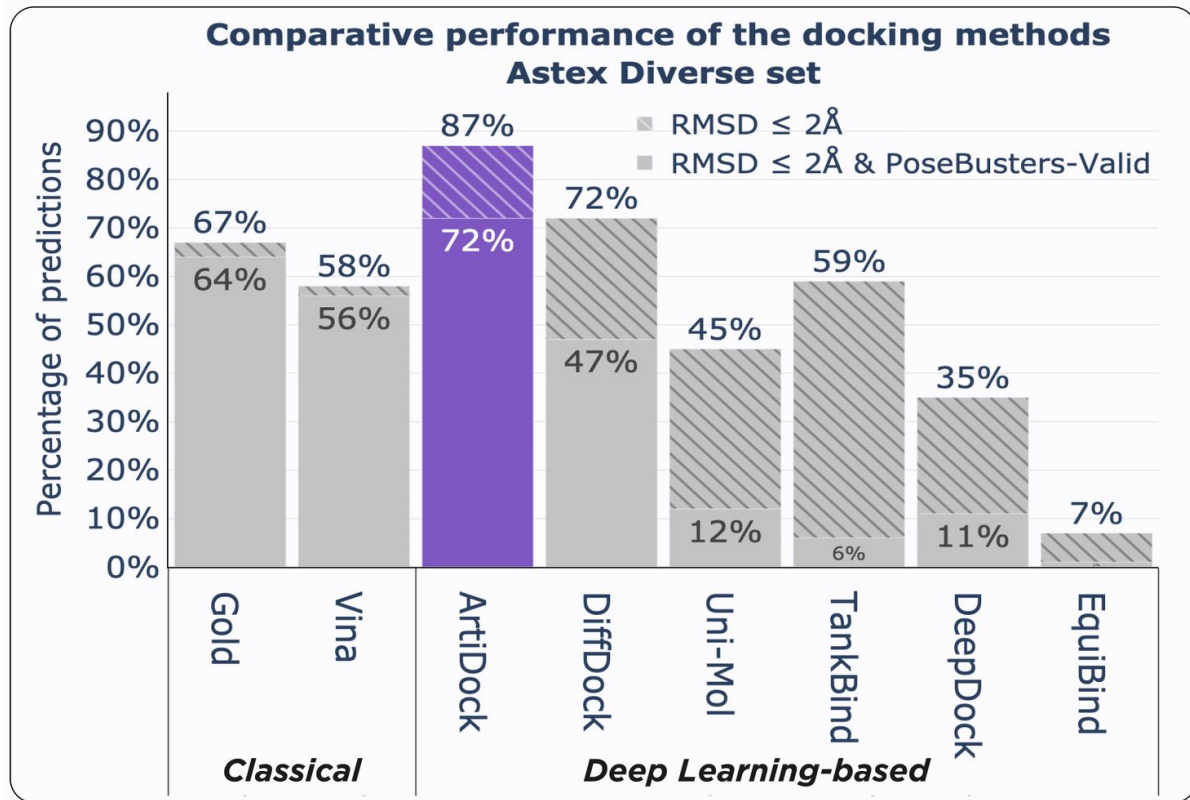


- Reasonable correspondence of distributions
- Potential of improvement at the cost of model training time
- Potential to add explicit ions and cofactors

# ArtiDock: next-gen ligand binding pose prediction

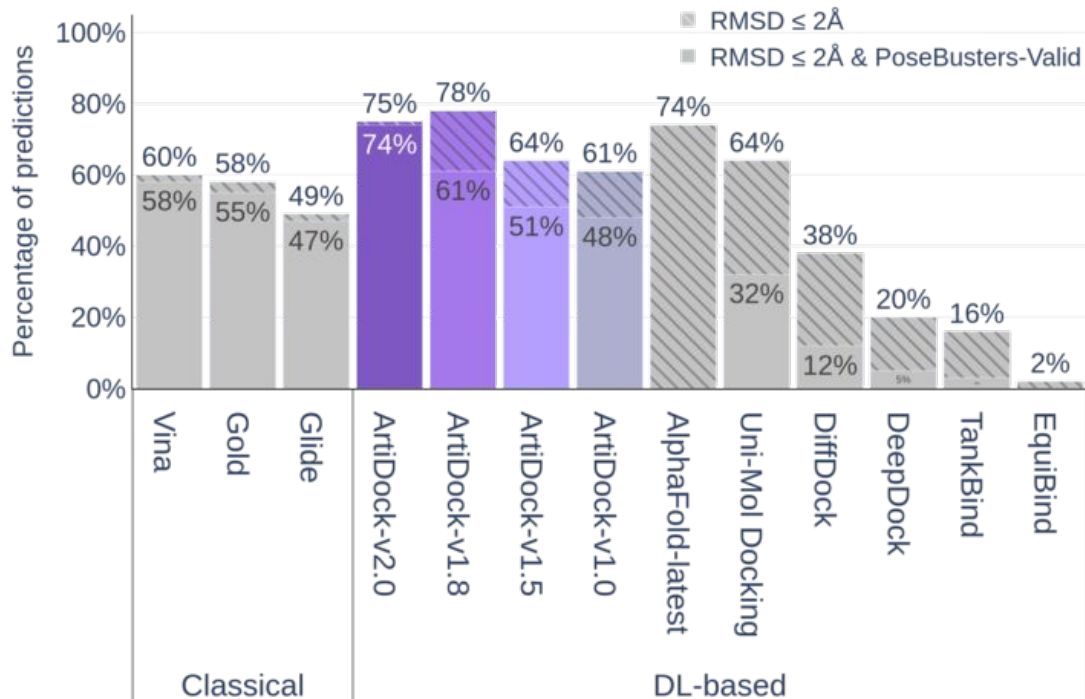
- Small model based on proprietary lightweight GNN architecture
  - Fast training and inference.
- Includes only the binding pocket
  - Less structural noise.
  - Much smaller and faster model.
- Augmenting limited data on protein-ligand complexes with artificial pockets
  - Algorithmic technique for 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 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](https://doi.org/10.1039/D3SC04185A)
- Includes multiple structure quality metrics beyond RMSD
- Designed to shame AI docking
- Ashamed by the next-gen AI docking 😊

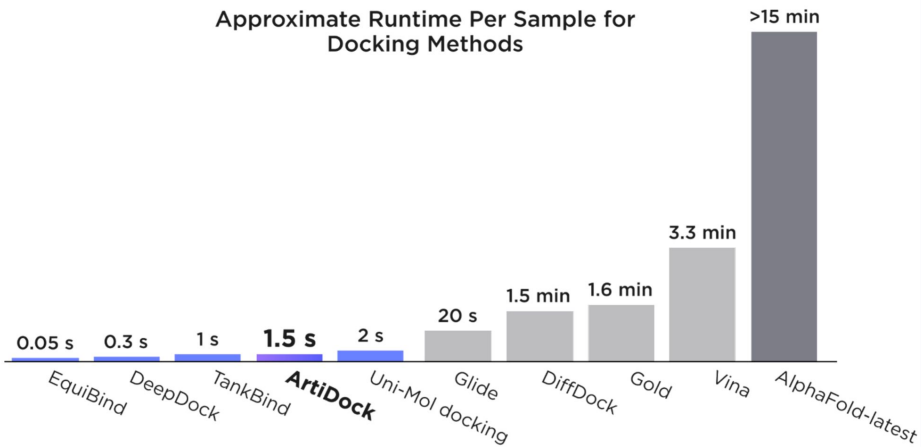
## PoseBusters versions

- 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)
- Latest AI models in 2025 seem to be overfitted against it!

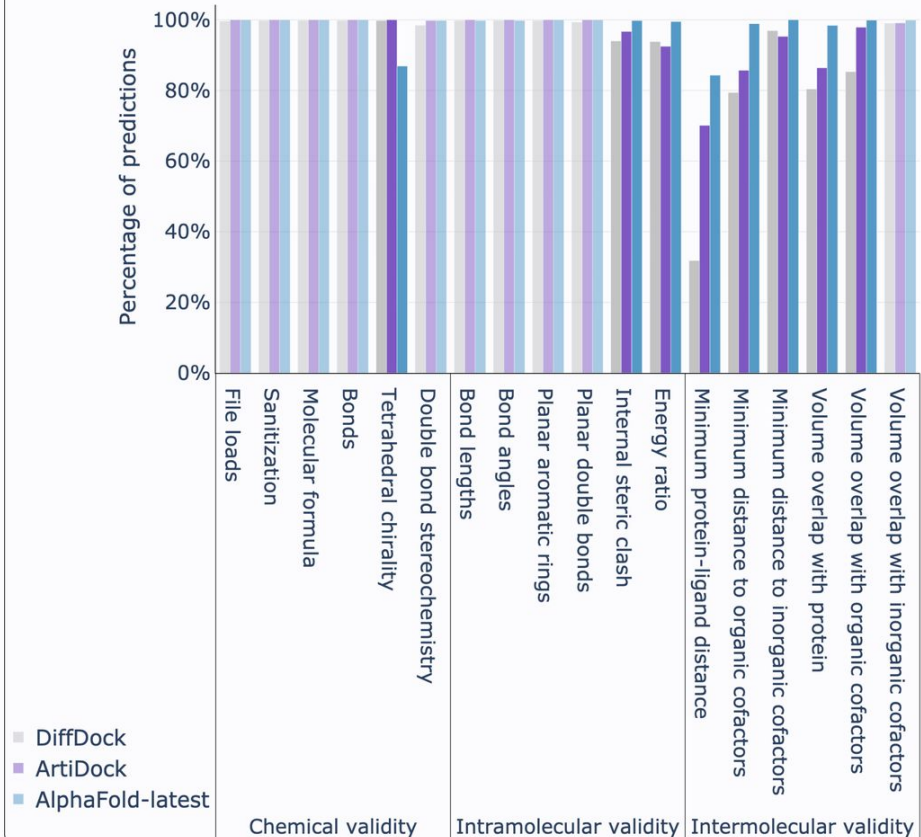
# ArtiDock performance

- Outperforms comparable ML methods.
- On par with conventional docking.
- Faster than anything else of comparable quality.

Approximate Runtime Per Sample for Docking Methods

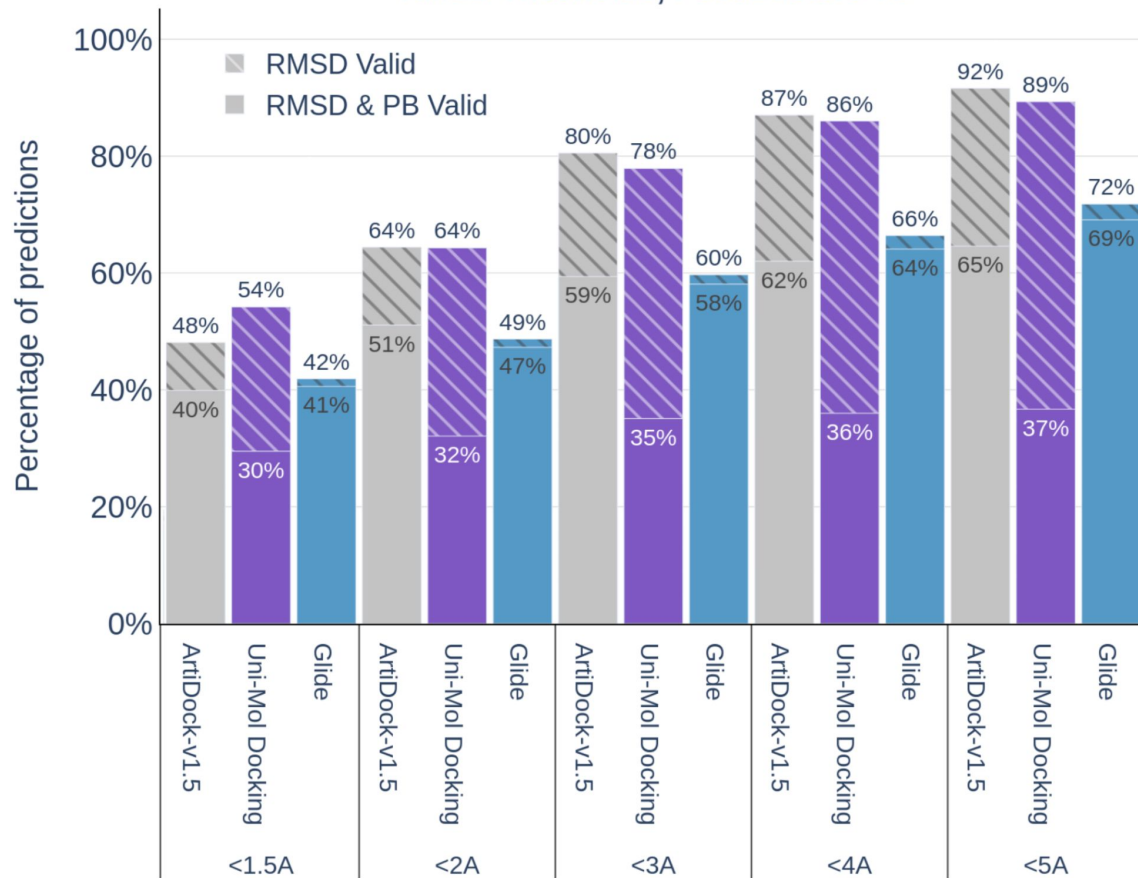


Percentage of predictions passing quality check from the PoseBusters



# Detailed comparison with Glide and UniMol

RMSD Thesholds, PoseBusters v3

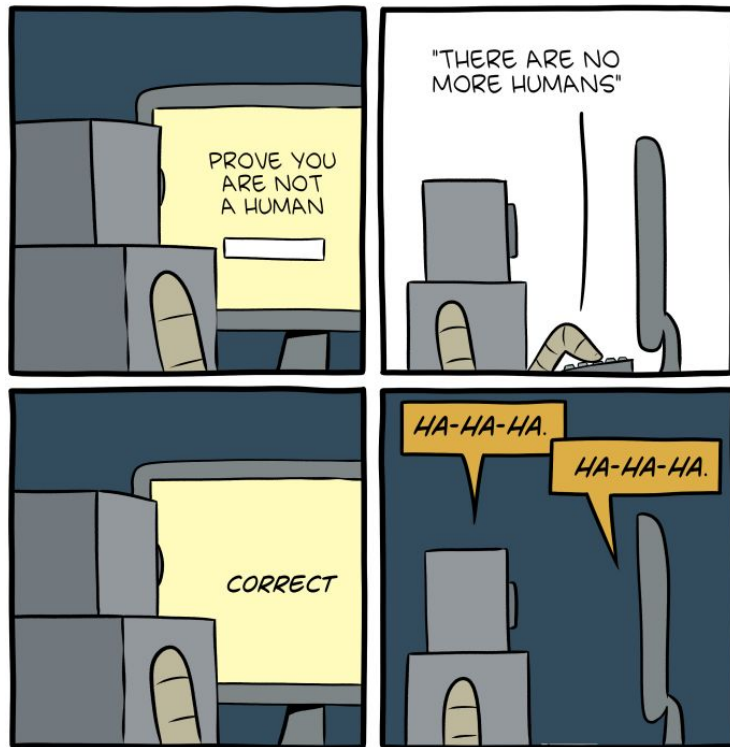


- PB-Valid scores dependence on RMSD cutoff:
  - ArtiDock and Glide: *increase*
  - Uni-Mol: *constant*
- Absolute PB-Valid scores:
  - ArtiDock and Glide: *comparable*
  - Uni-Mol: *low*
- Scores: ArtiDock ~ Glide
- Speed: ArtiDock >> Glide
- Uni-Mol prioritizes RMSD but fails miserably on PB-Valid



# Conclusions

- AI drug discovery techniques are here to stay.
- Pharma companies adoption increases.
- Data mining and analysis going 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)



**FOR YOUR ATTENTION**



**THANK YOU**

imgflip.com