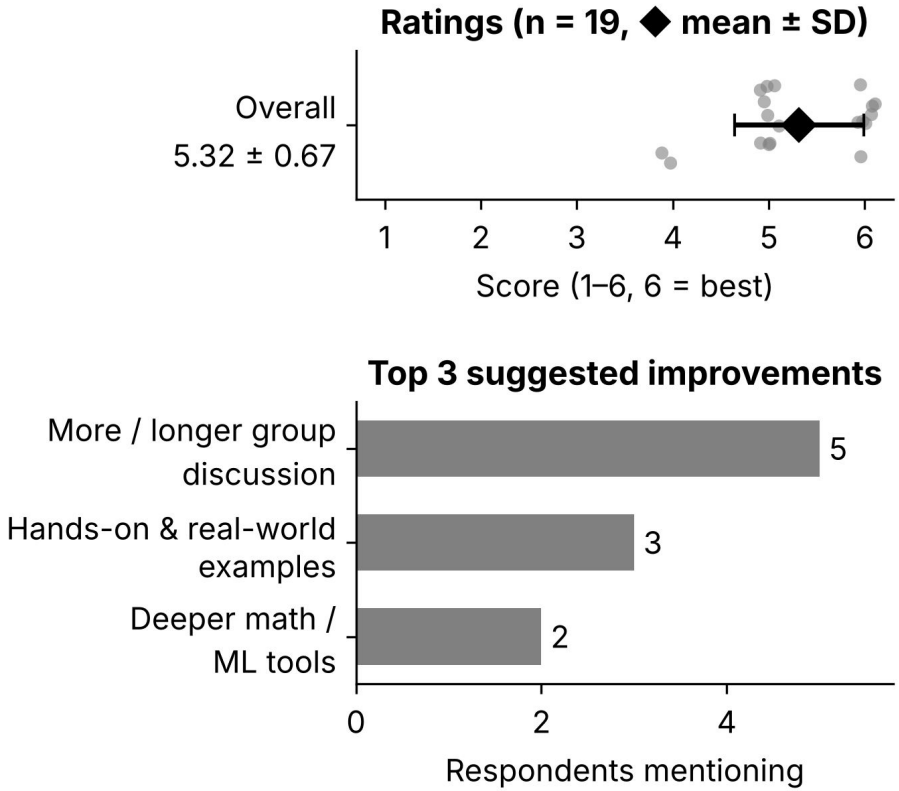
Anonymous feedback about previous lectures



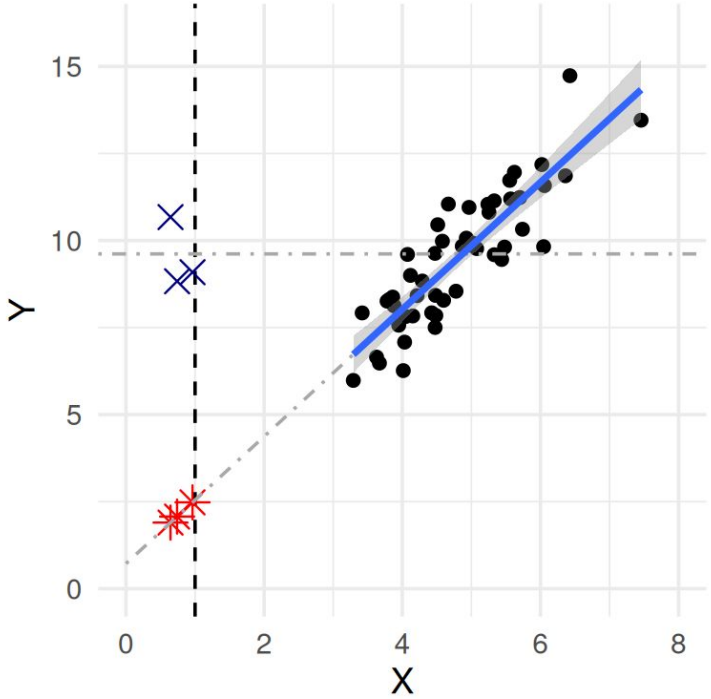
A few things that left strong impressions

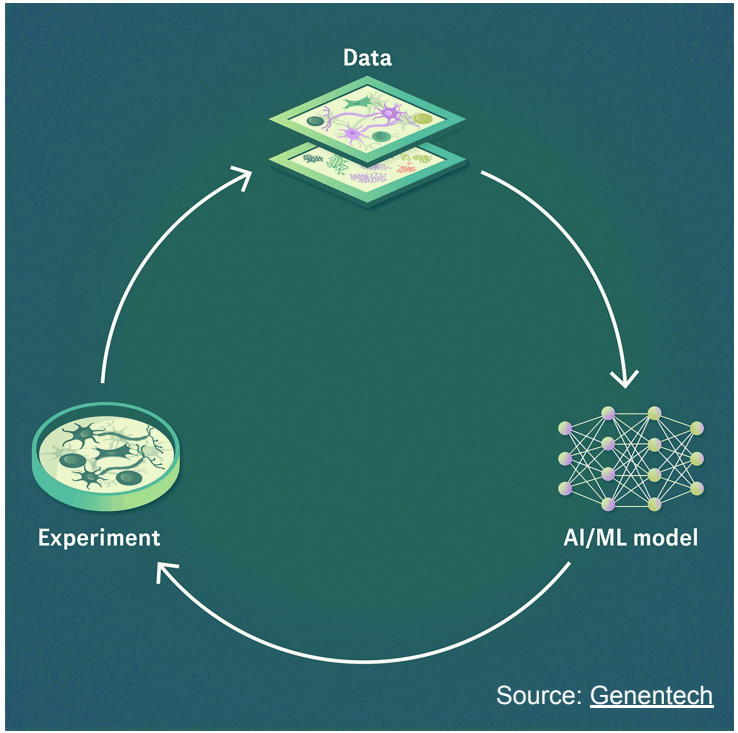
- **Role-play in lecture 4.**
- **How drug discovery really works.** *“How the pharma development world actually works. I had a pretty wrong impression about it in the beginning”*.
- **How little we *really* understand:** *“we don’t even know how Aspirin really works” and “every model is wrong, some are useful” – both facts affected the way I see my chemistry studies.*
- **Feedback loop.** *How much our feedback [has] been [taken] into consideration*
- **Interdisciplinarity.** *“The importance of communication between the fields, various professions.”*

Selected questions and comments from lecture 7

1. **Causal inference in industry practice & real examples:** *“I would find it interesting to see a practical example where a conventional ML model and a causal model lead to different conclusions about a potential drug target.”, “How pharmaceutical companies balance the strict assumptions needed for causal inference models with the messy, incomplete nature of real-world clinical data when deciding whether to advance a drug target.”, “I know Causal Inference is quite big in policy especially in the economy discipline. Is this the same in drug discovery, is it incorporated in the drug discovery pipeline? […] Is it an important factor that is checked before drug approval?”, “Is it possible to discuss such ML/AI use from your own experience. Sort of detailed A to Z steps rather than generic discussion”*
2. **Hill’s criteria & the standard of proof:** *“Do all […] factors need to be satisfied before we can establish a causal relationship, or can causality reasonably be inferred if only 6 […] hold true? In other words, is there a threshold, or does the importance of each factor depend on the context?”, “Since these are just guidelines and not definite criteria for establishing causation, what would be a sure way of telling something is causation and not just correlation?”*
3. **Telling causation from correlation in ML / DAGs:** *“Who draws the DAG? Two honest experts draw two different graphs, and the whole method depends on it.”*, *“Biology is full of feedback loops, and a DAG cannot have cycles. For me that is the biggest open gap.”, “I was wondering if GNNs are currently used in the field of Drug Discovery?”*
4. **Mechanism & plausibility in today’s drug development:** *“How much is understanding the mechanism of action valued during drug development and in the approval process today? In the 2023 paper it is stated as one of the conditions of causality whereas Hill was less strict about it.”, “But today, isn’t it necessary to understand the biological process in order to develop and market new drugs?”,“But if animals are not affected by something how will one even be able to find a drug in today’s time.”*
5. **Individual patient vs. average effect:** *“If randomised trials only show an average effect, how can we ever claim that a drug caused remission in an individual patient?”, “How can we determine whether a drug truly causes remission in an individual patient when the same treatment works for some patients but not others?”*
6. **RCTs: are they enough, and is a placebo ethical?** *“If you have a well designed RCT don’t you get rid of all these problems we learned in these papers?”, “Is a drug believed to cure a lethal disease also tested in trials against a placebo?”*

AMIDD 2026 Lecture 8: Causal inference and AI





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

Causal models are critical when we intend to intervene

Imagine that X is the amount of a particular protein in the blood plasma of a patient with a disease, and Y is the disease severity of that patient.

A central question: *should we develop a drug to reduce the levels of this protein?*

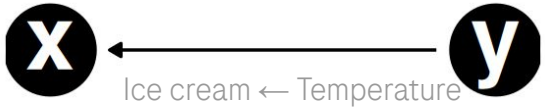
We can use Directed Acyclic Graphs (DAGs) as causal models to identify causality.

Model 1



Temperature → Heat stroke

Model 2

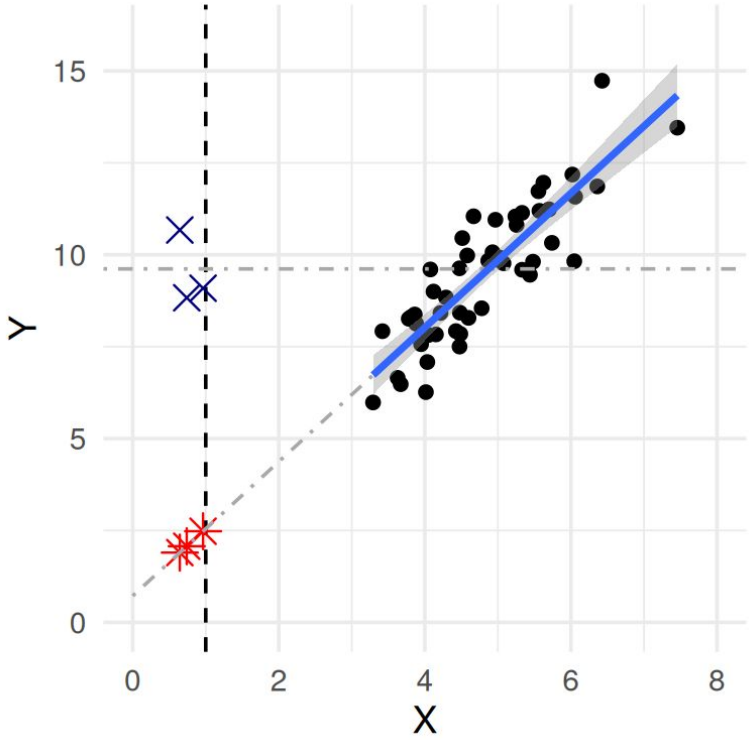


Ice cream ← Temperature

Model 3



Drowning ← Temperature → Ice cream



Common structures of Directed Acyclic Graphs (DAGs)

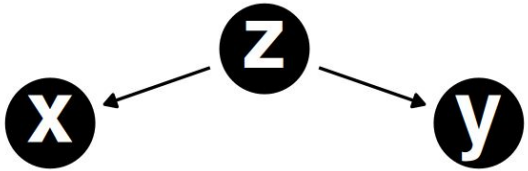
Pipe



X: The quality of a movie
Z: Review of an influencer
Y: Follower watching the movie

X and Y are marginally correlated. Intervening Z changes Y (but not X), removing the correlation between X and Y.

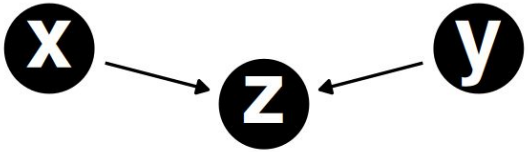
Fork



Z: Dog ownership
X: Number of daily walks
Y: Poop bags used

X and Y are correlated. Intervening Z changes both X and Y. Within each subgroup of Z, X and Y are not correlated.

Collider



Z: Late arrival at work
X: Traffic congestion status
Y: Amount of oversleeping

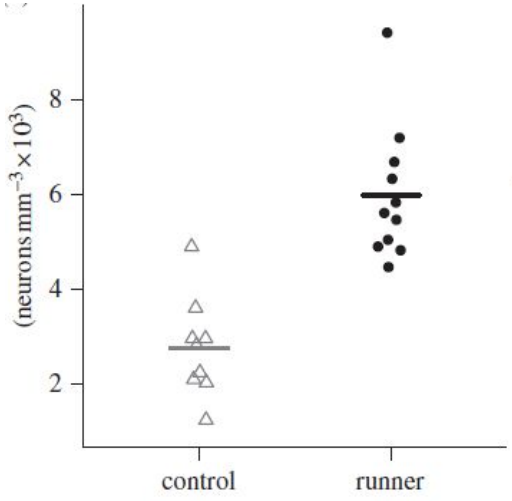
X and Y are not correlated. Intervening Z has no effect on X or Y. Within each subgroup of Z, X and Y are (anti-)correlated.

Causal structures and interventions can create or abolish correlations

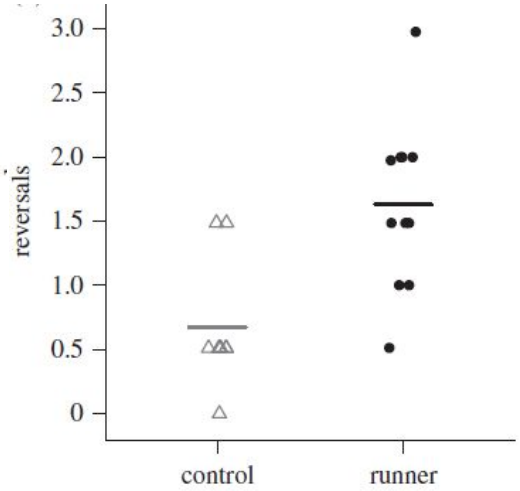
Case study: Running enhances spatial pattern separation in mice

Creer et al., PNAS 2010

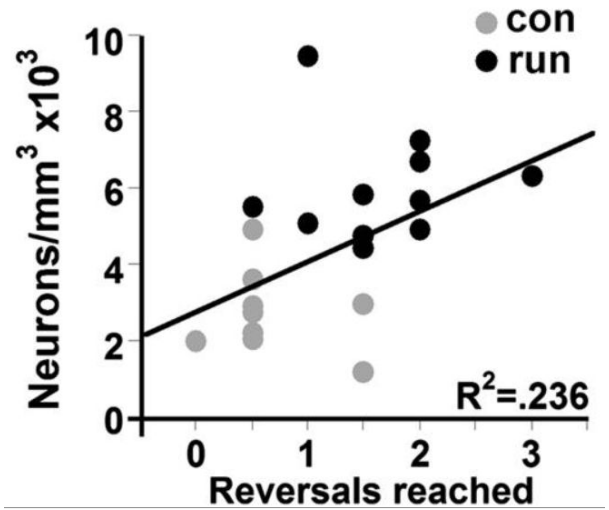
Neurogenesis in mouse hippocampus



Ability to discriminate locations of adjacent stimuli

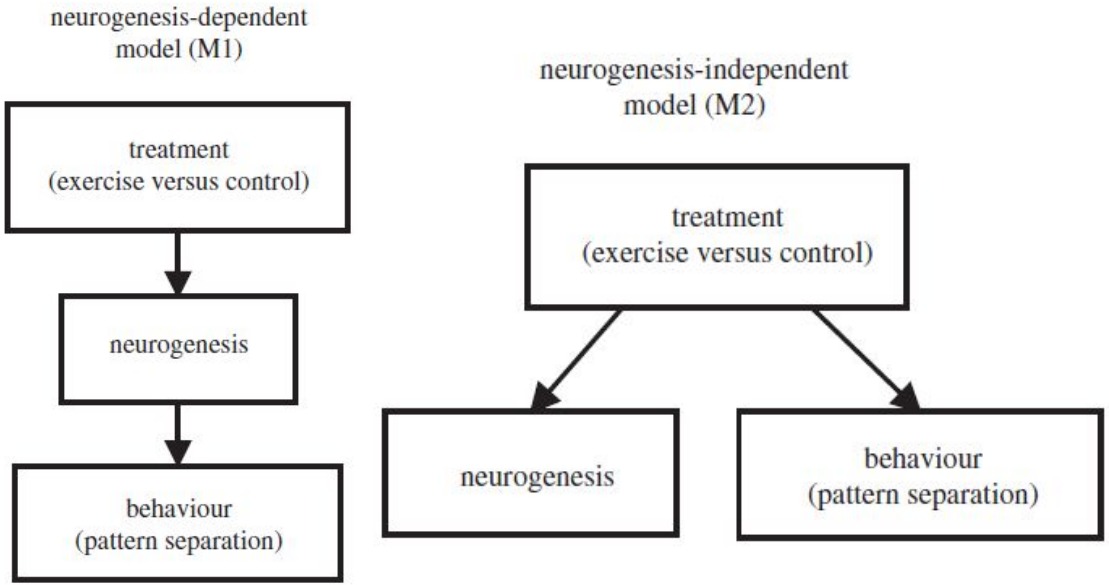


Correlation between the two measurements



Creer, David J. et al. "Running Enhances Spatial Pattern Separation in Mice." Proceedings of the National Academy of Sciences 107, no. 5 (February 2, 2010): 2367–72.

The billion-dollar question: should we develop a drug enhancing neurogenesis to improve brain health?



M1 (the **pipe** model) implies that neurogenesis and behavior would be *positively* correlated within each subgroup of treatment.

M2 (the **fork** model) implies that neurogenesis and behavior would not be correlated within each subgroup of treatment.

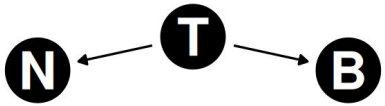
While M1 supports initiating a discovery program, M2 discourages us from doing so.

Lazic Stanley E. "Using Causal Models to Distinguish between Neurogenesis-Dependent and -Independent Effects on behavior." Journal of The Royal Society Interface 9, no. 70 (May 7, 2012): 907–17.

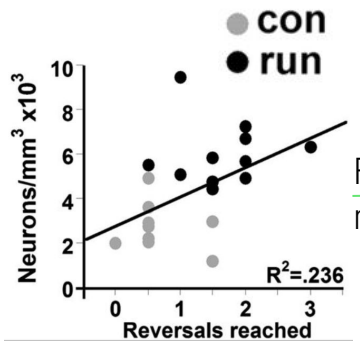
Causal thinking reduces cognitive bias by weighing models explicitly

M1
(neurogenesis-dependent model; T: Treatment; N: Neurogenesis; B: Behavior)

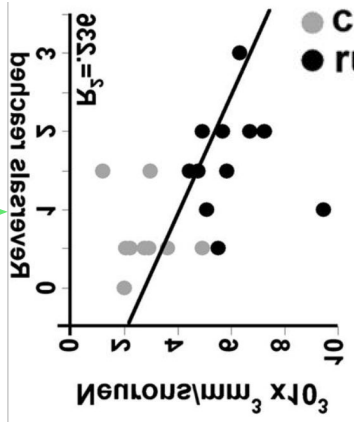




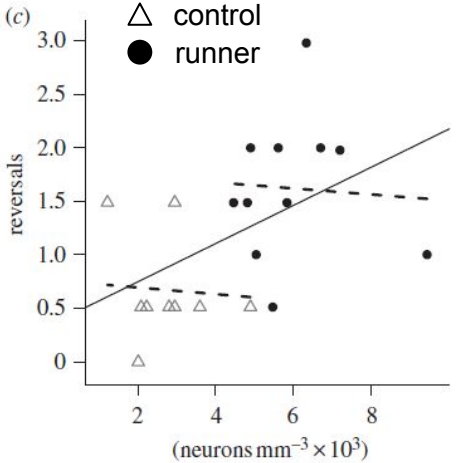
M2
(neurogenesis-independent model)



Rotating & mirroring



Regression within control and exercise groups

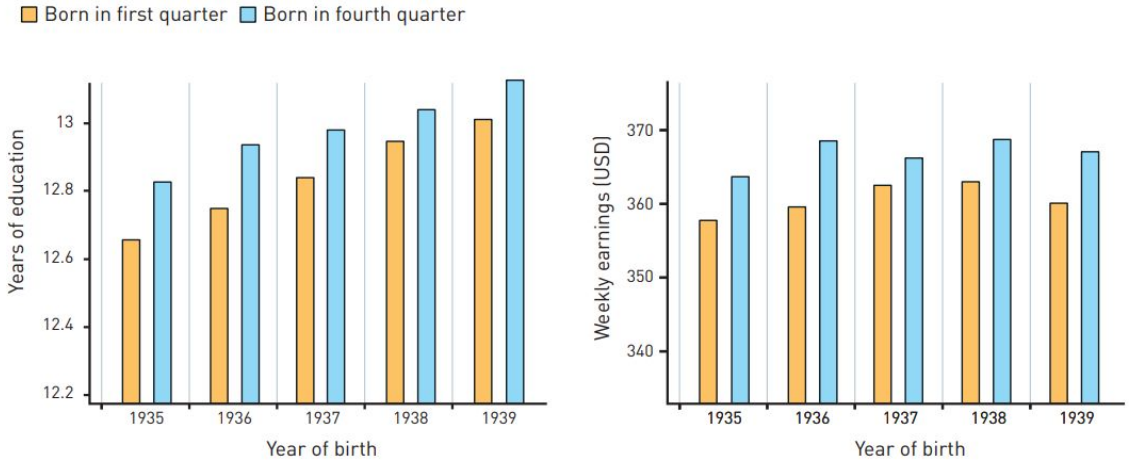


Neurogenesis (N) and behavior (B) are barely correlated (if anything, negatively) in both control and exercise groups, which refutes M1 (the pipe model) while strengthening M2 (the fork model). This piece of evidence speaks *against* starting a discovery program.

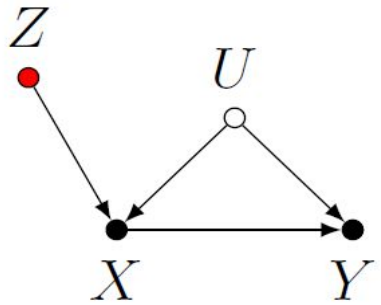
We note that several industrial projects aiming at promoting neurogenesis have been terminated.

We can quantify causal relationships that are otherwise difficult to measure experimentally

In the US, children can choose to leave school when they turn 16 or 17. Therefore, children born early in the year can legally leave school earlier than those born late. Angrist and Krueger (1991) observed that people born late in the year have more years of education and higher incomes.



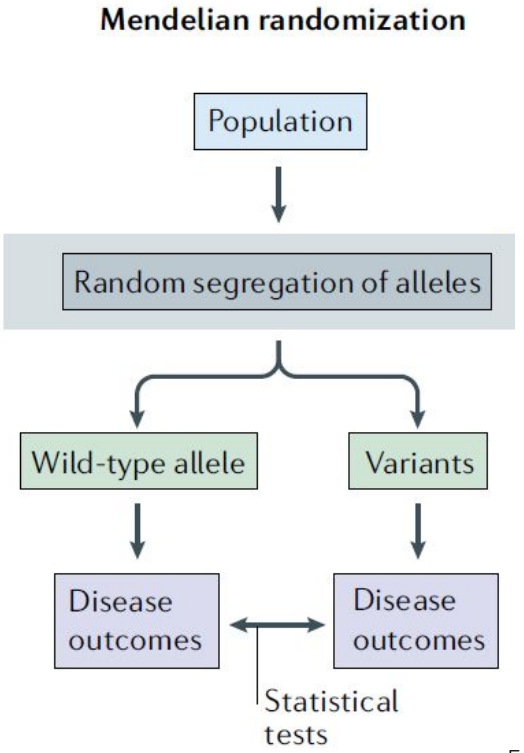
Natural experiments help answer important questions; the Prize in Economic Sciences in Memory of Alfred Nobel, 2021, which was awarded to David Card, Joshua Angrist and Guido Imbens. J. Angrist and Alan B. Krueger. (1991), "Does Compulsory School Attendance Affect Schooling and Earnings?," The Quarterly Journal of Economics.



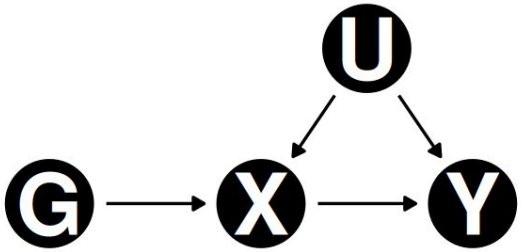
Z: Birthday (the *instrumental variable)*
X: Education
Y: Income
U: Socioeconomic/ individual factors

The causal effect is estimated by the ratio between the difference in earnings between late-born and early-born, and the difference in schooling (~9% per additional year of education).

Mendelian Randomization (MR) applies causal thinking to identify targets for drug discovery



Each of us carries millions of genetic variants (SNPs) that are randomly assigned. We can use them as instrumental variables to test the causality between genes and diseases.



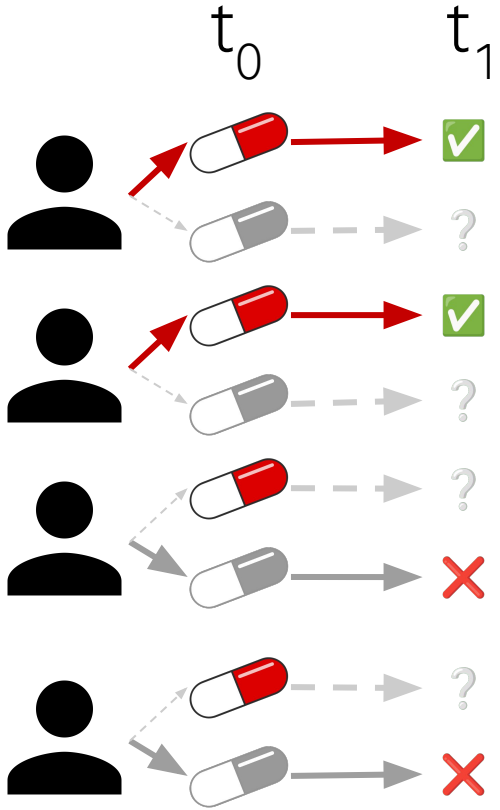
G: Genetic variant(s) associated with changes in X, the *instrumental variable*
X: Exposure, *e.g.* protein level
Y: Outcome, *e.g.* disease incidence/severity
U: Confounding factors, *e.g.* sex, diet, age

Example: Circulating FGF21 protein levels are correlated with symptoms of the liver disease MASH (metabolic dysfunction-associated steatohepatitis). MR confirms a causal relationship between FGF21 and several risk factors of MASH, supporting FGF21 as a potential therapeutic target.

E. Sanderson et al. (2022), "Mendelian randomization," Nature Reviews Methods Primers; Larsson *et al.* (2022), Metabolism

Causal inference is important for both randomized experiments and observational studies

- In drug discovery and development, we often care about ***potential outcomes*** or ***counterfactuals***: what had if the patient received the alternative treatment, keeping everything else constant?
- **Randomized experiments** and **controlled trials** are gold-standard methods to address causal questions. Non-compliance and intermittent events call for causal analysis of the data even in randomized trials.
- Given causal models, it is sometimes possible to learn causal relationships from observational data as well.



Causal inference is a missing data problem

Individual	Treatment	Value (AU)
1	Control	75
2	Control	73
3	Control	74
4	Treatment	55
5	Treatment	45
6	Treatment	60

A classical textbook

Individual	Value (AU) with Control	Value (AU) with Treatment
1	75	?
2	73	?
3	74	?
4	?	55
5	?	45
6	?	60

A classical textbook in 50 years

Assignment mechanism determines which data are missing, and determine the statistical technique to be used

Individual	Value (AU) with Control	Value (AU) with Treatment
1	75	?
2	73	?
3	74	?
4	?	55
5	?	45
6	?	60

- **Classical Randomized Experiments**: we control the assignment mechanism;
- **Regular assignment (*observational studies)***: we know part but not all of the assignment mechanism;
- **Regular assignment with non-compliance**: we need an *instrumental variable*.

Causality is crucial for drug discovery





?

?

Biomarker, tox study, pathology, omics data, real-world data, …

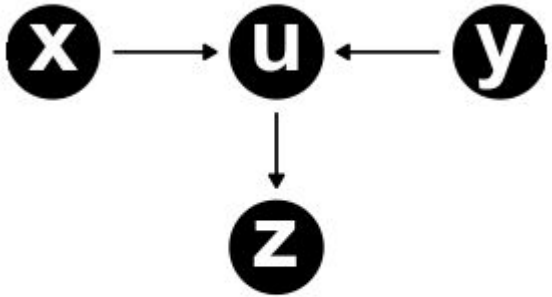
	x	z	y
1	0.835386320	1	-0.73897252
2	-0.005354014	-1	-0.82972315
3	0.058788286	1	0.76213369
4	-1.015602246	-1	-0.05951719
5	-0.339569780	-1	-0.11745910
6	-0.041077979	-1	-1.28243716
7	0.363740407	1	-0.30570762
8	0.119496314	-1	-1.19932461
9	0.257108454	-1	-1.06044066
10	0.304537158	-1	-0.43396492



?

?

The descendant



We need both models (knowledge + assumptions) and data to infer causality.

Accurate predictions on small data with a tabular foundation model

Almost too good to be true?



"I knew the indoor pool was too good to be true."

Tabular Prior-data Fitted Network (TabPFN)

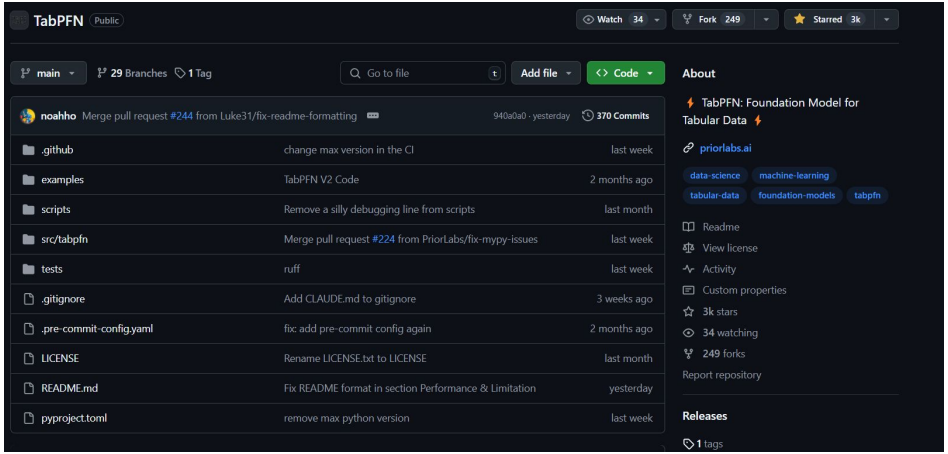
Article

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]

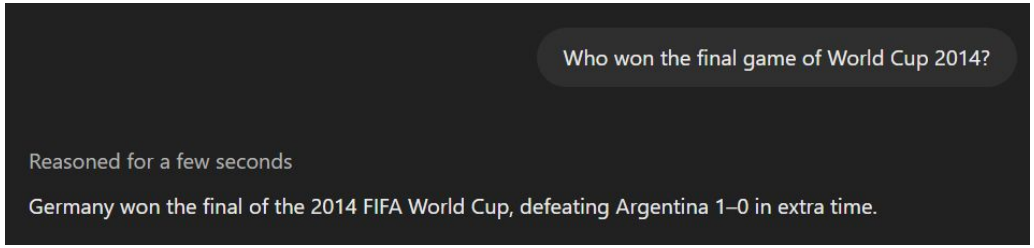
Tabular data, spreadsheets organized in rows and columns, are ubiquitous across scientific fields, from biomedicine to particle physics to economics and climate science[1,2]. The fundamental prediction task of filling in missing values of a label column based on the rest of the columns is essential for 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).

In-Context Learning (ICL)

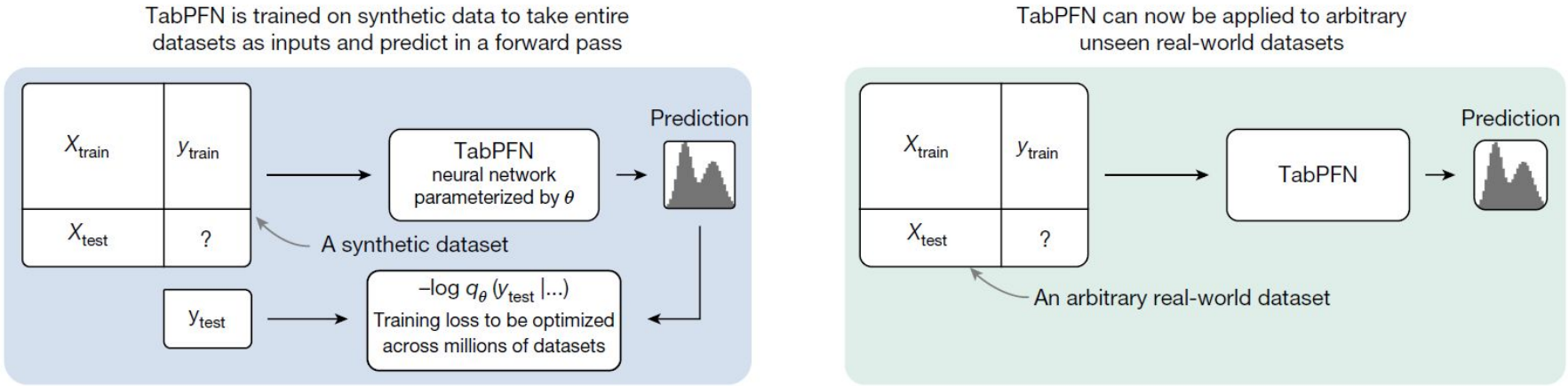
The model isn't explicitly trained on a particular task ahead of time. Instead, the prompt itself provides examples that enable the model to perform the task on the fly. It is one of the intriguing abilities of deep transformer models.





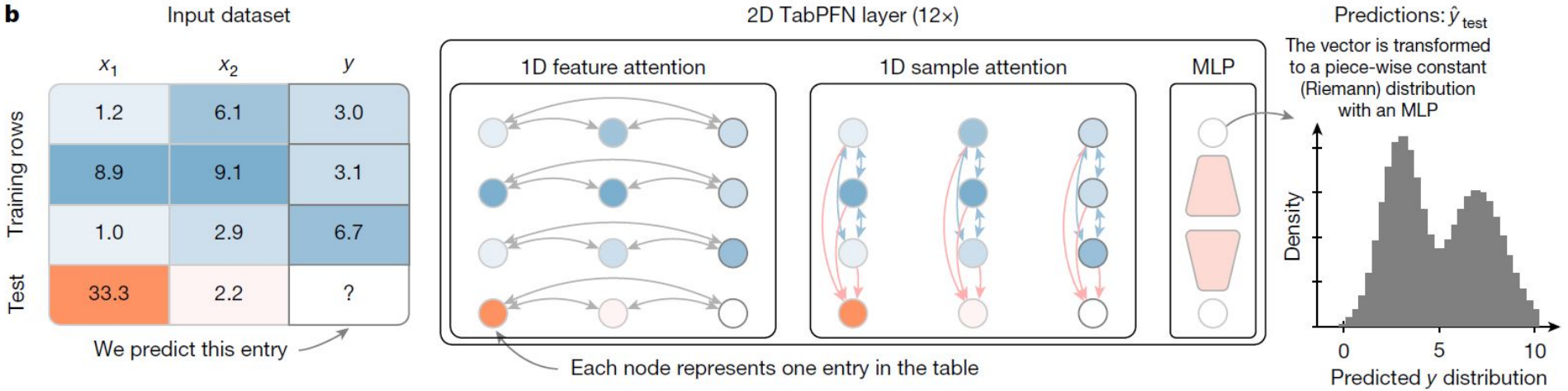
Left: chat history (ChatGPT o1 model, tested on 18.03.2025). Right: The ground truth.

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

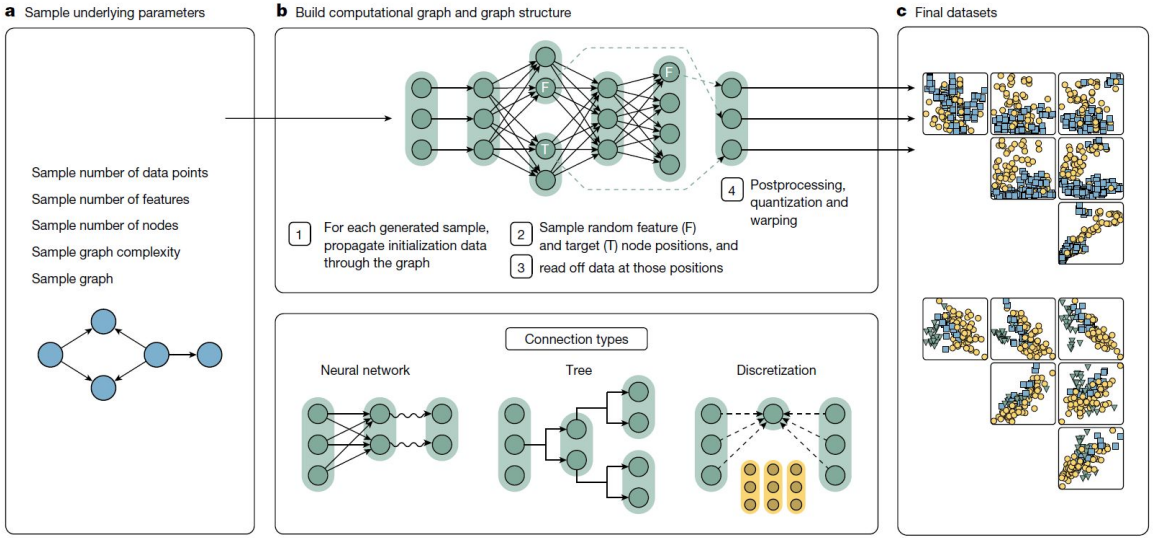


A metaphor: Imagine a lab where millions of billiard games are played simultaneously: in each game, different numbers of balls are placed randomly, and a white ball starts with a random velocity at a random position. By learning the trajectory of all balls of all games, one may learn to predict the trajectory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

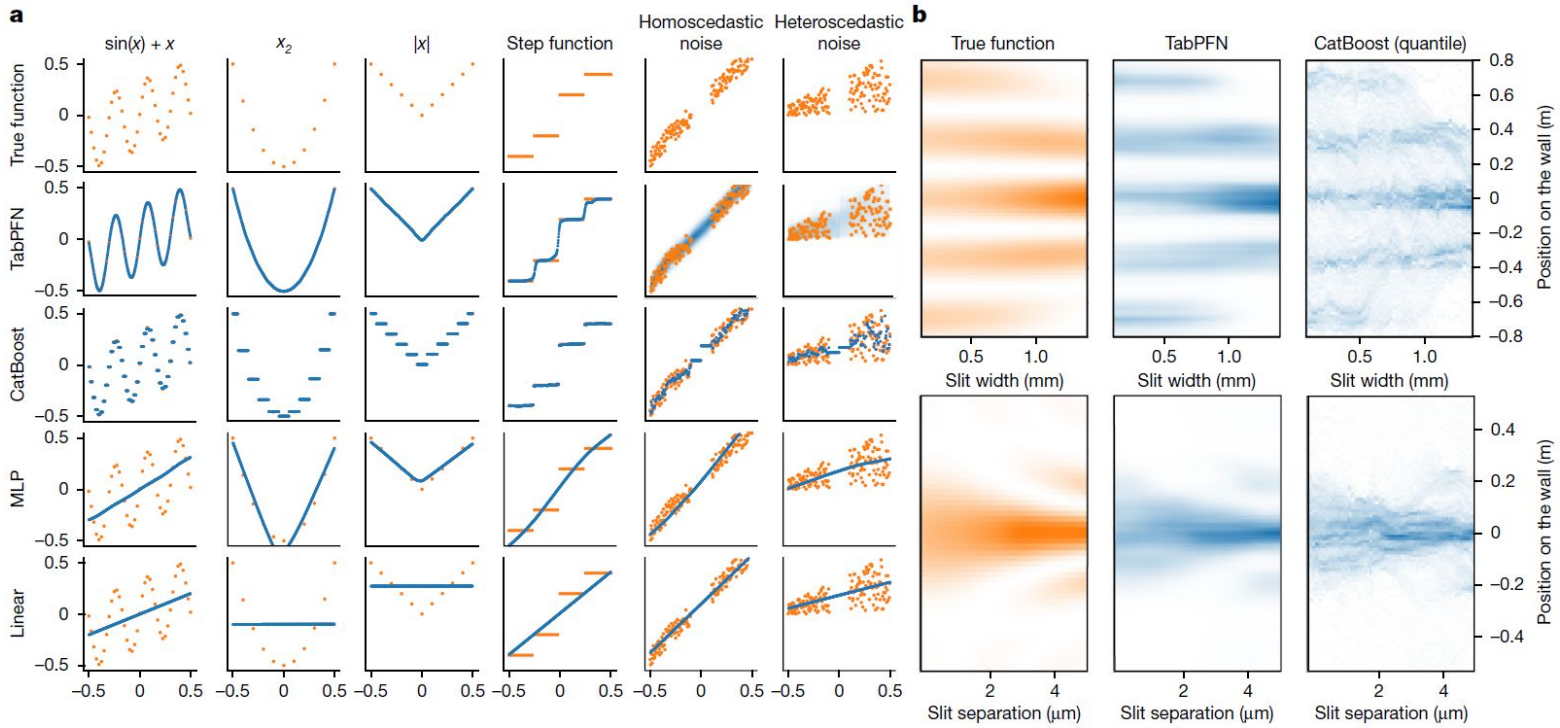


Generating synthesized data with causal models

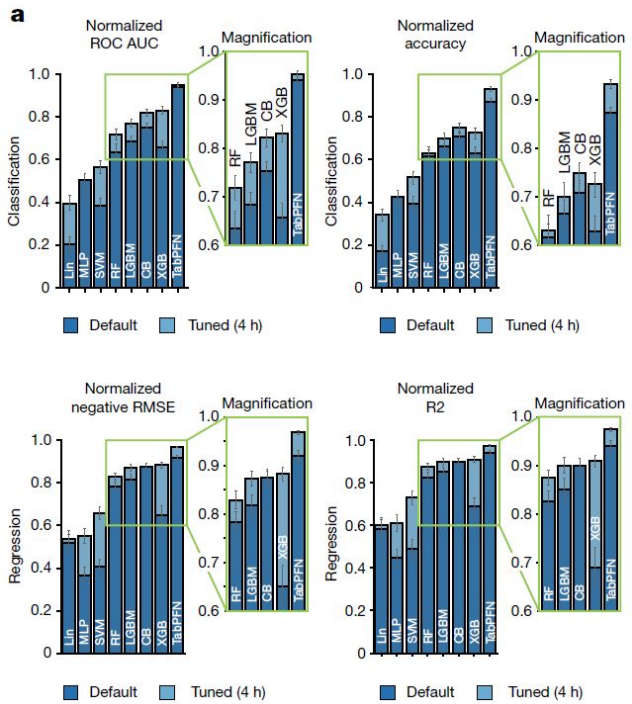


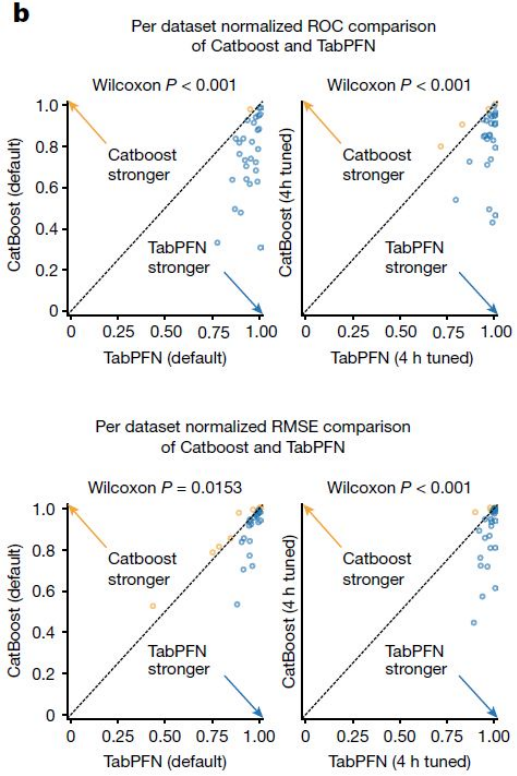
1. We sample 100-million structurally unique directed acyclic graphs (DAGs) as causal graphs that generate data.
2. Random data is assigned to the root node. Other nodes are propagated by rules specified by the edges, plus a Gaussian noise.
3. The values of sampled features (input to the model) and targets (output of the model) are extracted.
4. Data are post-processed (non-linear distortion, binning, random missing) to reflect real-world data processing.
5. Synthesized data from 10^8 causal experiments are used to train *TabPFN*.

TabPFN's prediction on data generated by simple functions



Performance benchmark against popular methods





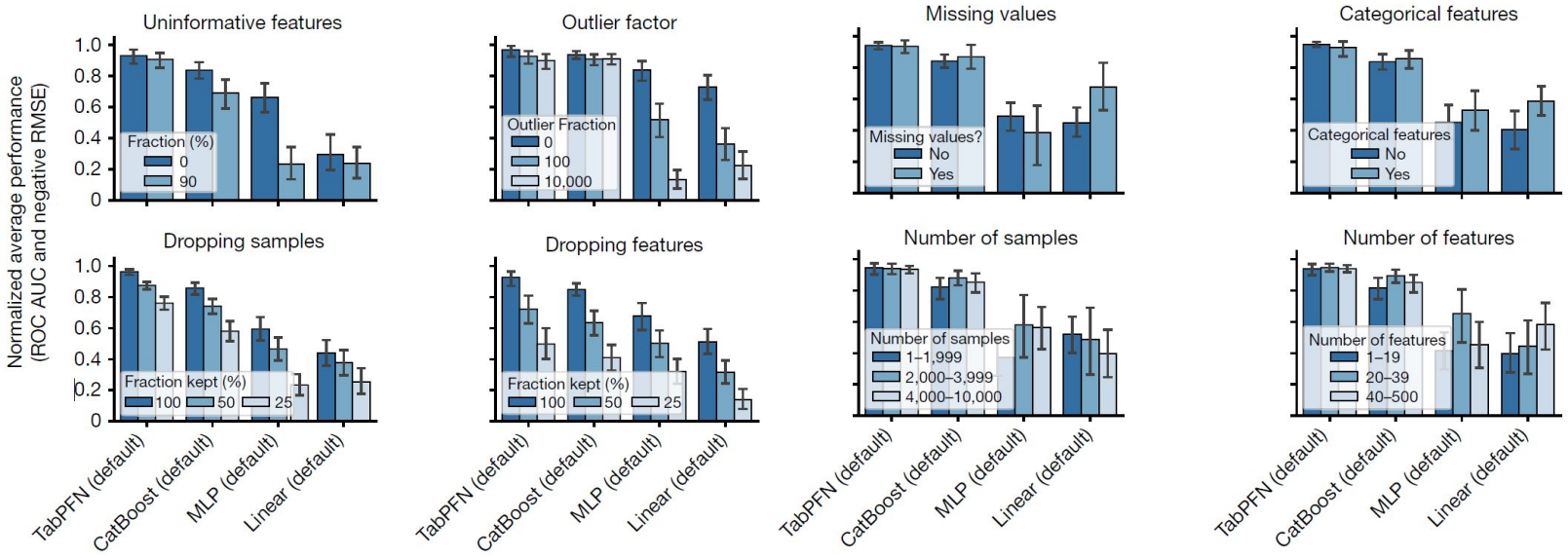
Left: Performance for classification (top) and regression (bottom) tasks.

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

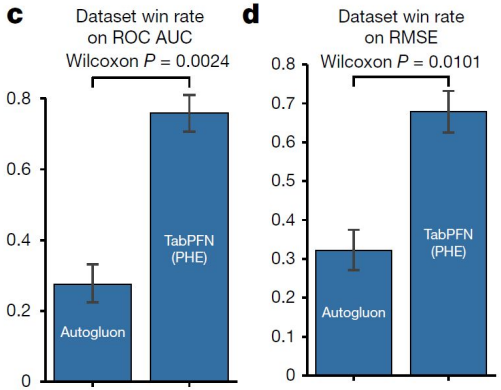
Hollmann, Noah, Samuel Müller, …, Frank Hutter. 2025. “Accurate Predictions on Small Data with a Tabular Foundation Model.” Nature 637 (8045): 319–26.

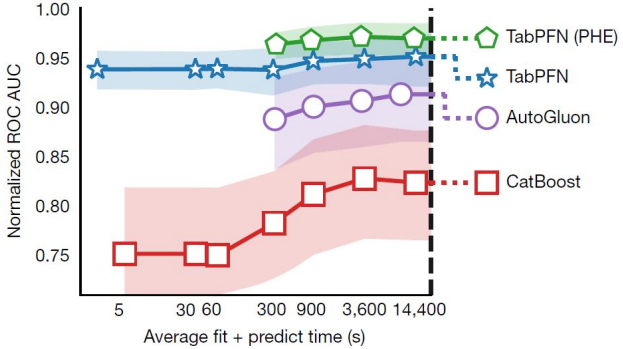
Lin: linear; MLP: multi-layer perceptrons; SVM: support vector machines; RF: random forest; LGBM: Light Gradient Boosting Machine; CB: CatBoost; XGB: XGBoost; ROC: receiver operating characteristic; RMSE: root mean square error.

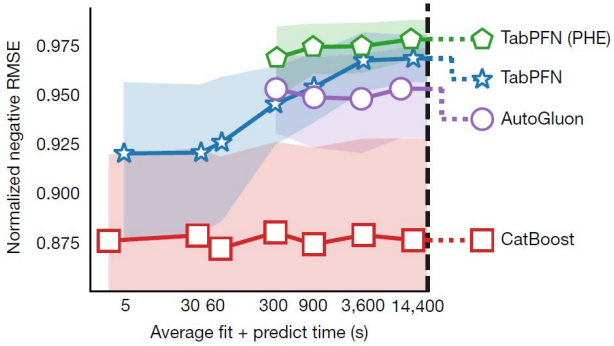
Robustness against common caveats and problems



Benchmark against AutoML and CatBoost







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

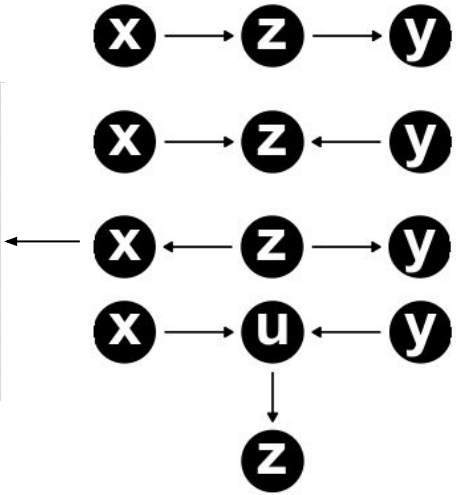
All regression tasks from the AutoML[36] and OpenML-CTR23[37] Benchmarks with fewer 10,000 samples and 500 features. The benchmark comprises diverse real-world tabular datasets, curated for complexity, relevance, and domain diversity.

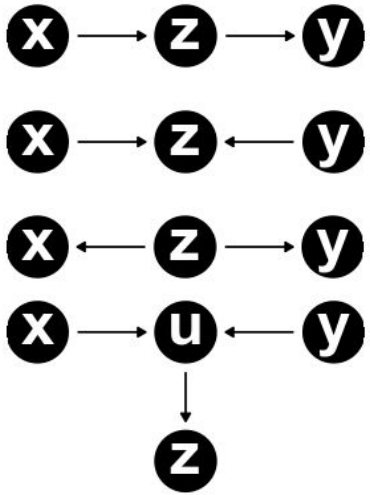
We witness a paradigm shift in analysing tabular data: from *causal model selection* to *causality-assisted prediction*

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

From now on: Sample a large number of causal models, synthesize data from them, and train ML models trained on synthesized

	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.

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 unlimited approaches in relatively short time (though not without possible 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.