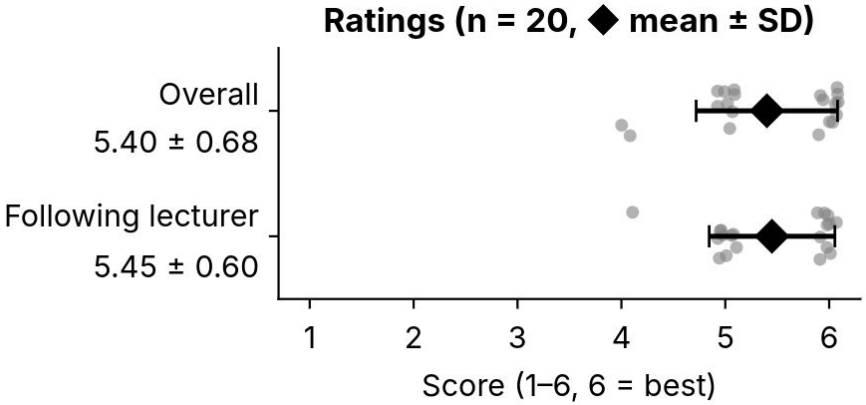
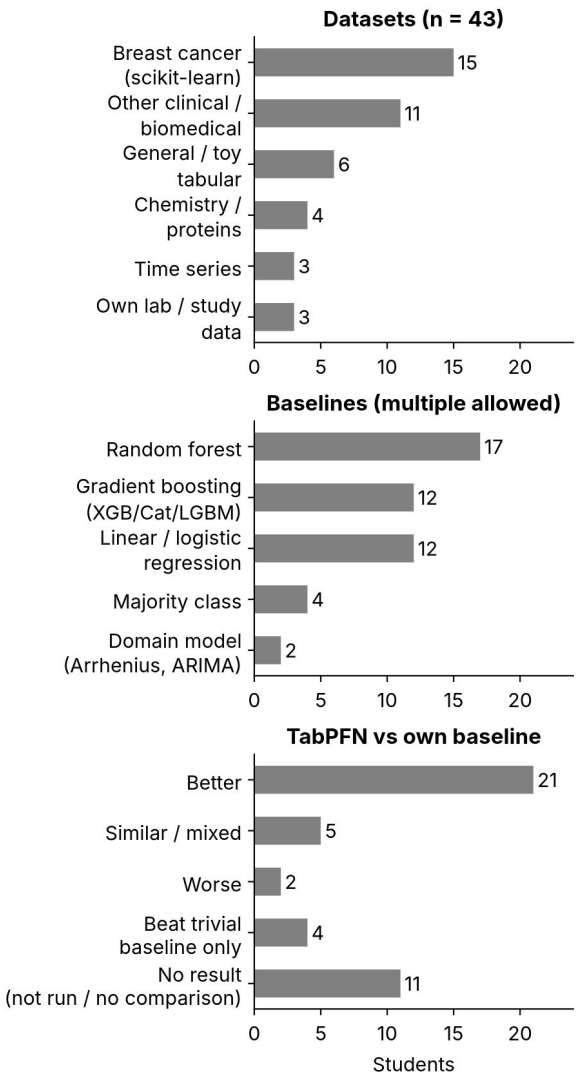
Feedback and questions from lecture 8



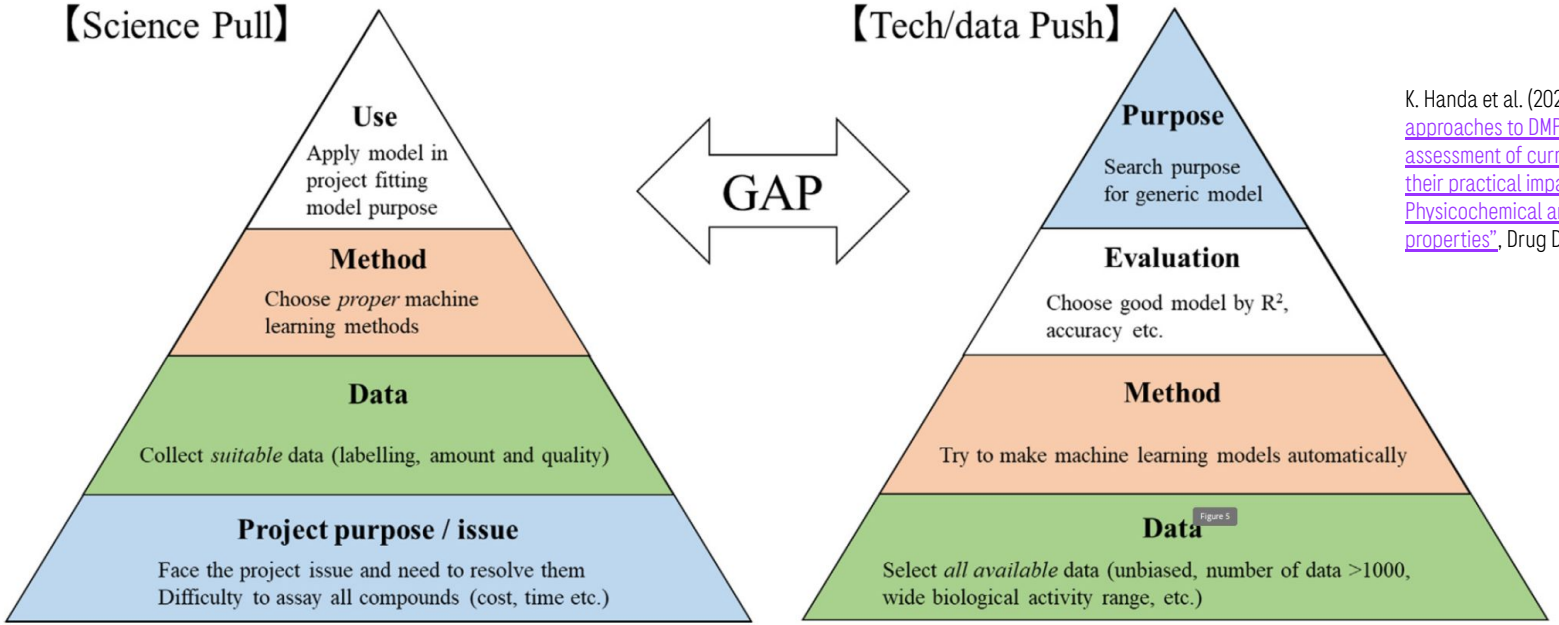
Common questions:

- How does TabPFN handle molecular fingerprints, or generally more features than samples (n<<p)?
- How reliable is TabPFN? Does its advantage become more evident as data gets scarcer?
- Is it used in industry or drug discovery?

Some couldn’t finish, had access issue, or have aversion to proprietary software. Others find the hands-on exercise useful.



AMIDD 2026 Lecture 9: AI in drug discovery



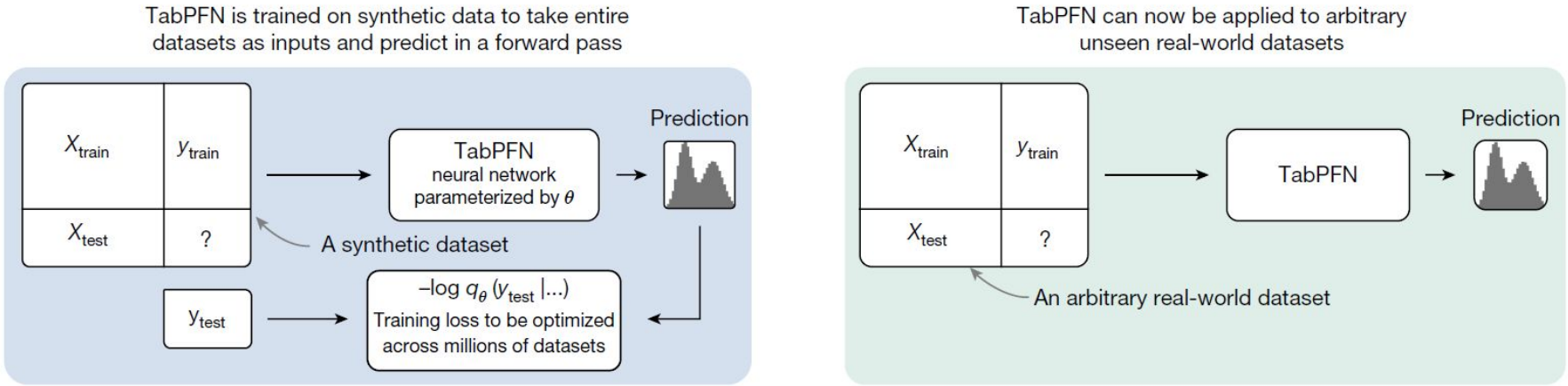
K. Handa et al. (2025), "Computational approaches to DMPK: A realistic assessment of current methods and their practical impact. Part I: Physicochemical and in vitro properties", Drug Discovery Today.

PD Dr. Jitao David Zhang, Computational Biologist

[1] Pharmaceutical Sciences, Pharma Research and Early Development, Roche Innovation Center Basel, F. Hoffmann-La Roche

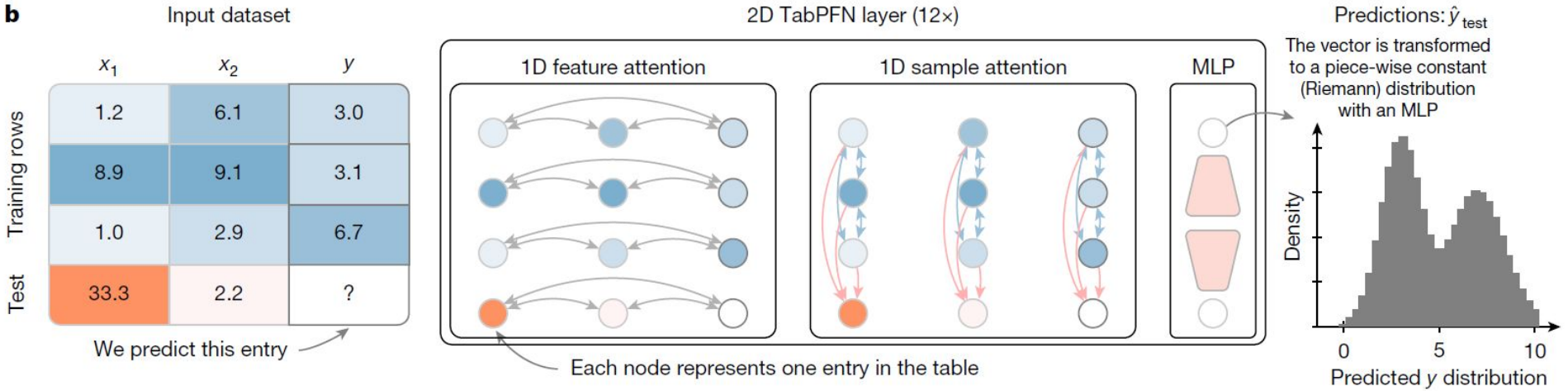
[2] Department of Mathematics and Informatics, University of Basel

Tabular Prior-data Fitted Network (TabPFN) is trained with synthesized data and predicts missing value in user data

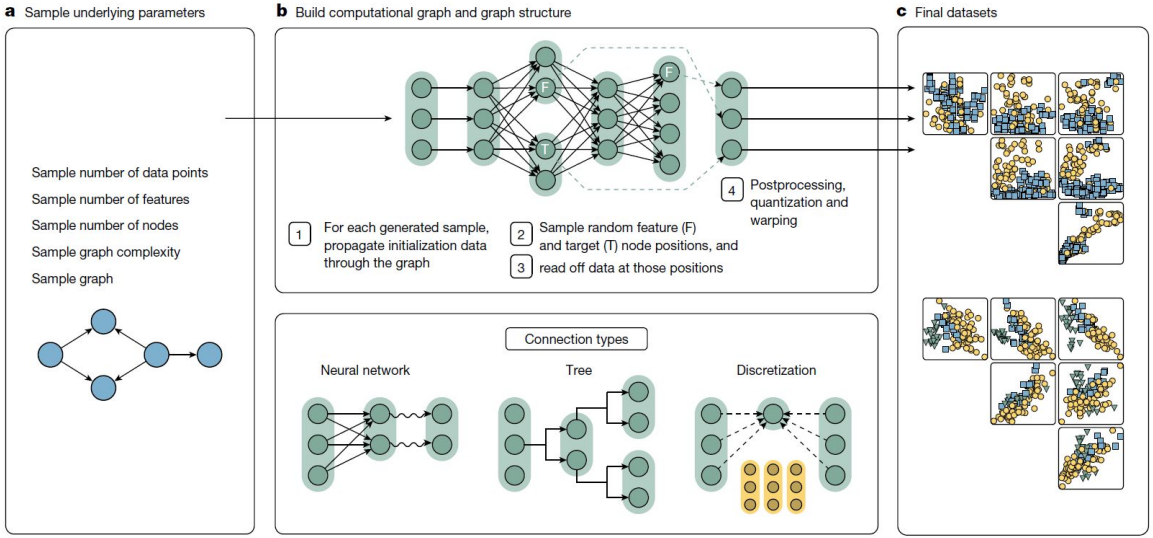


A metaphor: Imagine a lab where millions of billiard games are played simultaneously: in each game, different numbers of balls are placed randomly, and a white ball starts with a random velocity at a random position. By learning the trajectory of all balls of all games, one may learn to predict the trajectory of any real billiard game, as long as the positions of all balls and the initial velocity of the white ball is known.

Architecture of TabPFN uses both row-wise sample and column-wise feature information to predict a distribution

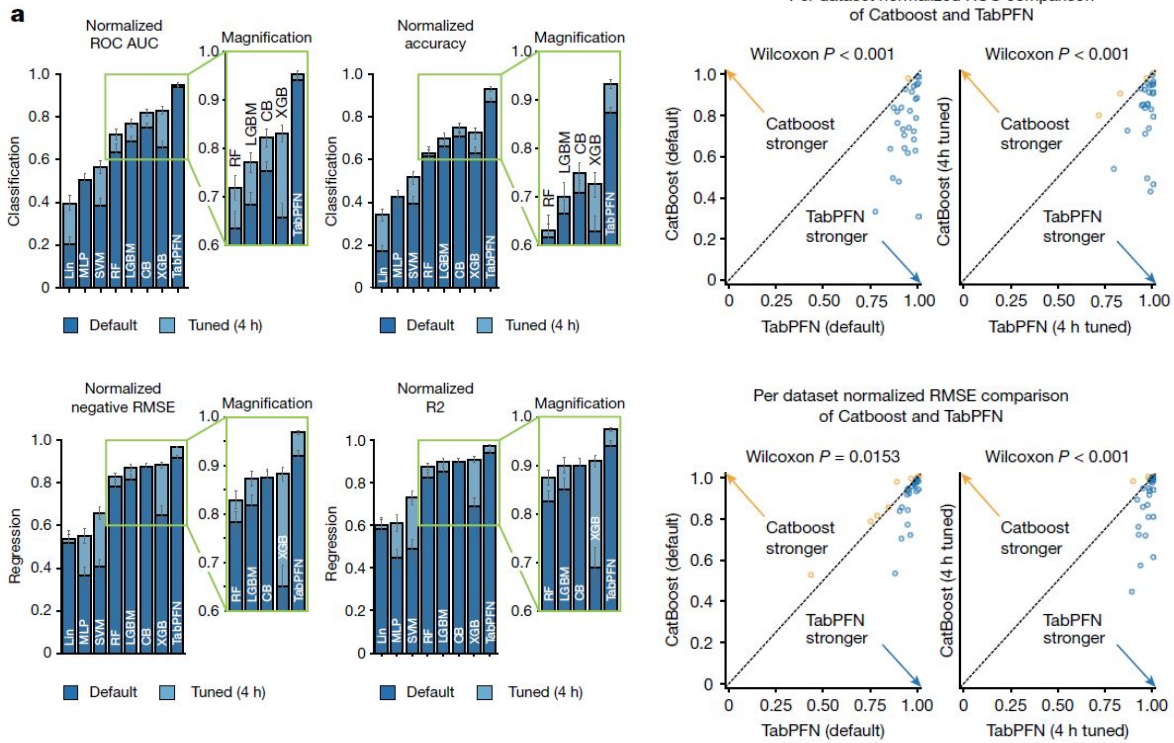


Generating synthesized data with causal models



1. We sample 100-million structurally unique directed acyclic graphs (DAGs) as causal graphs that generate data.
2. Random data is assigned to the root node. Other nodes are propagated by rules specified by the edges, plus a Gaussian noise.
3. The values of sampled features (input to the model) and targets (output of the model) are extracted.
4. Data are post-processed (non-linear distortion, binning, random missing) to reflect real-world data processing.
5. We run 100 million causal simulations to synthesize data, and use the data to train *TabPFN*.

Performance benchmark against popular methods



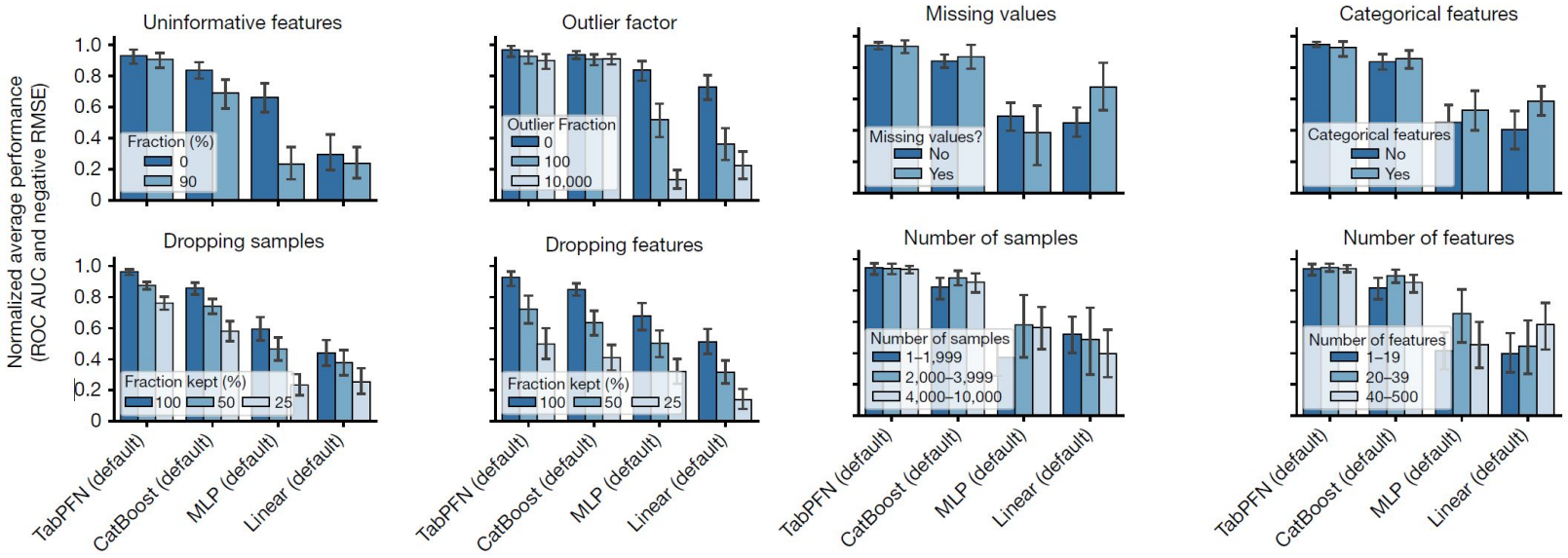
Left: Performance for classification (top) and regression (bottom) tasks.

Right: Comparing performance of TabPFN and CatBoost, the top contendant, for classification (top) and regression (bottom) tasks.

Hollmann, Noah, Samuel Müller, …, Frank Hutter. 2025. “Accurate Predictions on Small Data with a Tabular Foundation Model.” Nature 637 (8045): 319–26.

Lin: linear; MLP: multi-layer perceptrons; SVM: support vector machines; RF: random forest; LGBM: Light Gradient Boosting Machine; CB: CatBoost; XGB: XGBoost; ROC: receiver operating characteristic; RMSE: root mean square error.

Robustness against common caveats and problems



Datasets that the authors tested for classification (left) and regression (right) tasks

Extended Data Table 3 | List of test datasets used for primary evaluation of classification tasks

Name	OpenML ID	Domain	Features	Samples	Targets	Categorical Feats.
ada	41156	Census	48	4147	2	0
Australian	40981	Finance	14	690	2	8
blood-transfusion-service-center	1464	Healthcare	4	748	2	0
car	40975	Automotive	6	1728	4	6
churn	40701	Telecommunication	20	5000	2	4
cmc	23	Public Health	9	1473	3	7
credit-g	31	Finance	20	1000	2	13
dna	40670	Biology	180	3186	3	180
eucalyptus	188	Agriculture	19	736	5	5
first-order-theorem-proving	1475	Computational Logic	51	6118	6	0
GesturePhase Segmentation Processed	4538	Human-Computer Interaction	32	9873	5	0
jasmine	41143	Natural Language Processing	144	2984	2	136
kc1	1067	Software Engineering	21	2109	2	0
kr-vs-kp	3	Game Strategy	36	3196	2	36
madeline	41144	Artificial	259	3140	2	0
mfeat-factors	12	Handwriting Recognition	216	2000	10	0
ozone-level-8hr	1487	Environmental	72	2534	2	0
pc4	1049	Software Engineering	37	1458	2	0
philippine	41145	Bioinformatics	308	5832	2	0
phoneme	1489	Audio	5	5404	2	0
qsar-biodeg	1494	Environmental	41	1055	2	0
Satellite	40900	Environmental Science	36	5100	2	0
segment	40984	Computer Vision	16	2310	7	0
steel-plates-fault	40982	Industrial	27	1941	7	0
sylvine	41146	Environmental Science	20	5124	2	0
vehicle	54	Image Classification	18	846	4	0
wilt	40983	Environmental	5	4839	2	0
wine-quality-white	40498	Food and Beverage	11	4898	7	0
yeast	181	Biology	8	1484	10	0

All classification tasks from the AutoML Benchmark[36] with fewer 10,000 samples and 500 features. The benchmark comprises diverse real-world tabular datasets, curated for complexity, relevance, and domain diversity.

Extended Data Table 4 | List of test datasets used for primary evaluation of regression tasks

Name	OpenML ID	Domain	Features	Samples	Categorical Features
abalone	42726	Marine Biology	8	4177	1
airfoil_self_noise	44957	Aerospace Engineering	5	1503	0
auction_verification	44958	Economics	7	2043	2
boston	531	Real Estate	13	506	2
cars	44994	Automotive Engineering	17	804	0
colleges	42727	Education	44	7063	12
concrete_compressive_ strength	44959	Materials Science	8	1030	0
cpu_activity	44978	Computer Engineering	21	8192	0
energy_efficiency	44960	Architectural Engineering	8	768	0
geographical_origin _of_music	44965	Music Information Retrieval	116	1059	0
grid_stability	44973	Power Systems Engineering	12	10000	0
house_prices_nominal	42563	Real Estate	79	1460	43
kin8nm	44980	Robotics	8	8192	0
Mercedes_Benz_ Greener_ Manufacturing	42570	Manufacturing	376	4209	8
MIP-2016-regression	43071	Operations Research	144	1090	1
Moneyball	41021	Sports Analytics	14	1232	6
pumadyn32nh	44981	Robotics	32	8192	0
QSAR_fish_toxicity	44970	Toxicology	6	908	0
quake	550	Geophysics	3	2178	0
SAT11-HAND-runtime-regression	41980	Computational Logic	116	4440	1
sensory	546	Food Science	11	576	11
socmob	541	Sociology	5	1156	4
space_ga	507	Political Science	6	3107	0
student_performance	44967	Education	30	649	17
tecator	505	Food Science	124	240	0
topo_2_1	422	Cheminformatics	266	8885	0
us_crime	42730	Criminology	126	1994	0
yprop_4_1	416	Cheminformatics	251	8885	0

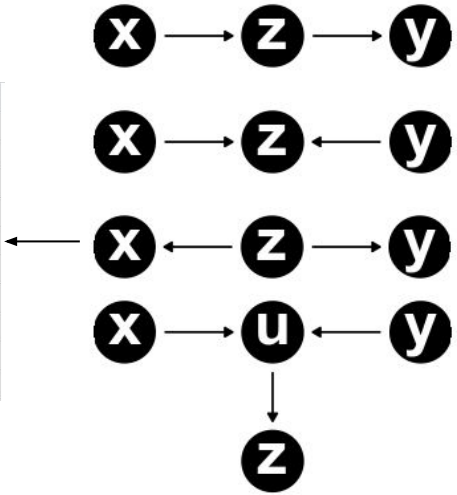
All regression tasks from the AutoML[36] and OpenML-CTR23[37] Benchmarks with fewer 10,000 samples and 500 features. The benchmark comprises diverse real-world tabular datasets, curated for complexity, relevance, and domain diversity.

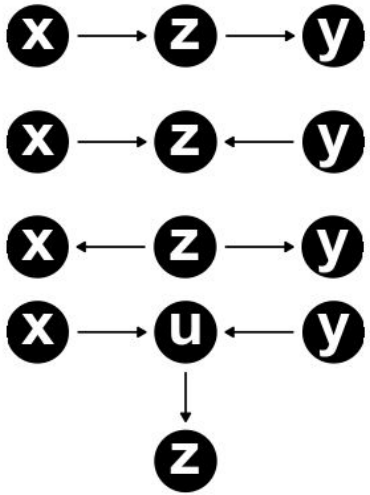
We witness a paradigm shift in analysing tabular data: from *causal model selection* to *causality-assisted prediction*

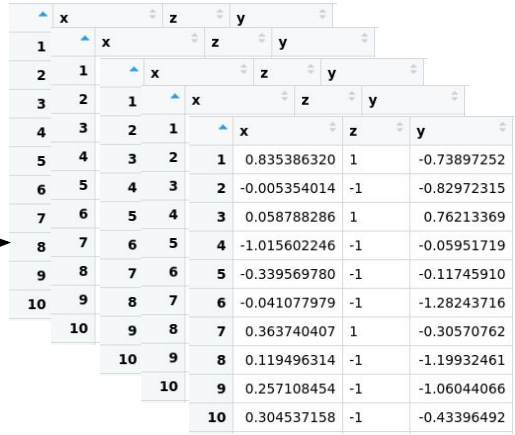
Until TabPFN: Which model(s) could have generated this piece of data? Use them for prediction.

Since then: Sample a large number of causal models to generate synthesized data, and train ML models for prediction.

	x	z	y
1	0.835386320	1	-0.73897252
2	-0.005354014	-1	-0.82972315
3	0.058788286	1	0.76213369
4	-1.015602246	-1	-0.05951719
5	-0.339569780	-1	-0.11745910
6	-0.041077979	-1	-1.28243716
7	0.363740407	1	-0.30570762
8	0.119496314	-1	-1.19932461
9	0.257108454	-1	-1.06044066
10	0.304537158	-1	-0.43396492







The difficulty of asserting causality remains. However, thanks to causal models and transformers, we now have better tools for prediction.

nature reviews drug discovery

https://doi.org/10.1038/s41573-026-01496-2

Perspective

Check for updates

Artificial intelligence in drug discovery — what it is, where we stand and the path forward

Andreas Bender[1,2,3,4], Morgan C. Thomas[1], Jack W. Scannell[5,6], David A. Shaywitz[7], Gian Marco Ghiandoni[8], Joe G. Greener[9], Lavinia-Lorena Pruteanu[1,10], Rachel DeVay Jacobson[11], Koichi Handa[12,13], Mariko Hirano[13], Srijit Seal[2,14], Manas Mahale[15], Marco F. Schmidt[16], Tim Ahfeldt[17,18], Francesca Grisoni[19] & Isidro Cortes-Ciriano[20]

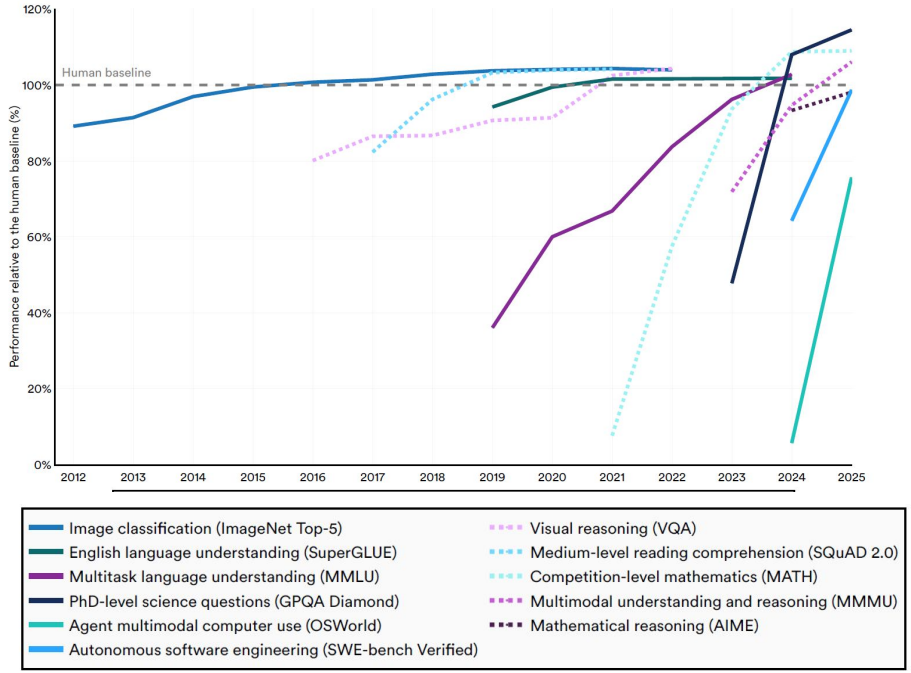
The authors argue that AI's clinical relevant impacts are disappointingly limited. They discuss potential reasons, and recommend ways to increase AI's translational relevance.

Bender, Andreas, Morgan C. Thomas, Jack W. Scannell, et al. "Artificial Intelligence in Drug Discovery — What It Is, Where We Stand and the Path Forward." Nature Reviews Drug Discovery, August 7, 2026, 1–22.

Soaring benchmarks and soaring investments form a positive feedback, self-reinforcing loop

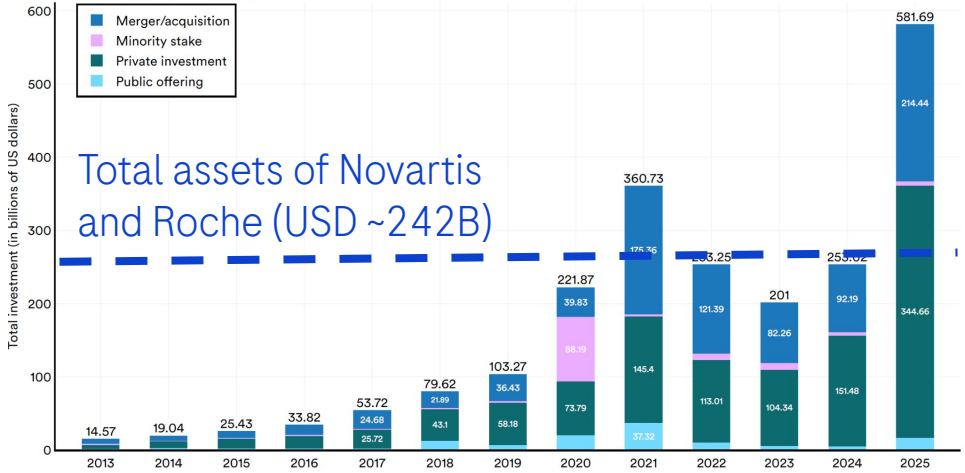
Select AI Index technical performance benchmarks vs. human performance

Source: AI Index, 2026 | Chart: 2026 AI Index report



Global corporate investment in AI by investment activity, 2013–25

Source: Quid, 2025 | Chart: 2026 AI Index report



Goldman Sachs predicts that global AI investment (augmented AI-related capital expenditure) will exceed $1 trillion in 2026 (source), roughly four times year-end 2025 total assets of Novartis and Roche combined, after currency conversion.

The 2026 AI Index Report, Stanford University Human-Centered Artificial Intelligence (HAI)

The linear model of drug discovery and development



Target identification, validation & assessment

	Target-to-hit	Hit-to-lead	Lead optimization	Preclinical	Phase I	Phase II	Phase III	Submission to launch	
p(TS)	80%	75%	85%	69%	54%	34%	70%	91%	Total≅4%
***WIP* needed for 1 launch**	24.3	19.4	14.6	12.4	8.6	4.6	1.6	1.1	1
Cost per WIP per Phase	$1	$2.5	$10	$5	$15	$40	$150	$40	
Cycle time (years)	1.0	1.5	2.0	1.0	1.5	2.5	2.5	1.5	Total≅13.5y
Cost per launch (out of pocket)	$24	$49	$146	$62	$128	$185	$235	$44	$873
% Total cost per NME	3%	6%	17%	7%	15%	21%	27%	5%	
Cost of capital	11%								
Cost per launch (capitalized)	$94	$166	$414	$150	$273	$319	$314	$48	$1,778

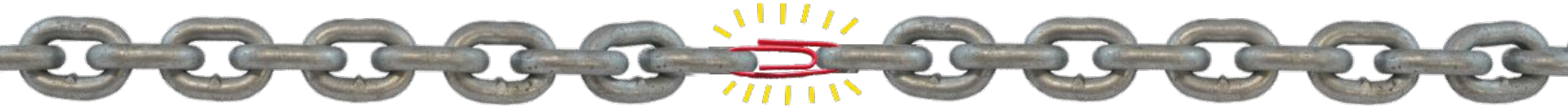
Launch

Discovery Development

Adapted from Paul, Steven M. *et al.* "How to Improve R&D Productivity: The Pharmaceutical Industry's Grand Challenge." Nature Reviews Drug Discovery, 2010. Costs are in 2008 US$ millions.

Every system has a limiting factor or constraint

A chain is only as strong as its weakest link. Similarly, every system must have a constraint that limits its output. In drug discovery, the biggest constraint is failures in Phase II trials, often due to lack of efficacy.

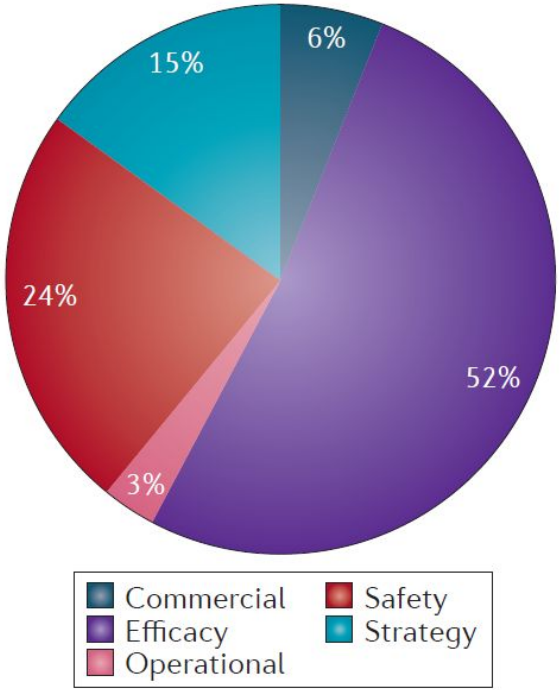


At a given point in time, a system often has a dominant bottleneck. Strengthening the other links does not increase the overall strength: it may even *decrease* the overall strength, for example due to added weight.

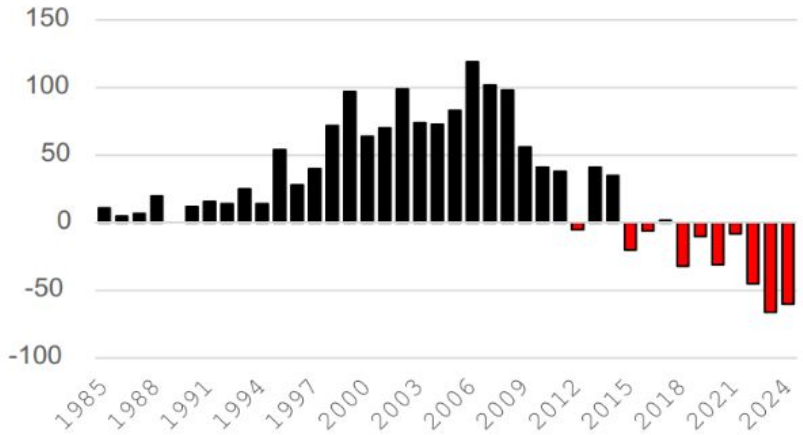


Courtesy of Eli Goldratt, the author of *The Goal*, and Theory of Constraints Institute, https://www.tocinstitute.org/theory-of-constraints.html

Efficacy drives most failures while we run out of targets



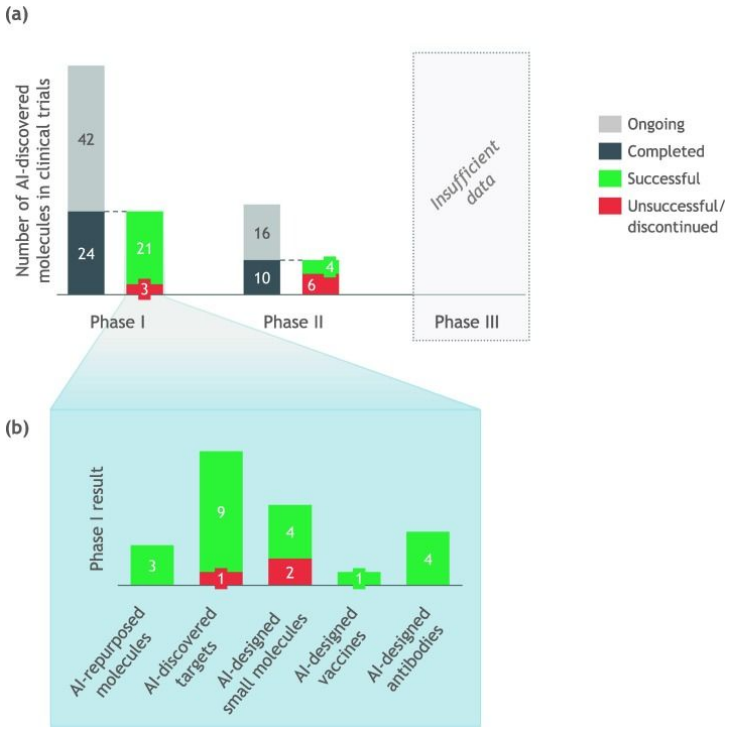
Net target-indication pairs tested in clinics



The *target paradox*
more research, fewer targets

Harrison, Richard K. "Phase II and Phase III Failures: 2013–2015." Nature Reviews Drug Discovery, 2016
Kinch, Michael S., Zachary Kraft, and Tyler Schwartz "Target & Mechanism Selection in Pharmaceutical Innovation: A Source for Concern or Optimism?" Drug Discovery Today, 2025

AI's clinical impact: emerging signals, but evidence and attribution remain limited



- *Absence of evidence,* not *evidence of absence*
- Disease relevance was demonstrated for most targets targeted by drug candidates labelled with ‘AI’, though the public seldomly know which AI methods were used or how they contributed.

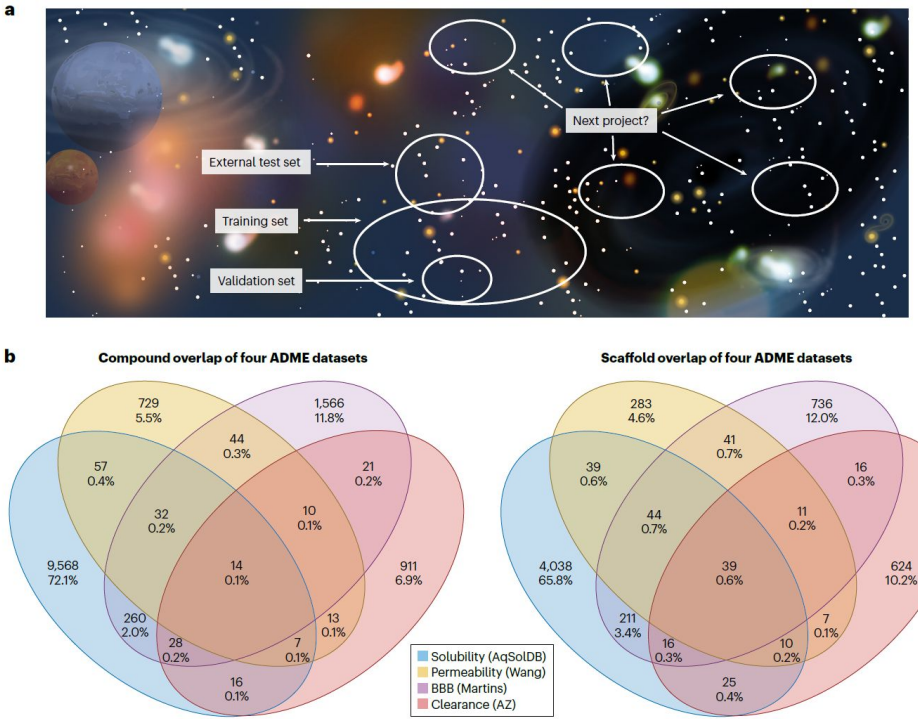
ATH-63, a "first-in-class, oral, small molecule, genomic regulator"
BEN-8744, a PDE10 inhibitor
BDTX-1535, an oral EGFR inhibitor
BDTX-4933, targeting various mutant forms of BRAF
CEL-383, an anti-TREM1 antibody for IBD
Bapotulimab, an anti-ILDR2 antibody for various tumor types
HFB301001, an anti-OX40 (CD134) antibody
HFB200301, an anti-TNFR2 antibody
HFB200603, an anti-BTLA monoclonal antibody
HST-1011, a CBL-B inhibitor
HMBD-001, an anti-HER3 antibody
HMBD-002, an anti-VISTA antibody
INS018-055, a small molecule TNIK inhibitor (Phase III 2026)
Omilancor, a LANCL2 activator
NX-13, an NLRX1 agonist
NIM-1324, a LANCL2 activator
NEU411, an LRRK2 inhibitor
REC-994, a superoxide scavenger
REC-3964, a *C. difficile* toxin B inhibitor
VRG-50635, a PIKfyve inhibitor

M. KP Jayatunga et al. (2024), “How successful are AI-discovered drugs in clinical trials? A first analysis and emerging lessons,” , https://doi.org/10.1016/j.drudis.2024.104009, Drug Discovery Today. Note: the authors are all employees of Boston Consulting Group.

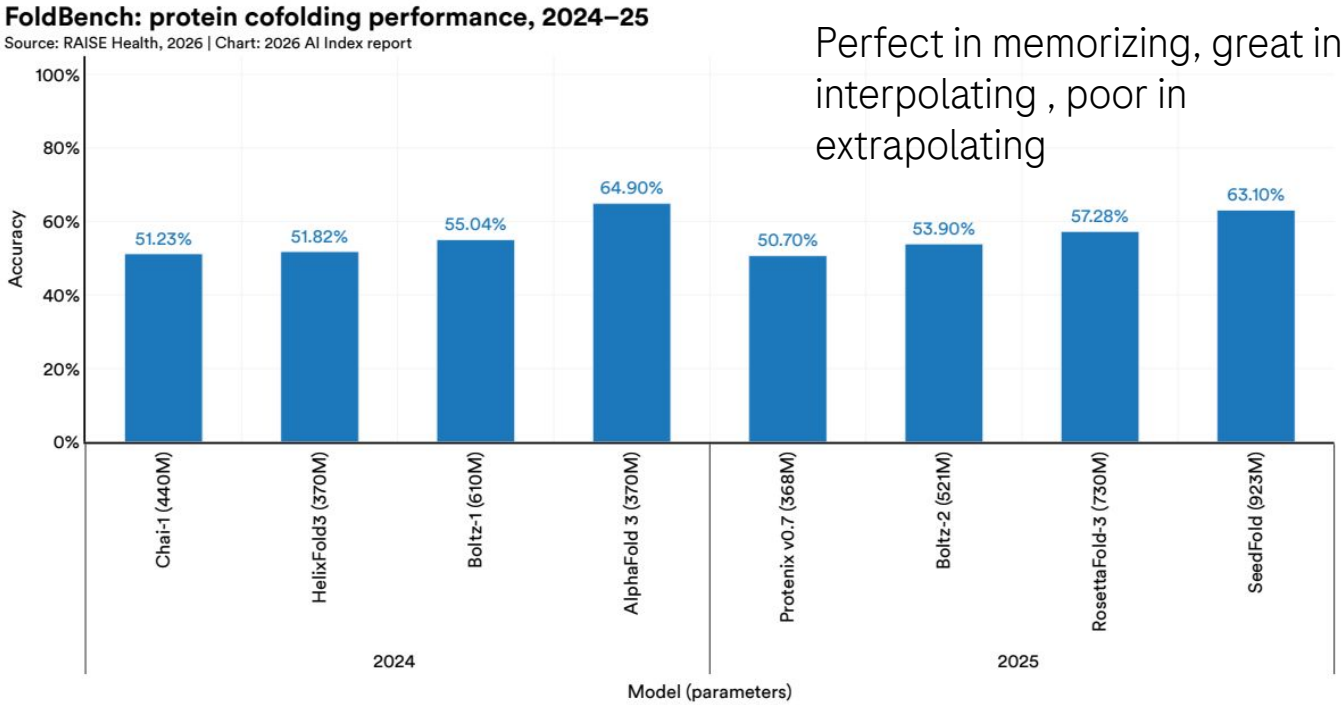
Fact check: Derek Lowe, *In the Pipeline,* https://www.science.org/content/blog-post/ai-drugs-so-far, information on INS018-055 (rentosertib) was added.

Challenges and opportunities of applying ML/AI to chemical data

1. The chemical space is huge (10^{60} drug-like molecules, top right)
2. Test-set performance can overstate future performance, particularly under distribution shift.
3. Models trained for different properties cover different chemical spaces (bottom right)
4. Artefacts and bias exist due to causal structures: Why presence of sugar rings predicts bitterness of compounds?

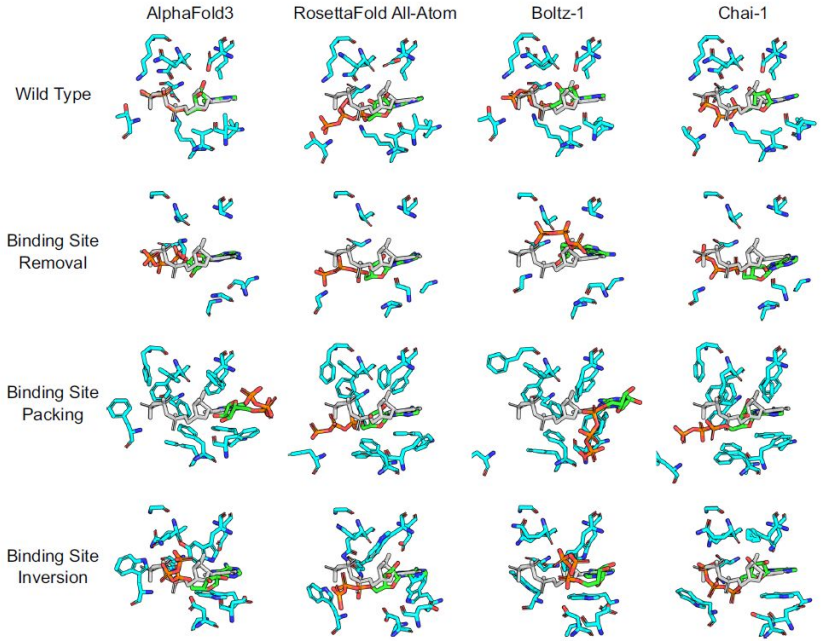


Protein-ligand cofolding has improved rapidly, but robust translation into drug-discovery decisions remains limited.



Perfect in memorizing, great in interpolating , poor in extrapolating

Deep learning models for co-folding does not always learn the physics of protein-ligand interactions



M. R. Masters, A. H. Mahmoud and M. A. Lill. (2025), “Investigating whether deep learning models for co-folding learn the physics of protein-ligand interactions,” , https://doi.org/10.1038/s41467-025-63947-5, Nature Communications

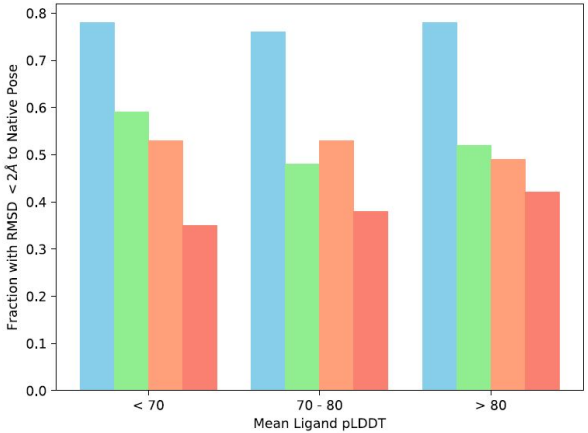
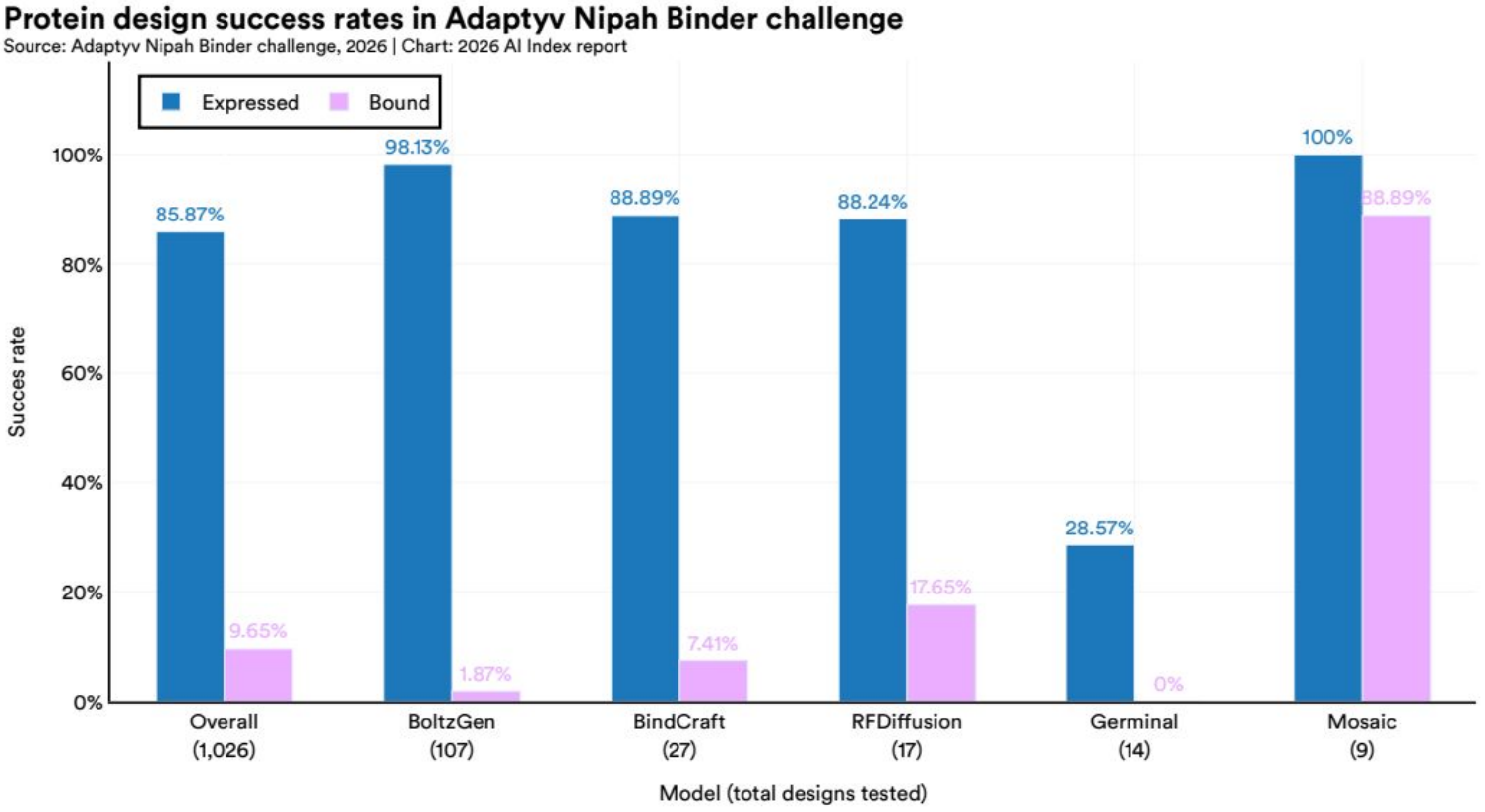


Fig. 3 | Systematic binding site mutation challenges performed across the CASF-2016 protein-ligand dataset (n=285 complexes). Bar represents fraction of systems with RMSD < 2 Å to the native pose for the unmutated wild-type (blue), binding site removal (green), binding site packing (orange), and binding site inversion (red) challenges.

Our findings underscore the models’ limitations in generalizing effectively across diverse protein-ligand structures… The results advocate a measured reliance on deep-learning-based models for critical applications in drug discovery and protein engineering, where a deep understanding of the underlying physical and chemical properties is crucial.

Protein and peptide design has progressed impressively: wet-lab validation remains essential

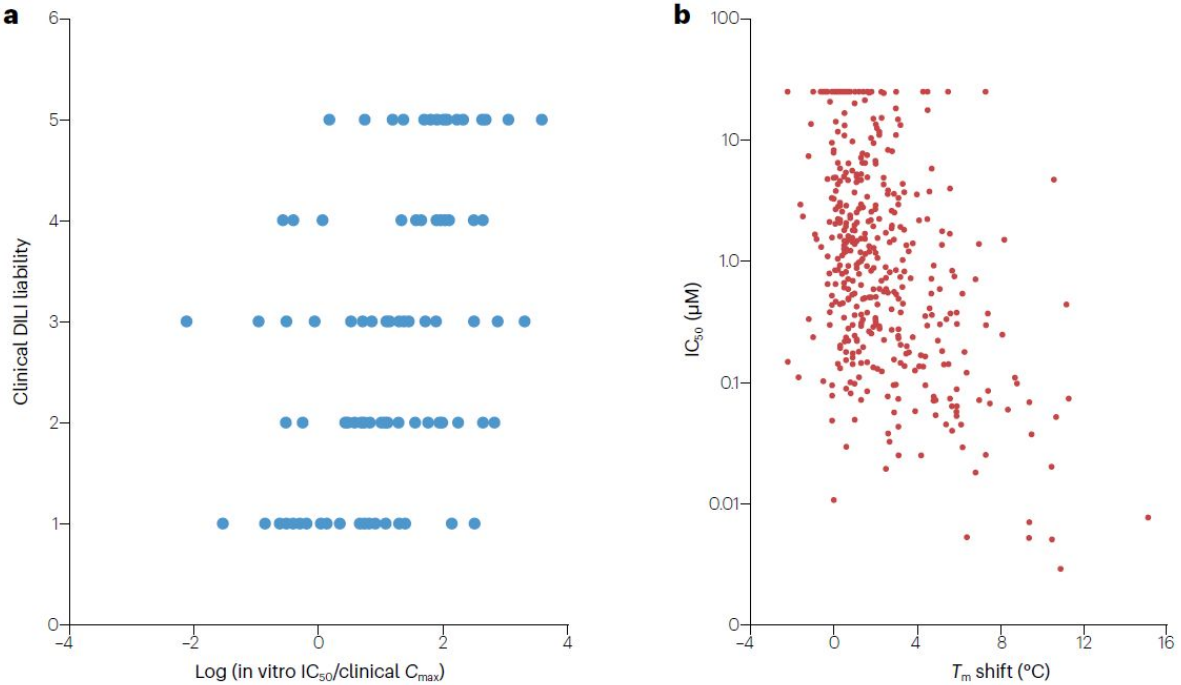


Proxy assays may not predict things that really matter

Left: an organoid assay for drug-induced liver injury is associated with actual clinical liver injury, but the correlation can be, depending on the context and use case, too weak to be useful.

Right: Thermal-shift measurements of ligand-target pairs correlate only weakly with the enzymatic activity they are intended to approximate.

A model cannot recover information that is absent from, or only weakly related to, its endpoint. AI cannot compensate for an endpoint that itself has weak predictive validity.



Conditionality presents fundamental challenges for biological data

Measured biological phenotypes at macroscopic, microscopic and molecular scales can change substantially with changes in many factors, including but not limited to

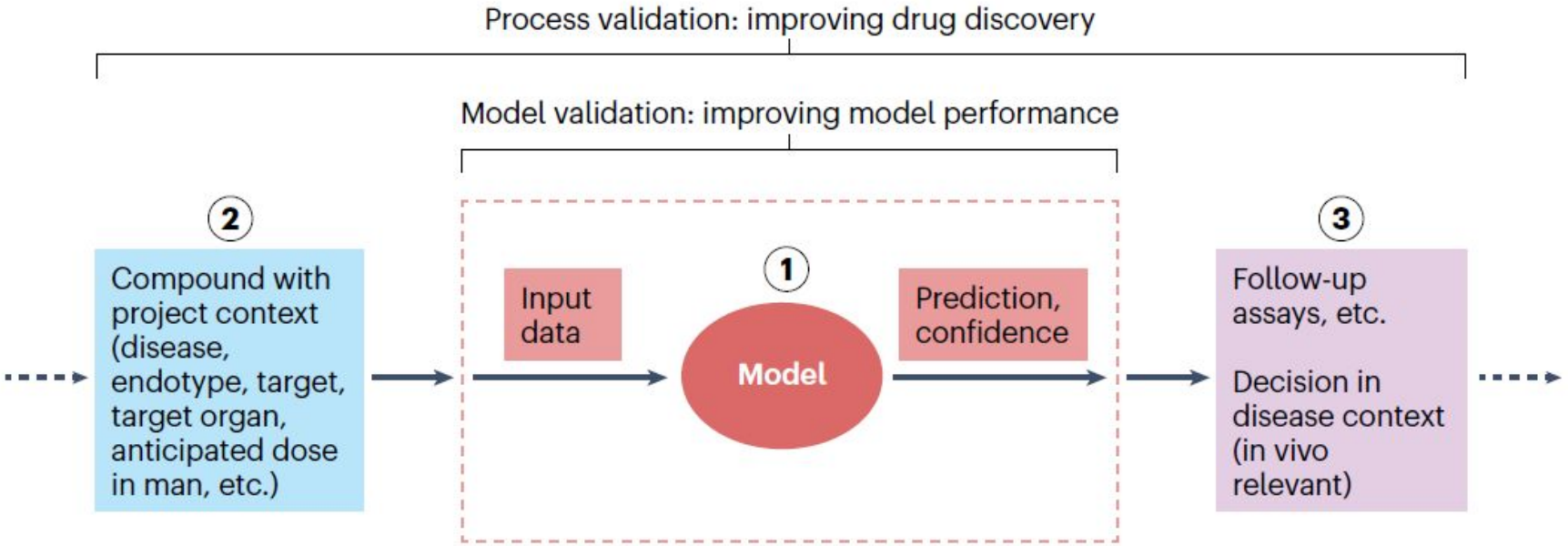
- Genotype
- Lifestyle
- Metabolism
- Microbiome
- Cell type and in case of culture, passage number
- Experimental conditions
- For diseases: disease subtype and stage
- For treatments: dose and treatment duration
- Time point
- Variance across devices, assays, or even consumables batches.

Consequences:

- The same perturbation can produce different outcomes depending on the biological system
- Labels such as *active/inactive*, *toxic/non-toxic*, *responder/non-responder* are often arbitrary, often ignore unrecorded conditions, and can be misleading.

All models are wrong, some are useful … Which ones then?

Shifting from *model validation* to *process validation*



Model validation: Does the model predict a held-out dataset accurately?

Process validation: Does using this model improve the actual drug-discovery decision?

Endpoint prevalence and downstream activities dictate desired assay performance: a case study with teratogenicity

	Teratogen	Non-teratogen	Sum
Predicted teratogenic	19	20	39
Predicted non-teratogenic	1	1980	1981
Sum	20	2000	2020

We take a hypothetical example of an assay with 95% sensitivity (True Positive Rate, TP/(TP+FN)) and 99% specificity (True Negative Rate, TN/(TN+FP)) (which is extremely good!).

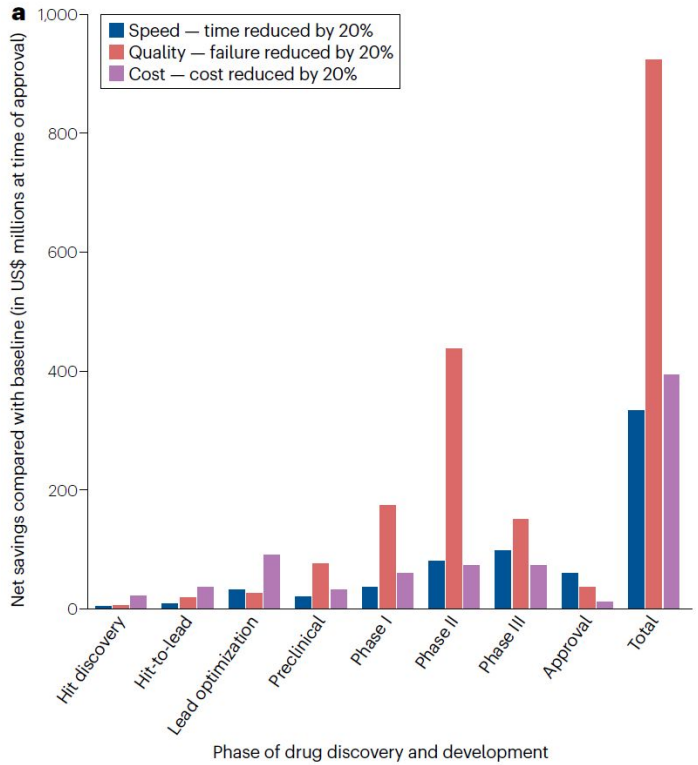
Given the imbalanced cases of teratogens *vs* non-teratogens (which is partly caused by the funneling nature of drug discovery), an ideal screening assay needs to have extremely high **specificity**, because otherwise many predicted teratogens are actually not, discarding potentially safe compounds or subjecting them to expensive animal tests. Even if so, the negative predictions are more valuable than positive ones because of the high false-positive rate.

I thank Giuliana Rossi, Manuela Jaklin, Mehmet Girgin, and Stefan Kustermann for discussions and input..

Higher quality around Phase I/II would save more than faster delivery

The largest savings would come from increased success rates of Phase I (safety, PK) and Phase II (dose finding, efficacy, and patient stratification) trials. Possibilities include:

1. Develop new MoAs and modalities for validated targets (degraders, molecular glues, macrocyclic peptides, etc.)
2. Identify better targets by leveraging genetics, functional genomics.
3. Phenotypic drug discovery in disease-relevant models.
4. Combination modalities and therapies.
5. Improve physicochemical properties, ADME/PK/drug-drug interaction (DDI) predictions and PBPK modeling capabilities.
6. Prioritize trial sites and accelerate patient enrollment.



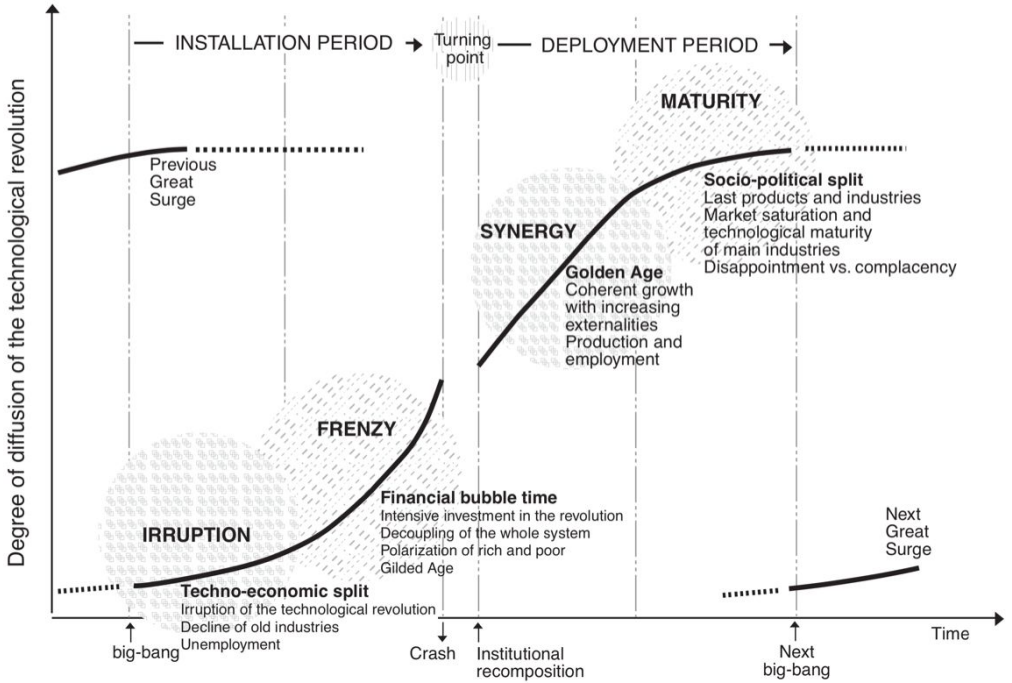
C. Weinmann et al. (2026), “DynlogP - A Novel High-Throughput LC−MS Assay Quantifying Dynamic Lipophilicity and Chameleonicity of bRo5 Macrocyclic Peptides,” , https://doi.org/10.1021/acs.molpharmaceut.6c00409, Molecular Pharmaceutics; S. Käser et al. (2026), “Predicting First-in-Human Pharmacokinetics: Comparative Evaluation of Standard PBPK, High-Throughput PBPK, and Machine Learning,” , https://doi.org/10.1021/acs.molpharmaceut.6c00429, Molecular Pharmaceutics

Technological revolutions and financial capital

In her book *Technological revolutions and financial capital*, Carlota Perez analyzed five technological revolutions from the Industrial Revolution (starting in 1771, Great Britain) to the Age of Information and Telecommunications (starting in 1971, USA), and described the connections between technological development and financial bubbles.

I encourage everyone make their own judgement of where we are now with AI. I expect that the technology to continue developing. We have the responsibility to develop mature and responsible applications for drug discovery.

Figure 5.1 Recurring phases of each great surge in the core countries



Six principles for evaluating capabilities of AI models for drug discovery, adapted from Melanie Mitchell (2026)

1. Be aware of your own cognitive biases and incentives;
2. Be sceptical of your own hypotheses and those of others: design control experiments to rule out alternative explanations;
3. Design novel variations of benchmarks to test robustness and generalization;
4. Be curious about mechanisms underlying performance;
5. Consider performance versus competence (e.g. getting correct answers through brittle shortcuts that fail to generalize);
6. Analyze failure types, and embrace ‘negative’ results.

M. Mitchell. (2026), “Six principles for evaluating cognitive capabilities in AI models,” , https://doi.org/10.1002/aaai.70061, AI Magazine.

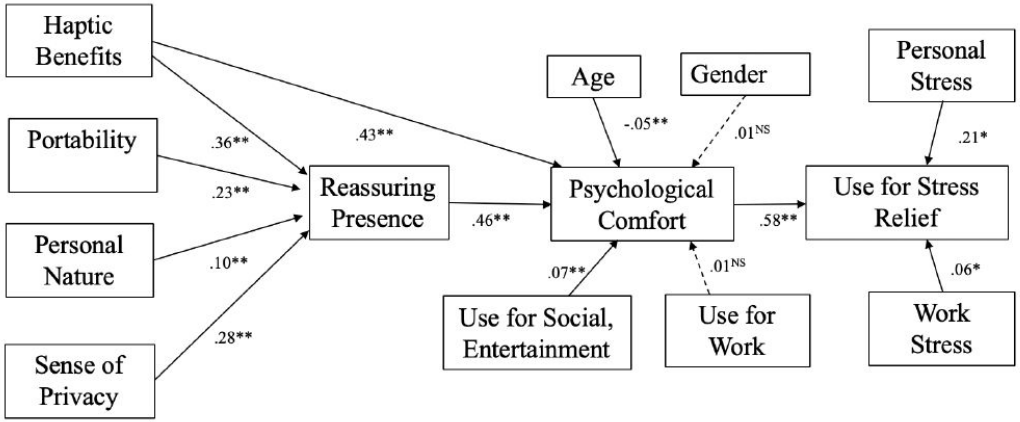
The Smartphone as a Pacifying Technology

We are part of the forces deciding how we shall use technologies

Melumad and Pham proposed that smartphones are not unlike “adult pacifiers”: users derive emotional benefits from their smartphone, including feelings of psychological comfort and stress relief.

Performing a field study and controlled experiments, they observed that

- In moments of stress, consumers exhibit a greater tendency to seek out their smartphone;
- Engaging with one’s smartphone provides greater stress relief than engaging in the same activity with own laptops or a similar smartphone belonging to someone else.



The example is not about AI. Technologies shape capability as well as behaviour. We have the technology; we also have the choice how to use it.

S. Melumad and M. T. Pham. (2020), “The Smartphone as a Pacifying Technology,” , https://doi.org/10.1093/jcr/ucaa005, Journal of Consumer Research.

Conclusions

- Causal models encode our knowledge and assumptions in statistical models: we can apply them to data to confirm or challenge our beliefs.
- AI can offer unprecedented opportunities to work with huge amount of data with diverse approaches in relatively short time (though not without issues and costs).
- It is a challenge and an opportunity to combine causal models and AI with experiments and observations to deepen our understanding of disease and drug-body interactions.

Backup slides

A few fundamental challenges of drug discovery

1. (The System) We cannot experiment freely or comprehensively with the system in which the drugs finally work, i.e. the human body, and no surrogate system reproduces human biology exactly.
2. (The Retrospective) It is difficult to identify the precise reasons why a drug discovery project is successful or unsuccessful, because decisions are interdependent and made over a long period of time.
3. (The Unquantifiable) We can not fully quantify the quality of our work or the risk of failure, particularly in the presence of *unknown unknowns*.

Tabular Prior-data Fitted Network (TabPFN)

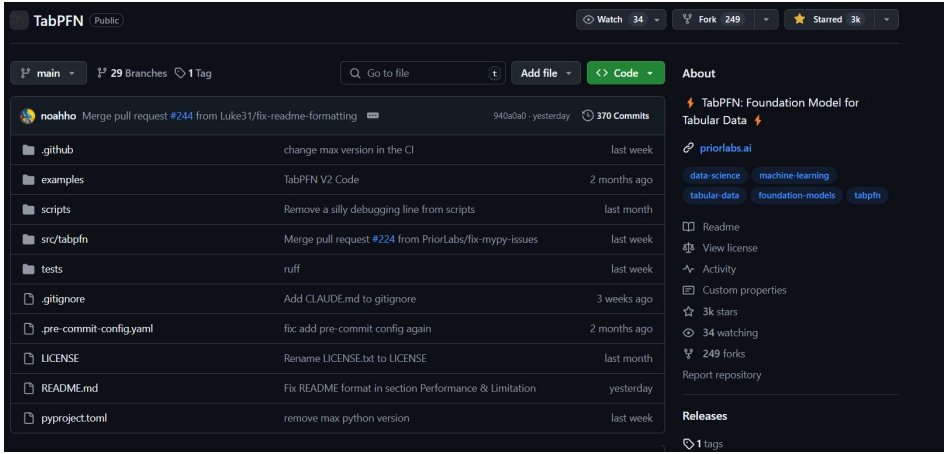
Article

Accurate predictions on small data with a tabular foundation model

https://doi.org/10.1038/s41586-024-08328-6
Received: 17 May 2024
Accepted: 31 October 2024
Published online: 8 January 2025
Open access
Check for updates

Noah Hollmann[1,2,3,7], Samuel Müller[1,7], Lennart Purucker[1], Arjun Krishnakumar[1], Max Körfer[1], Shi Bin Hoo[1], Robin Tibor Schirrmeister[4,5] & Frank Hutter[1,3,6]

Tabular data, spreadsheets organized in rows and columns, are ubiquitous across scientific fields, from biomedicine to particle physics to economics and climate science[1,2]. The fundamental prediction task of filling in missing values of a label column based on the rest of the columns is essential for various applications as diverse as biomedical risk models, drug discovery and materials science. Although deep learning has revolutionized learning from raw data and led to numerous high-profile success stories[3–5], gradient-boosted decision trees[6–9] have dominated tabular data for the past 20 years. Here we present the Tabular Prior-data Fitted Network (TabPFN), a tabular foundation model that outperforms all previous methods on datasets with up to 10,000 samples by a wide margin, using substantially less training time. In 2.8 s, TabPFN outperforms an ensemble of the strongest baselines tuned for 4 h in a classification setting. As a generative transformer-based foundation model, this model also allows fine-tuning, data generation, density estimation and learning reusable embeddings. TabPFN is a learning algorithm that is itself learned across millions of synthetic datasets, demonstrating the power of this approach for algorithm development. By improving modelling abilities across diverse fields, TabPFN has the potential to accelerate scientific discovery and enhance important decision-making in various domains.



Source code at https://github.com/PriorLabs/tabpfn, released in a license analogous to Apache 2.0 (commercial use friendly).

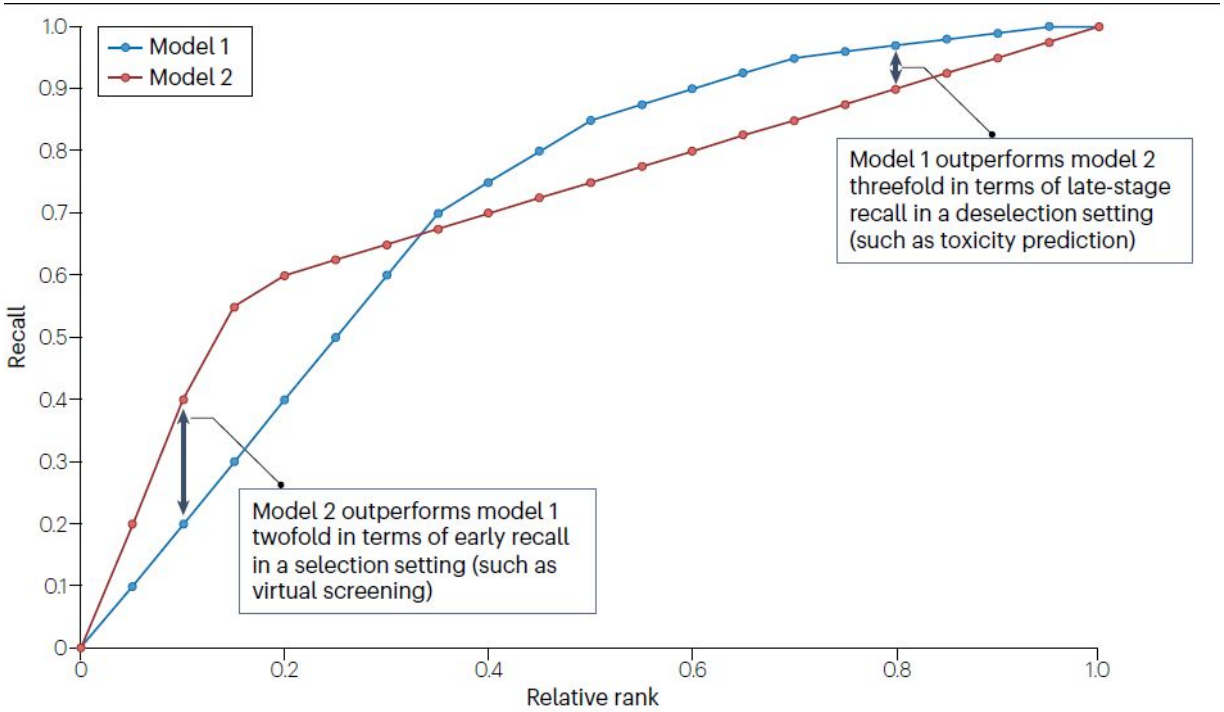
No *one size fits all*: the context and the downstream costs determine which models works best

Different decisions need different metrics:

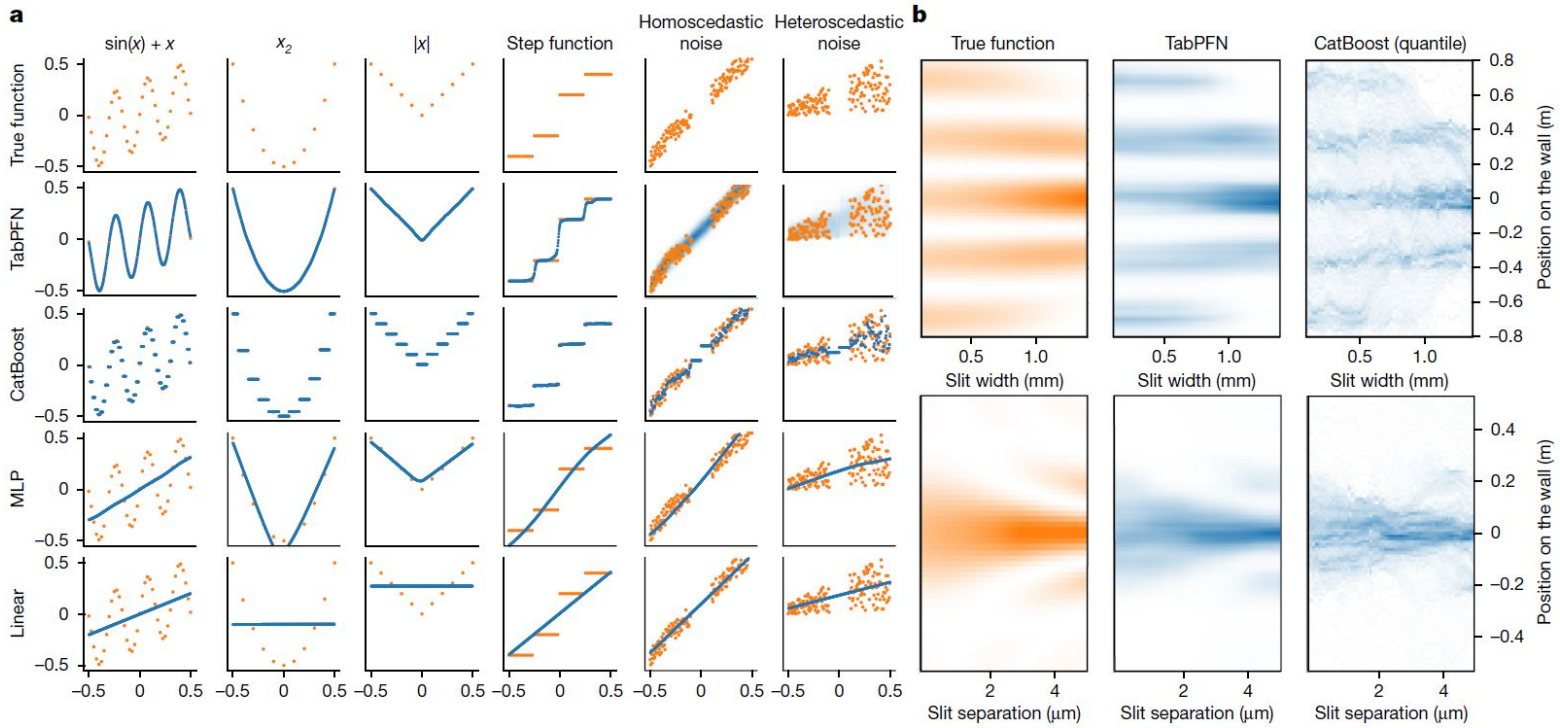
Selection: identify a few molecules from a large number of candidates (i.e. virtual screening). Precision (true positive of all predicted positive) matters.

Deselection: eliminate undesired compounds. Recall (low false negative rate) matters.

Quantification, *e.g.* estimate dose or PK. Actual numerical error matters.



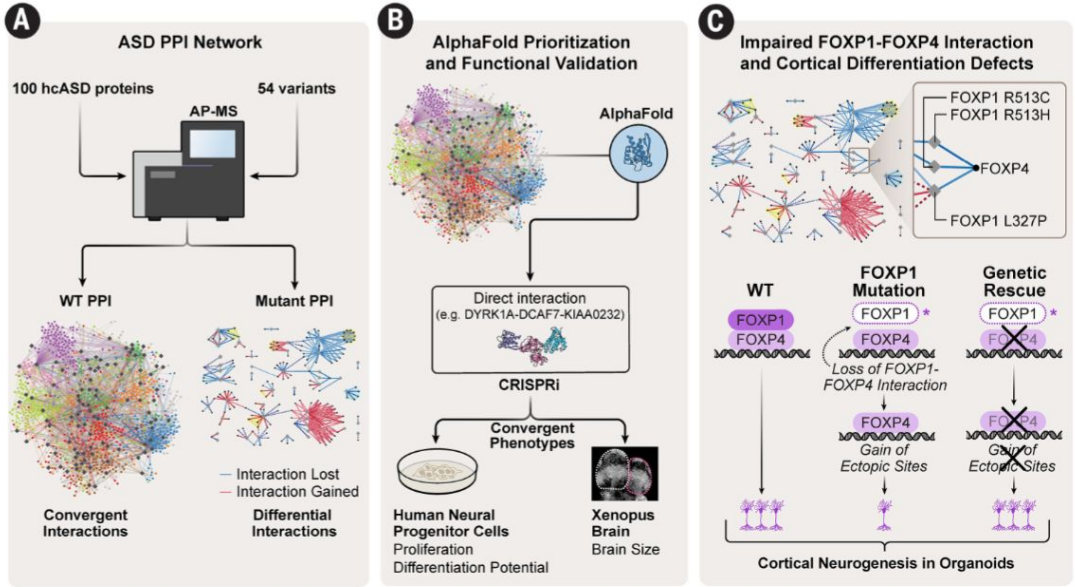
TabPFN's prediction on data generated by simple functions



Biology as multilayer and dynamic networks: high-quality data in relevant systems are often absent and hard to get

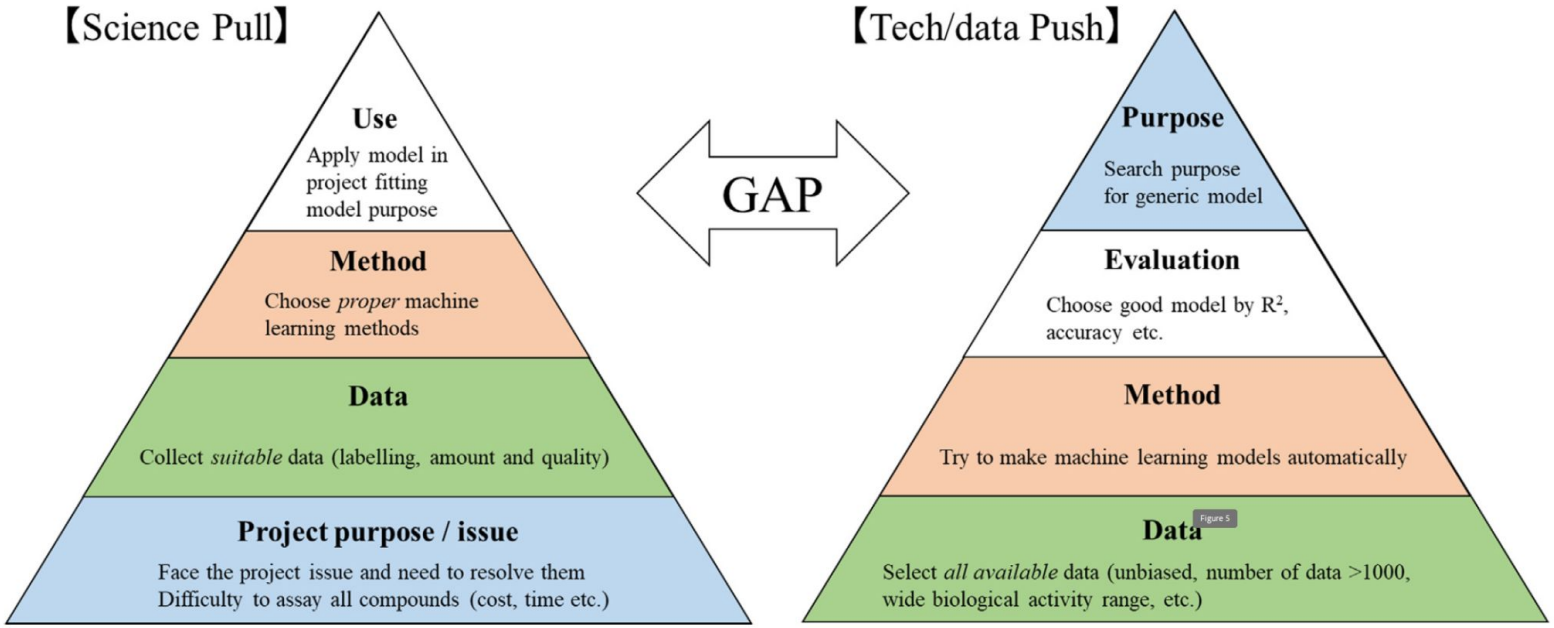
Besides sequences and abundances of biological molecules (RNA, protein, metabolites, *etc.*), their dynamics, spatial distribution, and interactions all determine the biological function of the system. For instance, in a recent study, Wang *et al.* reported that autism mutations rewire protein interaction networks to drive neurodevelopmental pathology.

B. Wang et al. (2026), "Autism mutations rewire protein interaction networks to drive neurodevelopmental pathology," , https://doi.org/10.1126/science.ady4523, Science



Often we do not know what data are relevant for disease treatment *a priori*. Even if we know that, it is challenging to get them: for instance, there can be about 10^6 protein-protein interactions, many of which are too transient and weak to be measured, yet many are critical for cellular functions. **A potential benefit of AI is to develop new measurement technologies/tools that better characterize biological systems.**

Science pull *versus* tech/data push



K. Handa et al. (2025), "Computational approaches to DMPK: A realistic assessment of current methods and their practical impact. Part I: Physicochemical and in vitro properties," , https://doi.org/10.1016/j.drudis.2025.104422, Drug Discovery Today.

#	The challenge	Implications
1.	**People tend to apply AI to problems where which data already exist, rather than what matters**	*Far better an approximate answer to the right question… than an exact answer to the wrong question* (J. Tukey) Start with clinically-relevant, project-driven questions: what prevents us from discovering safe and efficacious drugs, and testing them in relevant patients?
2.	**Biological data are conditional, and relevant data can be difficult to capture**	Accept the conditionality and make not only the model, but also the process robust with QC and curation.
3.	**Proxy measurements lack predictive validity**	Weak proxy measurements are dangerous for critical questions, and inform users and decision makers about limitations.
4.	Chemical space is enormous, and models are almost always local	Perform rigorous prospective studies, quantify prediction uncertainty, and continuously update the model with newly generated data
5.	Data are often extremely sparse, high-dimensional, and biased	*The first principle is that you must not fool yourself - and you are the easiest person to fool.* (R. Feynman)
6.	**Metrics are context-dependent, and benchmarks do not reflect translational and clinical utility**	Choose appropriate metrics and prioritize models based on contexts, and instead of focusing on improving SOTA, ask how the models help project teams and clinical outcome.
7.	Preclinical and clinical learnings are often disconnected	Perform rigorous QC and critical assessment of clinical data, and perform targeted, registered analysis. The outcome, positive or negative, should be reported to inform decisions.
8.	**Model validation is not process validation**	Consider AI model in the context of project need and downstream activities, and focus on using models to inform decisions given uncertainties and risks.
9.	Data quantity and algorithm complexity is often mistaken for information	Prefer high-quality data and parsimonious models to more data and fancier models.
10.	**AI is pushed by technology and incentives**	Adopt a ‘science-pull’ model besides ‘technology/data-push’ endeavors

Promises and risks of AI in drug discovery

- **Capability gain**: AI as a tool and sparring partner to spark critical thinking.
- **Productivity gain**: Navigating *known unknowns,* and surfacing *unknown unknowns*.
- **Social gain**: Enhancing human connection and mutual support.

- **Cognitive atrophy**: Knowledge debt and an illusion of progress if we give up critical thinking.
- **Productivity waste**: Reinforced biases in *known unknowns,* ignored *unknown unknowns*.
- **Social loss**: Concentration of power, loss of pluralism, erosion of trust.

We have the choice to focus on the process rather than only the outcome, keep researching and share what we learn, and prioritise interpersonal relationships.

Interesting ongoing initiatives

- OpenADMET, open data, open models, and community blind challenges for predicting drug absorption, distribution, metabolism, excretion, and toxicity.
- The *OpenBind* consortium, see an interview with Charlotte Deane, appeared in *Nature Reviews Drug Discovery,* Filling the massive AI drug discovery data gap
- Model Validation Protocols for Machine Learning in Small Molecule Drug Discovery