

Causal meets generative AI: from reasoning “why” to imagining “what If”

ELLIIT 2026 Focus Period – Halmstad University

Giorgos Papanastasiou, PhD

Associate Research Professor, Mathematics Research Centre, Academy of Athens

Lead Principal Investigator, Archimedes Unit, Athena Research Centre

Associated Faculty Member, Biomedical Research Foundation of Academy of Athens

Member of <https://ellis.eu/>

22.09.2026

Academy of Athens - Mathematics RC
(Associate Research Professor)



Pfizer HQs – New York (Research Scientist)



Pfizer CDI –
Thessaloniki



Archimedes Unit, Athena RC
(Lead PI)



University of Essex (Assistant Professor)



University of Edinburgh (Faculty Research Scientist)



University of Edinburgh (PhD)



University of Ioannina (MEng)



National Technical University of Athens (MSc)



University of Groningen (MSc)

Motivation (example 1): Challenges and biases in non-Causal AI/ ML

nature
machine intelligence

ANALYSIS
<https://doi.org/10.1038/s42256-021-00307-0>


OPEN

Common pitfalls and recommendations for using machine learning to detect and prognosticate for COVID-19 using chest radiographs and CT scans

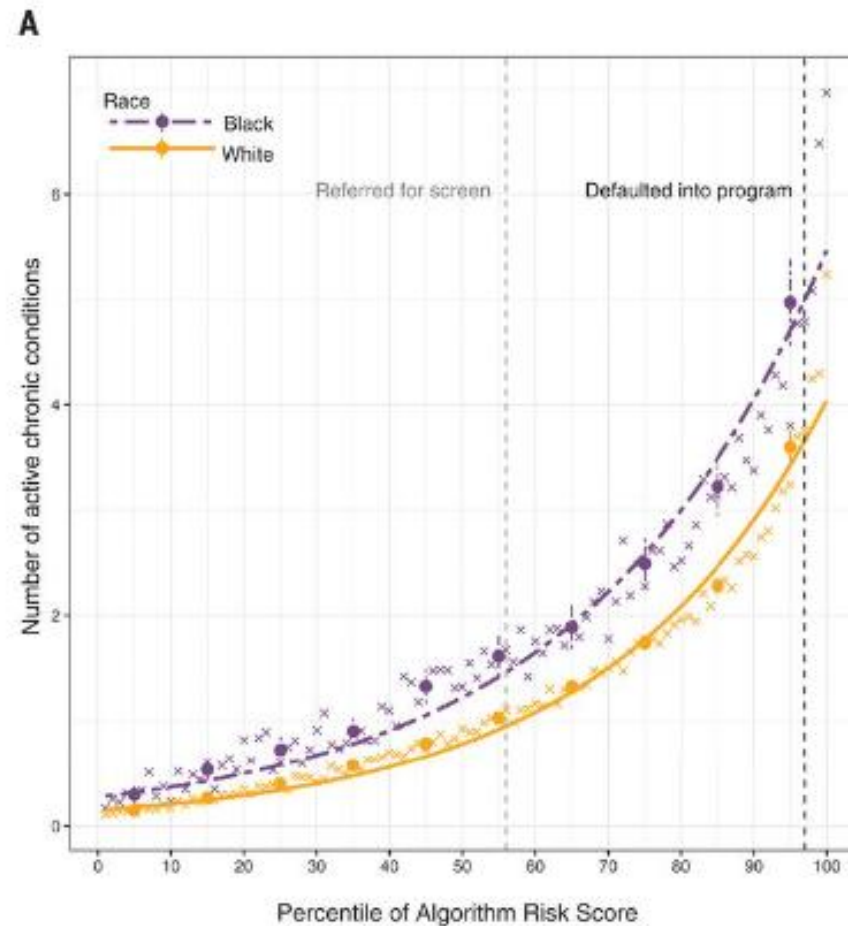
Michael Roberts ^{1,2}✉, Derek Driggs¹, Matthew Thorpe³, Julian Gilbey ¹, Michael Yeung ⁴,
Stephan Ursprung ^{4,5}, Angelica I. Aviles-Rivero¹, Christian Etmann¹, Cathal McCague^{4,5},
Lucian Beer⁴, Jonathan R. Weir-McCall ^{4,6}, Zhongzhao Teng⁴, Effrossyni Gkrania-Klotsas ⁷,
AIX-COVNET^{*}, James H. F. Rudd ^{8,36}, Evis Sala ^{4,5,36} and Carola-Bibiane Schönlieb^{1,36}

Machine learning methods offer great promise for fast and accurate detection and prognostication of coronavirus disease 2019 (COVID-19) from standard-of-care chest radiographs (CXR) and chest computed tomography (CT) images. Many articles have been published in 2020 describing new machine learning-based models for both of these tasks, but it is unclear which are of potential clinical utility. In this systematic review, we consider all published papers and preprints, for the period from 1 January 2020 to 3 October 2020, which describe new machine learning models for the diagnosis or prognosis of COVID-19 from CXR or CT images. All manuscripts uploaded to bioRxiv, medRxiv and arXiv along with all entries in EMBASE and MEDLINE in this timeframe are considered. Our search identified 2,212 studies, of which 415 were included after initial screening and, after quality screening, 62 studies were included in this systematic review. Our review finds that none of the models identified are of potential clinical use due to methodological flaws and/or underlying biases. This is a major weakness, given the urgency with which validated COVID-19 models are needed. To address this, we give many recommendations which, if followed, will solve these issues and lead to higher-quality model development and well-documented manuscripts.

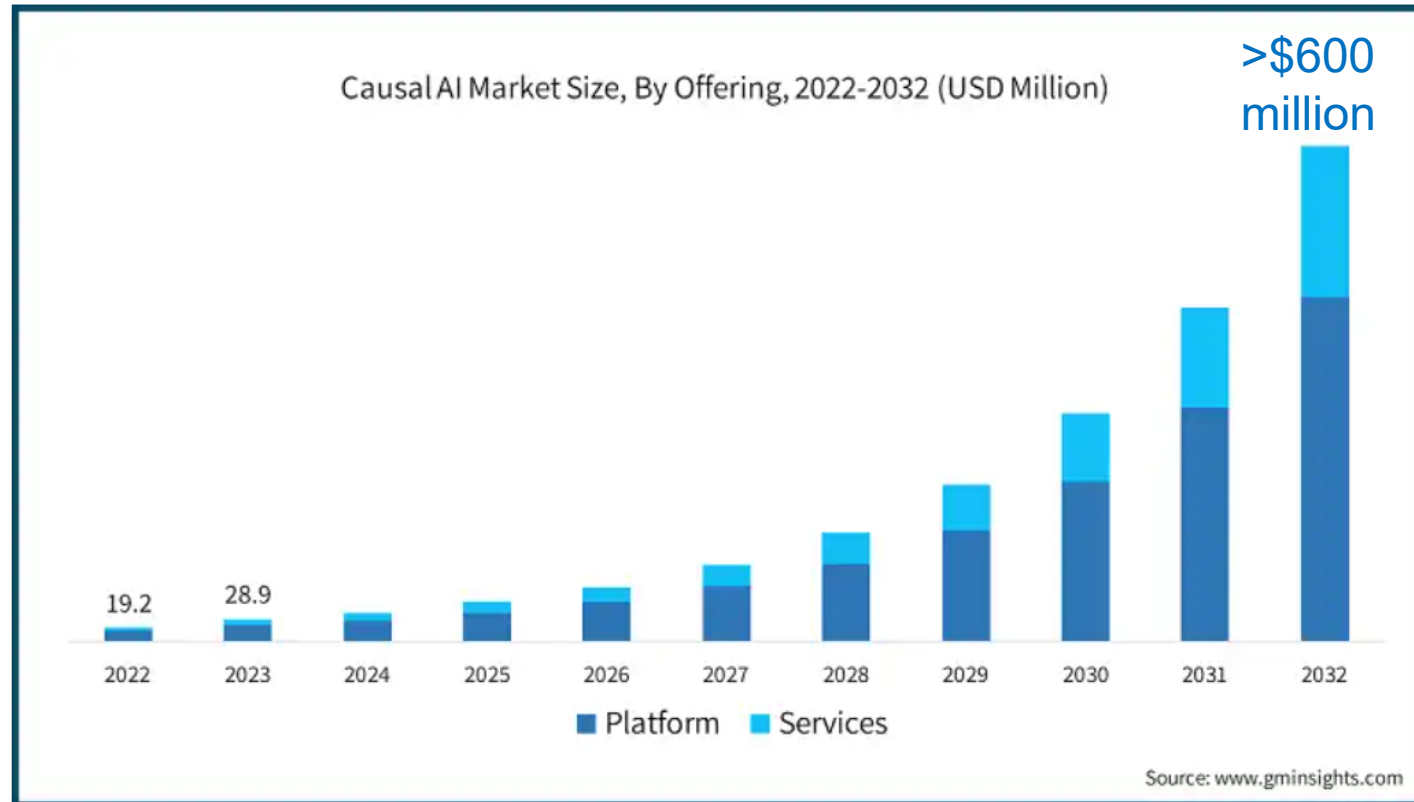
Motivation (example 2): Challenges and biases in AI/ ML

Dissecting racial bias in an algorithm used to manage the health of populations

ZIAD OBERMEYER , BRIAN POWERS, CHRISTINE VOGELI, AND SENDHIL MULLAINATHAN  [Authors Info & Affiliations](#)



Causal AI market growth



According to financial reports by

Global Market Insights: <https://www.gminsights.com/industry-analysis/causal-ai-market>

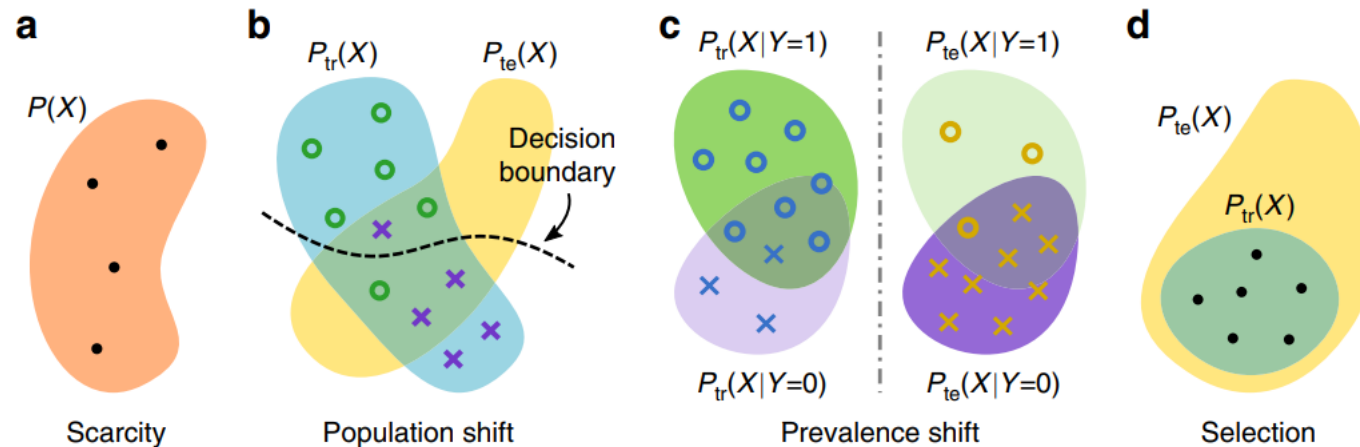
MarketsAndMarkets: <https://www.marketsandmarkets.com/Market-Reports/causal-ai-market-162494083.html>

MarkNtel Advisors: <https://www.marknteladvisors.com/research-library/causal-artificial-intelligence-market.html>

Many others: Harvard Business Review, McKinsey, Microsoft

Key open challenges in AI

Our **current AI** is based on **correlations, not causality**. **Key open challenges** that can be addressed via **causality**:



We need causality:

1. AI/ ML can not **generalize** when new “shifts” in data occur, and is **not interpretable**¹.
2. **Central to causality, is to unravel the relationship between cause and effect, via the detection of data confounders (spurious associations)**²

A logical perspective on AI & causality

- **Standard AI (non-causal)** relies on the **joint distribution** $P(x, y)$ where x are the data (labelled or not) and y is the output (AI prediction).
- However, **correlation is symmetric** ($x \sim y$), but **causation is directional** ($x \rightarrow y$). In other words, probability alone cannot distinguish between a cause (a drug) and a confounder (a patient's genetic profile, age, gender or socioeconomic status) “**living actively**” inside your dataset and “**biasing**” your AI analysis if not considered.

The causal hierarchy (by Pearl)

The three levels of causal reasoning:

- Association $P(y|x)$ (via causal discovery or defined by domain experts): **seeing directional patterns**. "If I see a patient taking this drug (data x), are they likely to recover (output y)?"
- Intervention $P(y|do(x))$: **doing**. "If I force a patient population to take drug A , will they recover?"
- Counterfactuals $P(y_x | x', y')$: **imagining**. "Given this patient took the drug x' and failed to recover y' , would they have recovered if they would have taken another drug?"

Step 1: Causal discovery and directed acyclic graphs (DAG)

A DAG³:

1. Is typically a **Bayesian network** whose nodes present variables and whose edges represent directed influence among or between variables:

$$P(X) = \prod_{i=1}^d P(X_i | X_{\pi i})$$

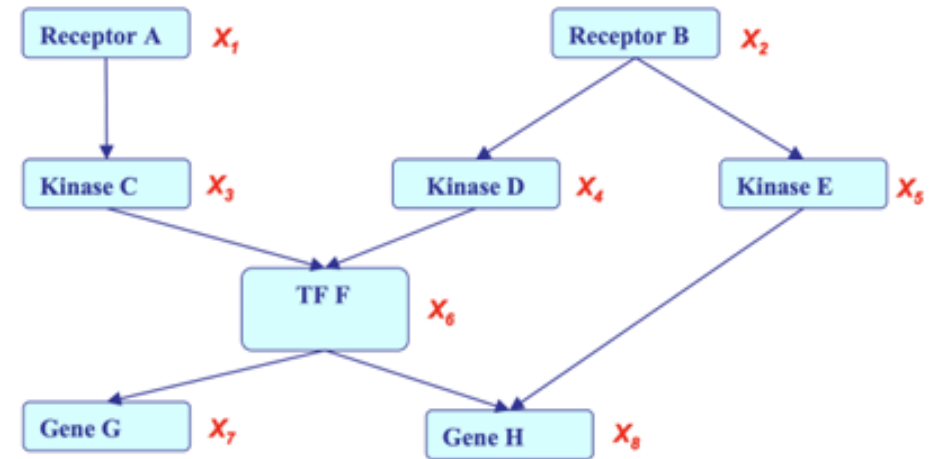
where $X_{\pi i}$ is the set of parents of X_i , d is the number of variables in the DAG.

Note that the **goal of causal discovery** is to identify:

$$X_i = f_i(Pa(X_i)) + \epsilon_i$$

Here, ϵ_i represents exogenous noise - the unique 'personal signature' of the individual(s). It accounts for unobserved factors (e.g., data acquisition parameters, genetics, lifestyle, history) that exist **outside your dataset**, but define how a specific subject responds to a change.

2. A DAG can be **calculated by developing causal discovery algorithms or defined by domain experts**.



The joint probability of the above directed graph can be written as:

$$P(X_1, X_2, X_3, X_4, X_5, X_6, X_7, X_8) = P(X_1)P(X_2)P(X_3|X_1)P(X_4|X_2)P(X_5|X_2)P(X_6|X_3, X_4)P(X_7|X_6)P(X_8|X_5, X_6)$$

3. Eric P.Xing. Probabilistic graphical models; 2020.

Step 2: What happens now if we want to perform an intervention?

The mechanics of intervention: do(X) When we intervene on a system (e.g., applying a drug to **Kinase D**, i.e., X_4), performing a “graph surgery”.

1. The graph surgery involves:

Cut incoming arrows: We delete all arrows pointing into the intervened node (X_4).

The Logic: Since the researcher now controls X_4 externally, its natural parents (Receptor B) no longer influence it.

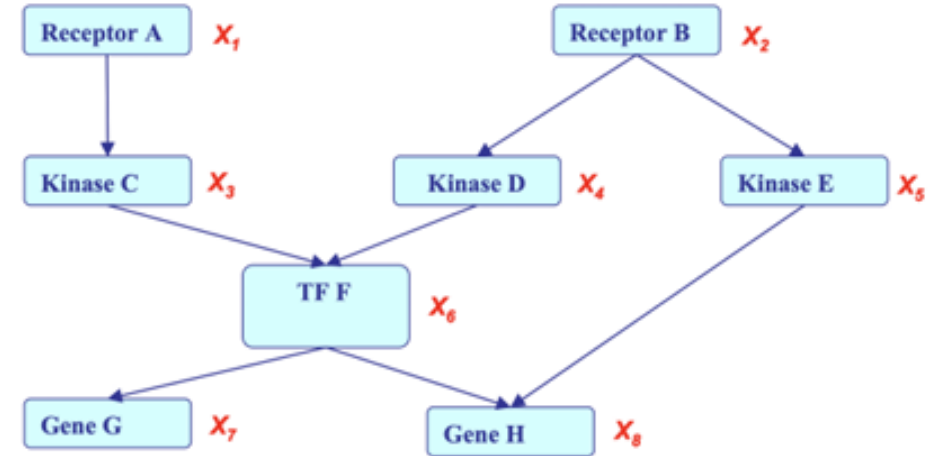
Preserve downstream flow: All outgoing arrows from X_4 remain. The forced change still flows down to **TF F** (X_6) and **Genes G** (X_7) and **H** (X_8).

2. The mathematical transformation involves:

Intervention changes how we factorize the system's probability. We remove the variable's "natural" behavior and replace it with our "forced" value.

Observational world (natural correlation): $P(X_1, X_2, X_3, X_4, X_5, X_6, X_7, X_8) = P(X_1) * P(X_2) * P(X_3|X_1) * P(X_4|X_2) * P(X_5|X_2) * P(X_6|X_3, X_4) * P(X_7|X_6) * P(X_8|X_5, X_6)$

Interventional world (surgical do-operation): $P(X_1, X_2, X_3, X_5, X_6, X_7, X_8 | do(X_4=x)) = P(X_1) * P(X_2) * P(X_3|X_1) * [1] * P(X_5|X_2) * P(X_6|X_3, x) * P(X_7|X_6) * P(X_8|X_5, X_6)$



Step 3: What happens if we want to perform a “what-if” counterfactual scenario?

The mechanics of counterfactual estimation:

In a DAG, we assume the world is **deterministic** once all background noise ϵ is accounted for, by solving:

$$X_i = f_i(Pa(X_i)) + \epsilon_i,$$

where ϵ_i is the “personal signature“, i.e., the sum of everything **exogenous & unique** to this specific dataset.

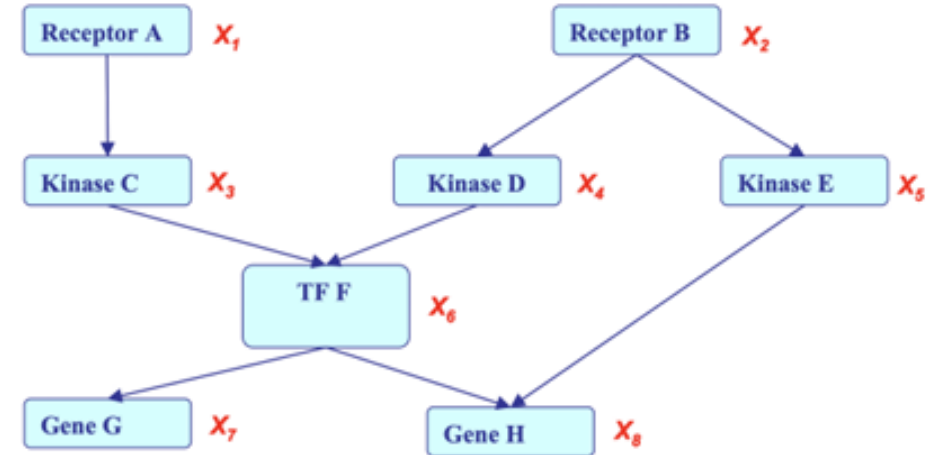
The three-step logic:

How the model “imagines“:

Abduction (capture the ϵ_i signature): We look at the factual data (what actually happened). We calculate the unique personal signature ϵ_i required to explain why this specific individual(s) behaved the way they did.

Action (apply the change): We perform a do-operation to change a single variable (e.g., as we did before with X_4). Crucially, we keep the ϵ_i fixed.

Prediction (generate the parallel result): We run the model forward. Because the signature is constant, we can see how that exact same individual(s) would have responded in a parallel world, imaging a different “past”.



Exemplar 1 - Large scale causal modeling to identify adults at risk for combined and common variable immunodeficiencies

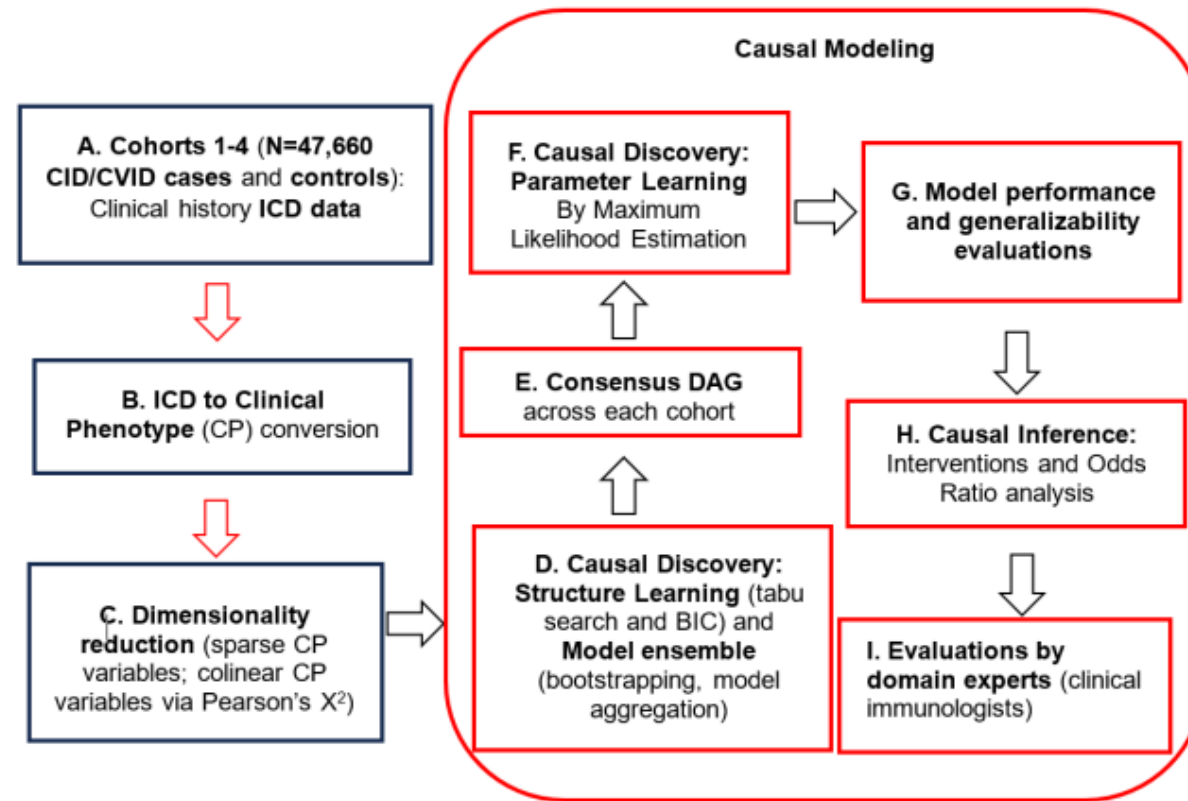


Figure 1) Study workflow. CID: Combined Immunodeficiency, CVID: Common Variable Immunodeficiency, BIC: Bayesian Information criterion, DAGs: directed acyclic graphs.

Estimated DAGs (Cohorts 3 and 4)

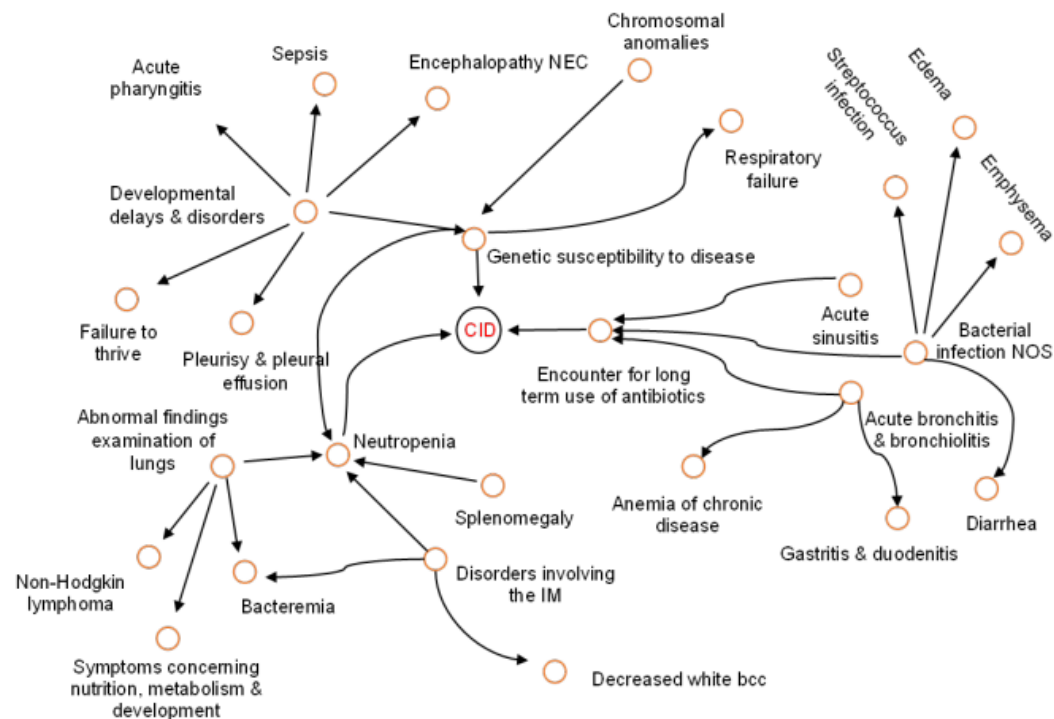


Figure 4) Consensus DAG calculated in Cohort 3. Cohort 3 involved N=2,312 CID cases (of which 797 with pneumonia) and 2,312 matched controls (with no PI), both with and without pneumonia. We visualize up to 2 direct parent levels and up to 2 direct child levels away from CID diagnosis. Up to 1 direct parent level for each direct child and up to 1 direct child level for each direct parent are included. DAG: directed acyclic graph; NEC: not elsewhere classified; NOS: not otherwise specified; IM: immune mechanism; bcc: blood cell count; CID: combined immunodeficiency; PI: primary immunodeficiency.

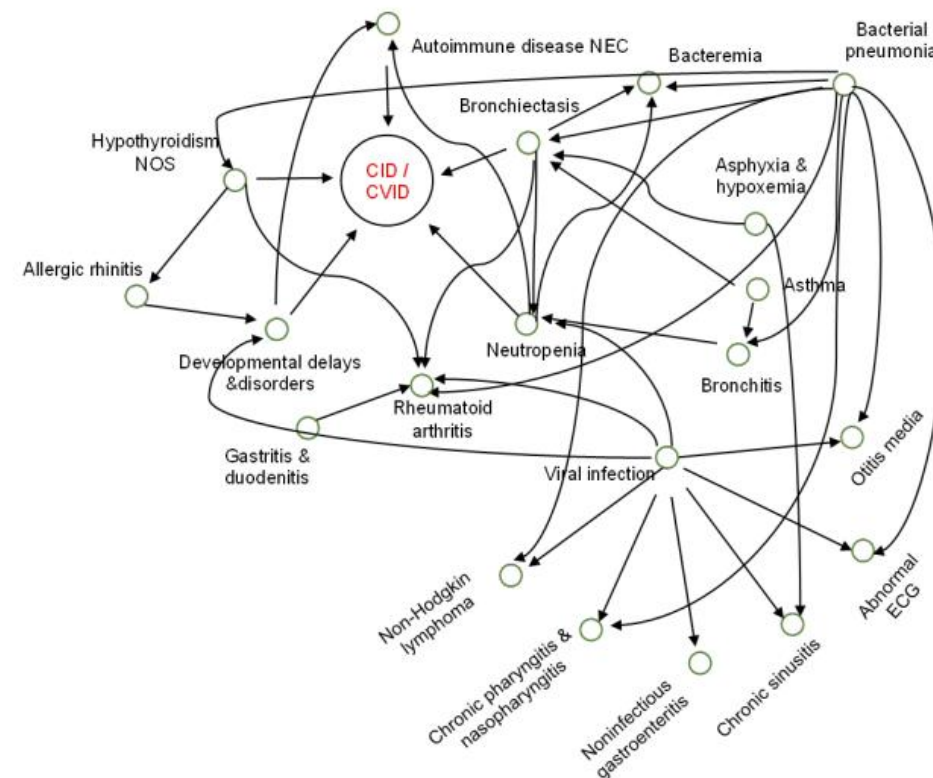
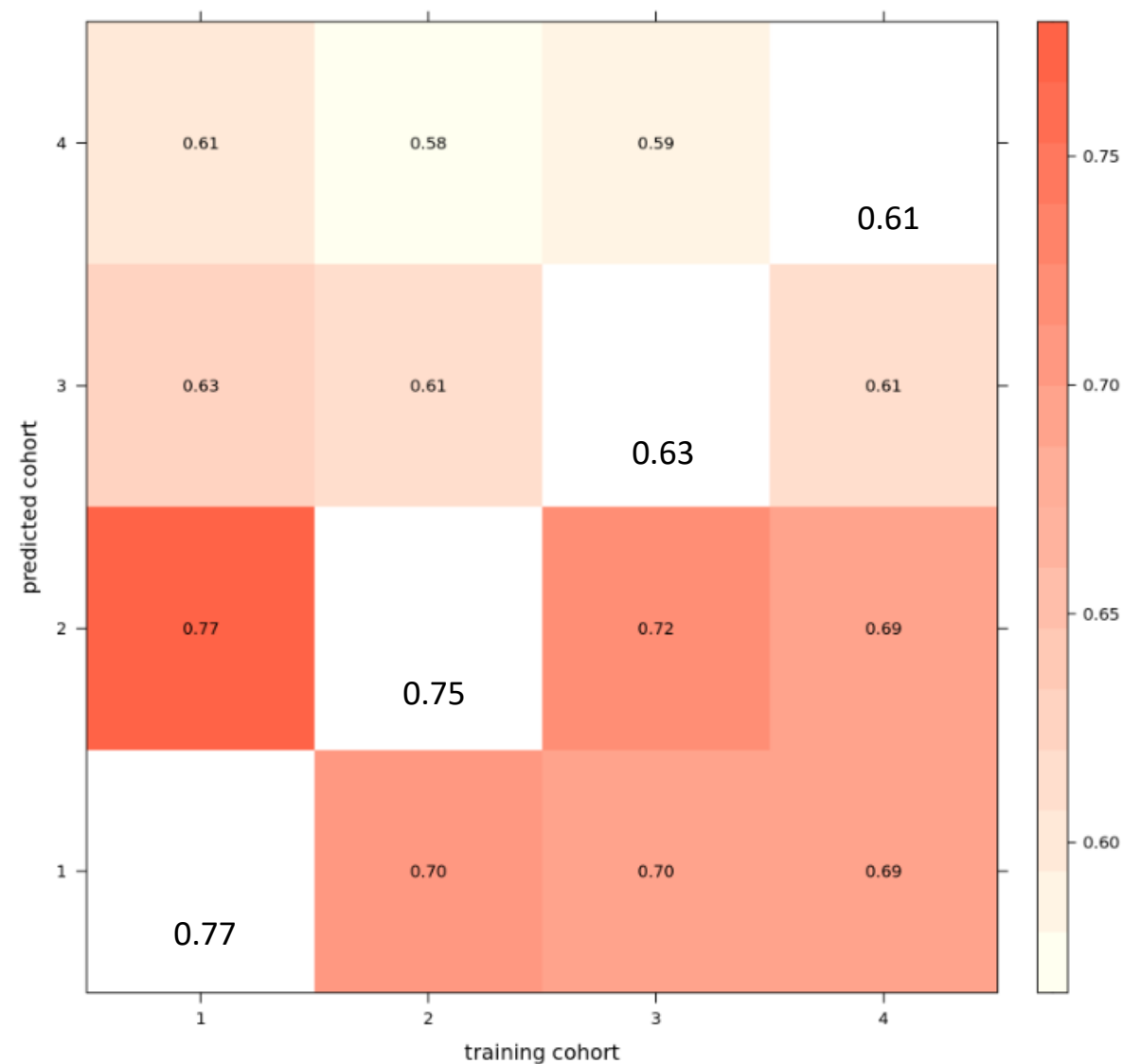
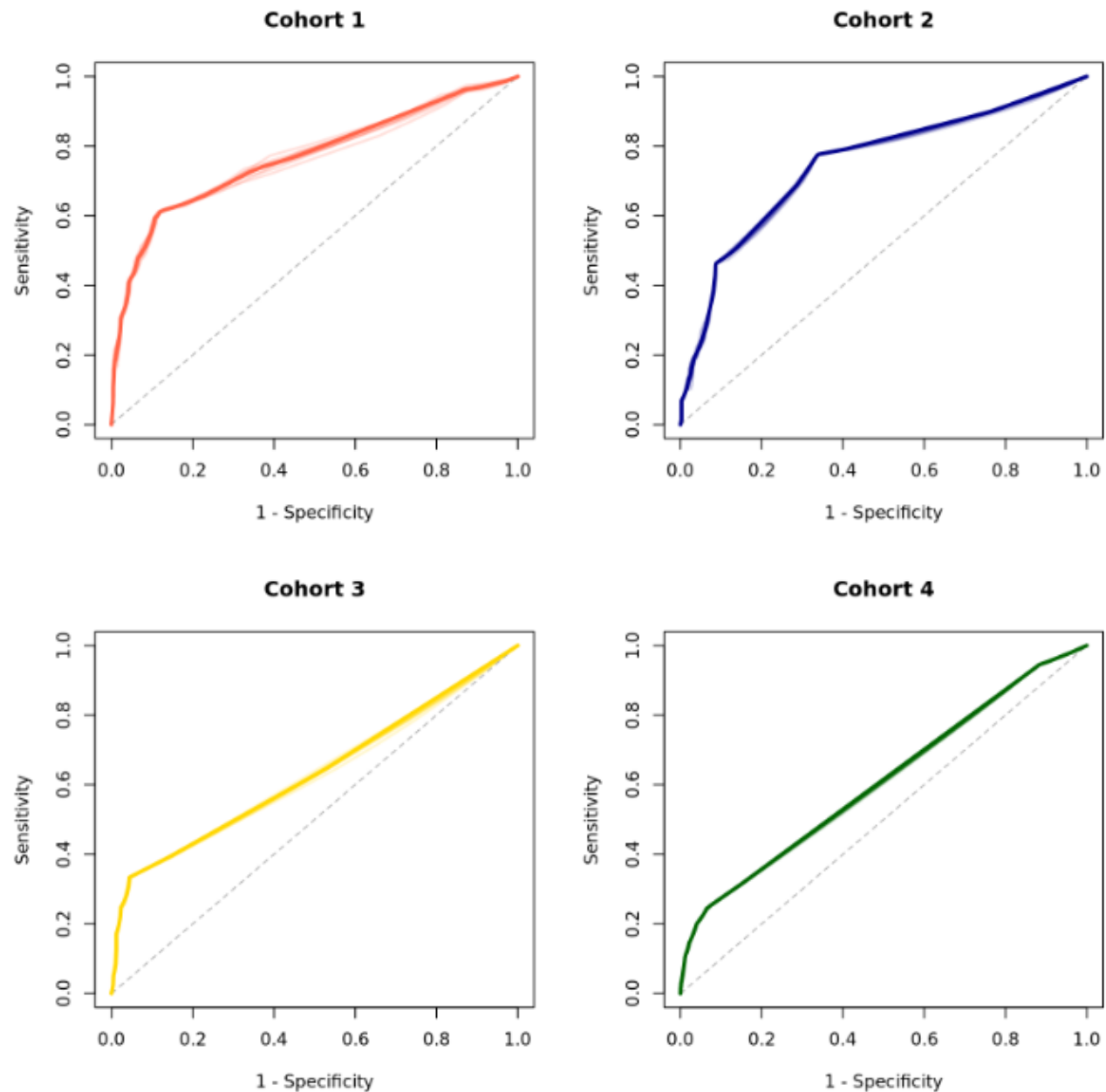


Figure 5) Consensus DAG calculated in Cohort 4. Cohort 4 involved N=19,924 CID and CVID cases (of which 2,350 with pneumonia) and 19,924 matched controls (with no PI), both with and without pneumonia. We visualize up to 2 direct parent levels and up to 2 direct child levels away from CID diagnosis. Up to 1 direct parent level for each direct child and up to 1 direct child level for each direct parent are included. DAG: directed acyclic graph; NEC: not elsewhere classified; NOS: not otherwise specified; IM: immune mechanism; ECG: electrocardiogram; CID: combined immunodeficiency; CVID: common variable immunodeficiency; PI: primary immunodeficiency.

Causal predictions & inference



Causal predictions & inference

Table 4) Temporal information and interventions (causal inference) in Cohorts 3 and 4. Temporal distributions were calculated by considering the first diagnosis of each phenotype in reference to the first CID/CVID diagnosis, in terms of Box and Whisker plots: median value and 50% interquartile range to the median (lower and higher interquartile value to the median, shown in parenthesis). All temporal information (median values and interquartile ranges) is expressed in months. Odds ratio represents the effect of each intervention on the CID/CVID diagnosis. IM: immune mechanism; bcc: blood cell count; NEC: not elsewhere classified.

Phenotype per cohort	Temporal Information (in months)	Intervention (Odds ratio)
Cohort 3		
Failure to thrive	-4.9 (-31.0, 0.7)	23.65
Genetic susceptibility to disease	-4.1 (-30.7, 1.2)	13.24
Disorders involving the IM	-8.2 (-35.2, -0.4)	8.74
Decreased white bcc	-3.03 (-25.4, 0.7)	5.14
Splenomegaly	-4.9 (-24.7, 1.8)	1.49
Nutritional, metabolic, and developmental symptoms	-10.0 (-34.9, 0.9)	1.46
Developmental delays & disorders	-5.8 (-33.0, 6.3)	1.26
Pancytopenia	-5.2 (-29.5, 0.7)	1.22
Pneumonia	-13.7 (-46.6, 0.0)	1.16
Autoimmune disease NEC	-28.8 (-61.2, -2.2)	1.16
Cohort 4		
Bronchiectasis	-4.2 (-29.6, 1.3)	9.44
Autoimmune disease NEC	-23.7 (-50.1, -3.2)	4.36
Neutropenia	-4.9 (-29.8, 0.9)	3.68
Developmental delays & disorders	-7.8 (-39.3, 5.7)	2.25
Bacterial pneumonia	-1.8 (-23.4, 5.8)	1.14
Asphyxia and hypoxemia	-2.1 (-25.3, 6.5)	1.13
Asthma	-27.9 (-62.8, -4.5)	1.10
Bronchitis	-7.8 (-19.8, -0.2)	1.08
Viral infection	-16.5 (-45.8, 0.3)	1.07
Rheumatoid arthritis	-17.2 (-61.8, -0.2)	1.01

- The odds ratio (OR) was calculated **by comparing the odds of diagnosis when the phenotype is present versus the odds when absent**, using the parameters learned by the model during training.
- The OR was computed over the entire graph (inference) in order to assess the effect of the presence or absence of each phenotype on the final probability of diagnosing the target disease.

Exemplar 2 - ORCD: ODE-residual causal decomposition to separate invariant physics from causal forcing

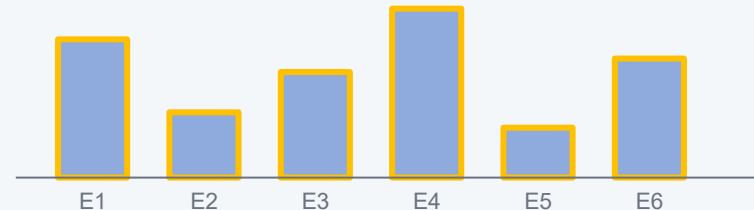
The entanglement problem

- K heterogeneous environments recorded over time (here: 6 English regions, 34 years of daily data) measuring air pollution and cardiorespiratory mortality.
- Slow-scale invariant physics is entangled with fast-scale instantaneous forcing (e.g., acute episodes, sensor and aggregation error).
- Data are close to Gaussian, and available causal discovery machine learning methods have no mechanism to decouple the two regimes, and conflate physics with causality.

1 One global ODE for all environments

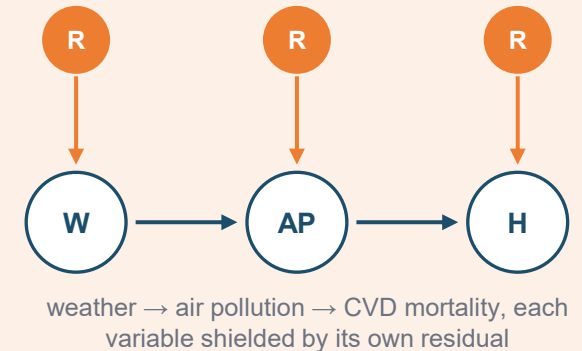
$$\mathbf{R}_{i,k} = d\mathbf{X}_{i,k}/dt - \Theta(\mathbf{X}_k) \Xi_i$$

- A single sparse ODE model is fitted on the pooled data.
- Local dynamics the parsimonious global law cannot represent are pushed into the structured residuals $\mathbf{R}_i$.
- AR and cross-environment whitening then remove autocorrelation while preserving that signal.



2 Augmented 2N-node discovery

- Whitened $\mathbf{R}_i$ enter as observed proxy covariates for environment-specific forcing, not as exogenous noise.
- A self-residual bottleneck and a weather $\rightarrow$ pollution $\rightarrow$ health prior fix every edge orientation.



Assumption 3.3

environments are dynamically distinct



Proposition 3.6

global-fit residuals are heteroscedastic across environments



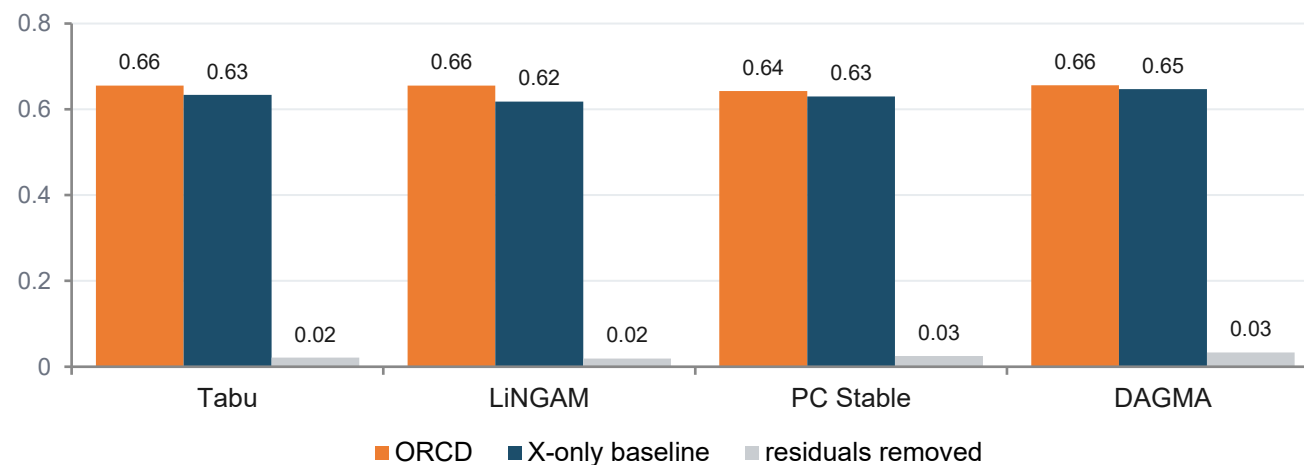
Theorem 3.10

DAG identifiability under Gaussian noise

The heteroscedasticity that makes the graph identifiable is engineered by the global fit.

ORCD in practice: structural gains, and causal signal that transfers

Spatial transfer (Experiment B): CVD60 (cardiovascular mortality rate) Spearman correlation, 30 region pairs



48 / 48

synthetic settings with lower SHD
(4 algorithms × 12 settings)

36 %

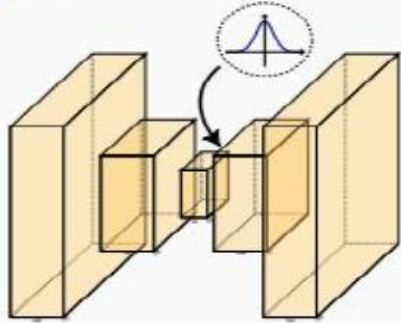
median SHD reduction

F1 0.97-0.99

structural accuracy on real data
(0.90–0.96 without residuals)

Causal Generative AI: full of “causal” promise?

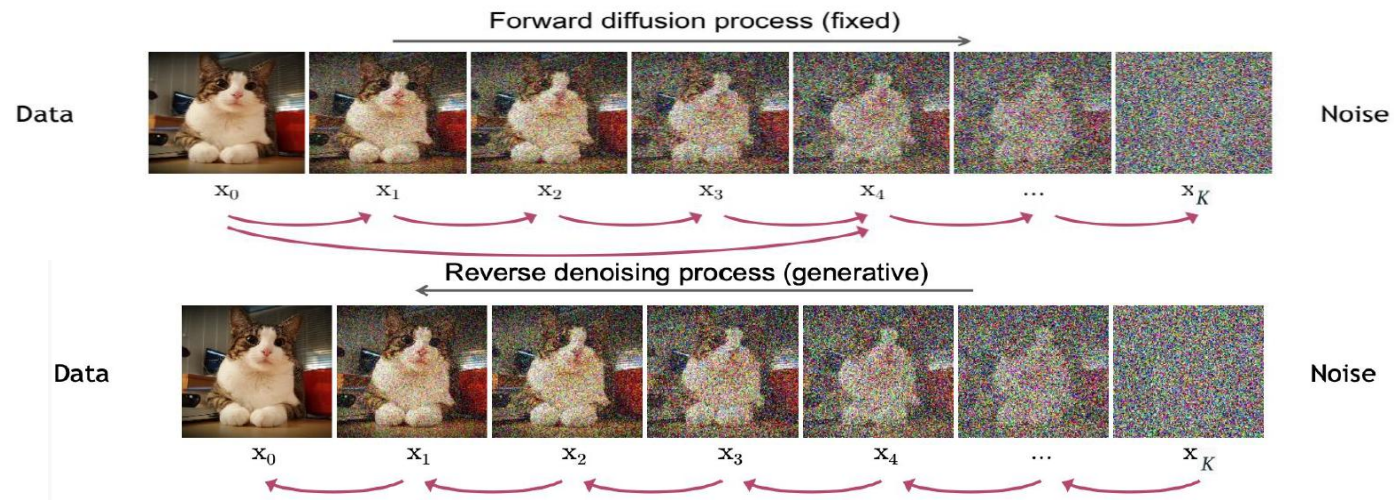
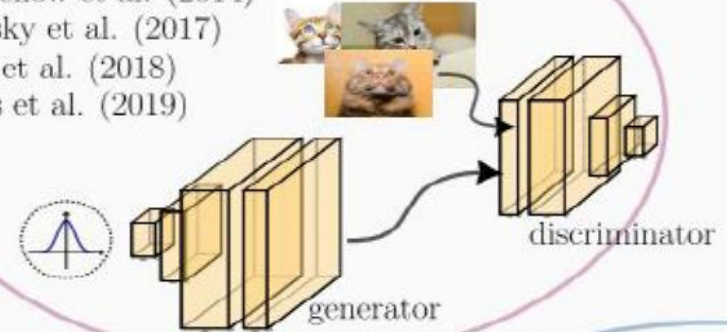
Variational AutoEncoder



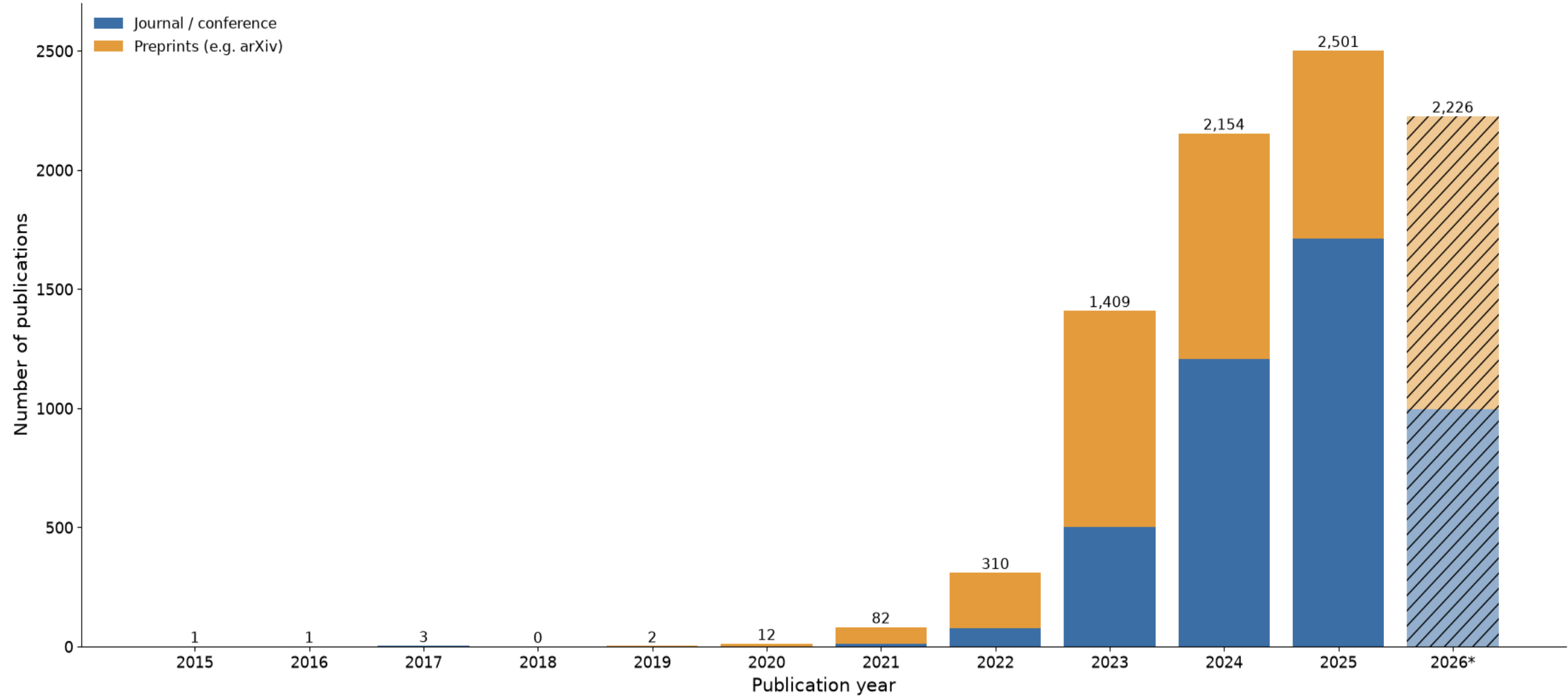
Kingma et al. (2014)
Rezende et al. (2014)
Ranganath et al. (2016)
Vahdat et al. (2021)

Generative Adversarial Network

Goodfellow et al. (2014)
Arjovsky et al. (2017)
Brock et al. (2018)
Karras et al. (2019)

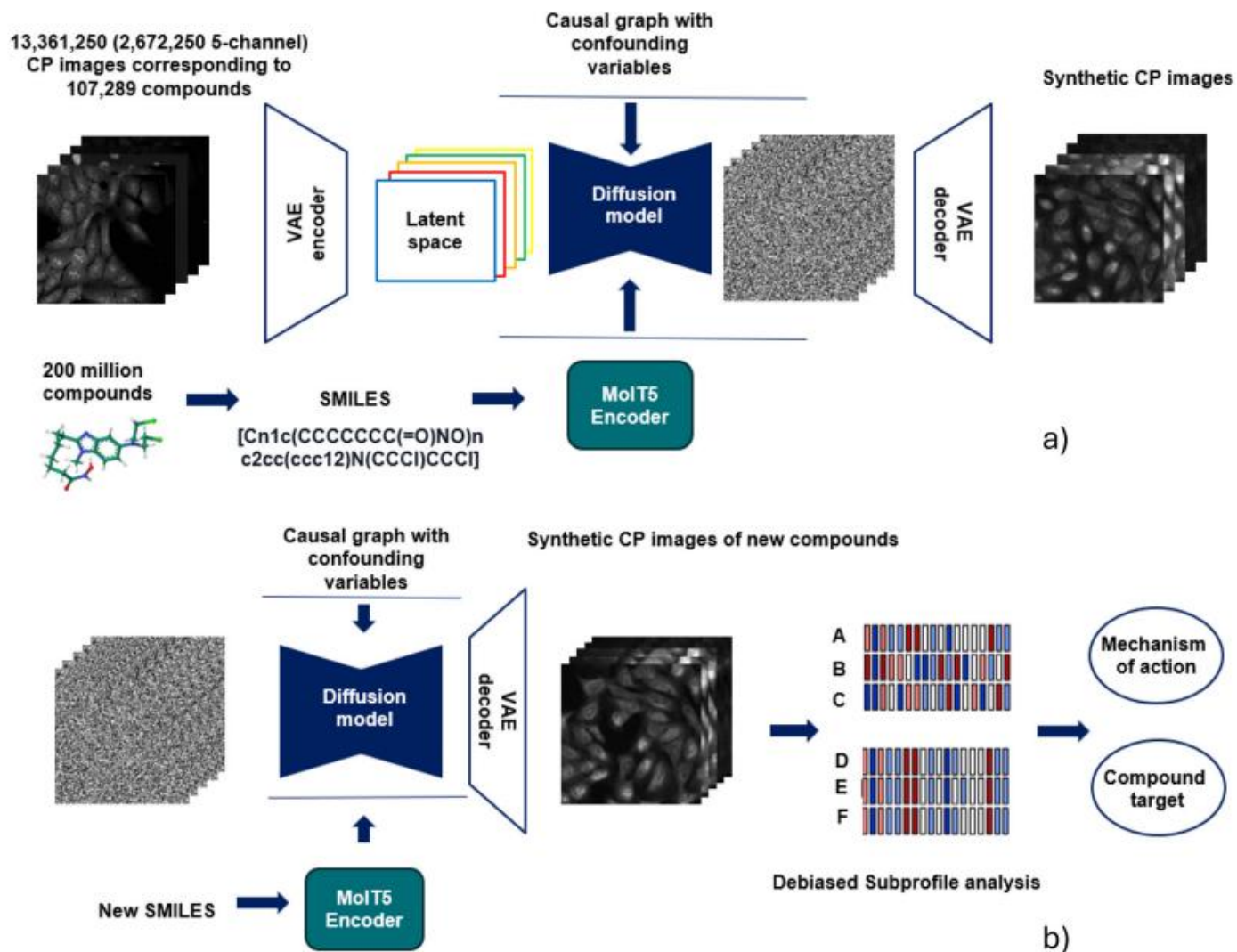


Growth of research on generative diffusion models

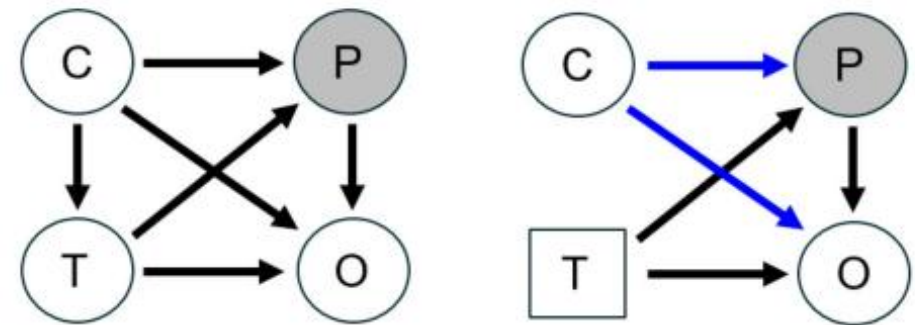
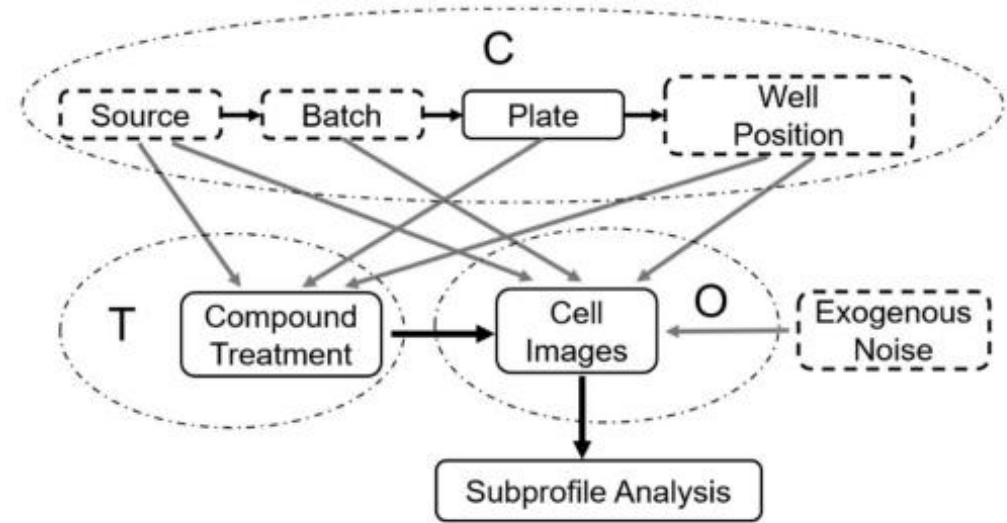


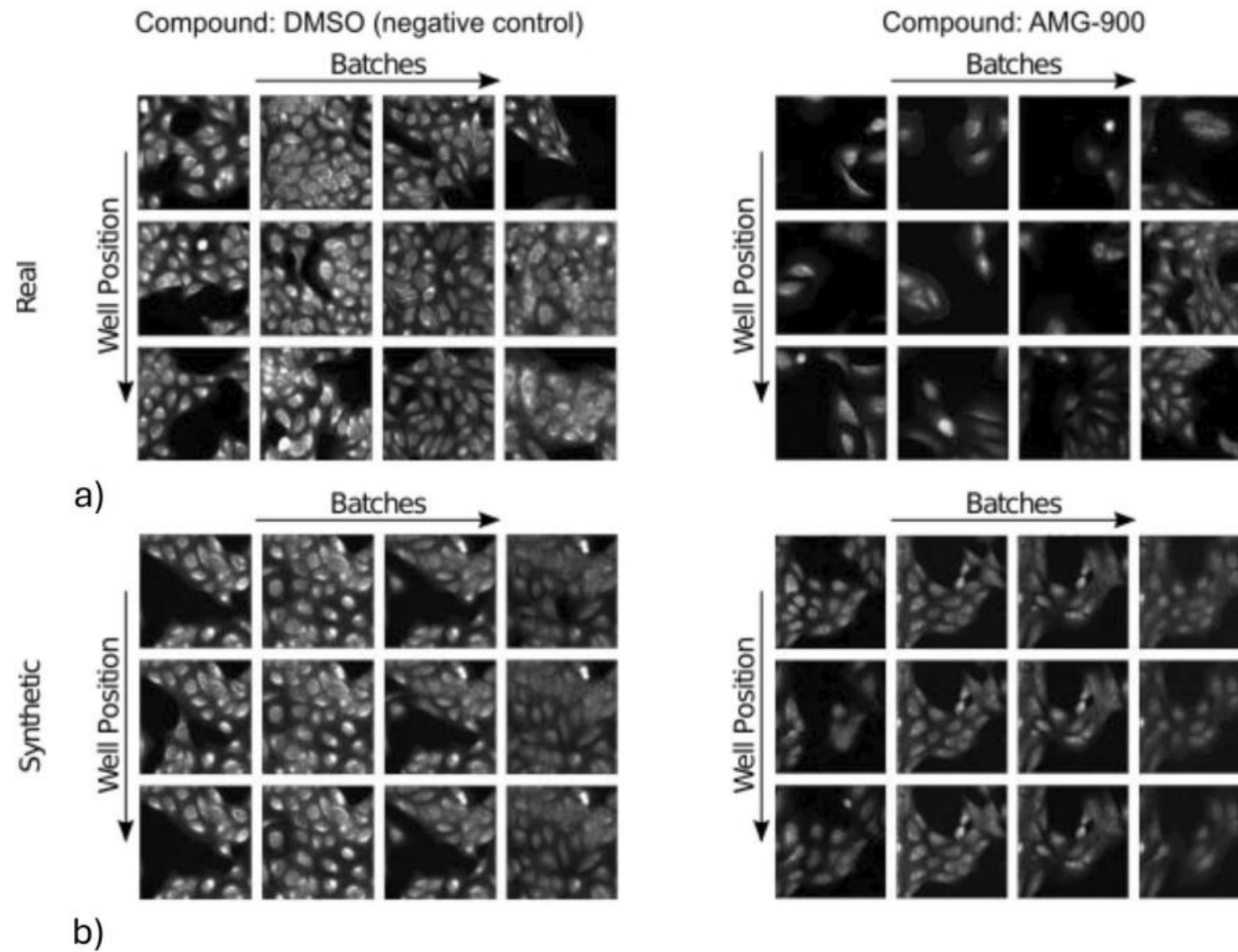
*2026: partial year (data to Sep 2026); recent papers still being indexed.
Source: OpenAlex, title/abstract search, Computer Science & Medicine; specific generative-diffusion terms (e.g. "denoising diffusion", "score-based generative", "latent diffusion model").

Exemplar 3 - Confounder-aware foundation modeling for accurate phenotype profiling in cell imaging

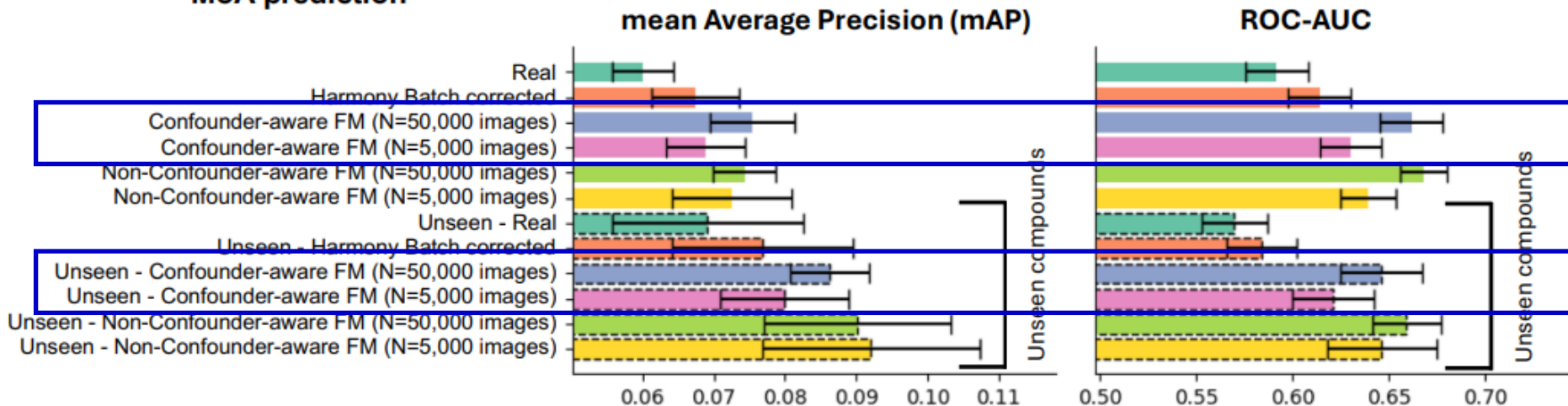


384-well plate for cell imaging

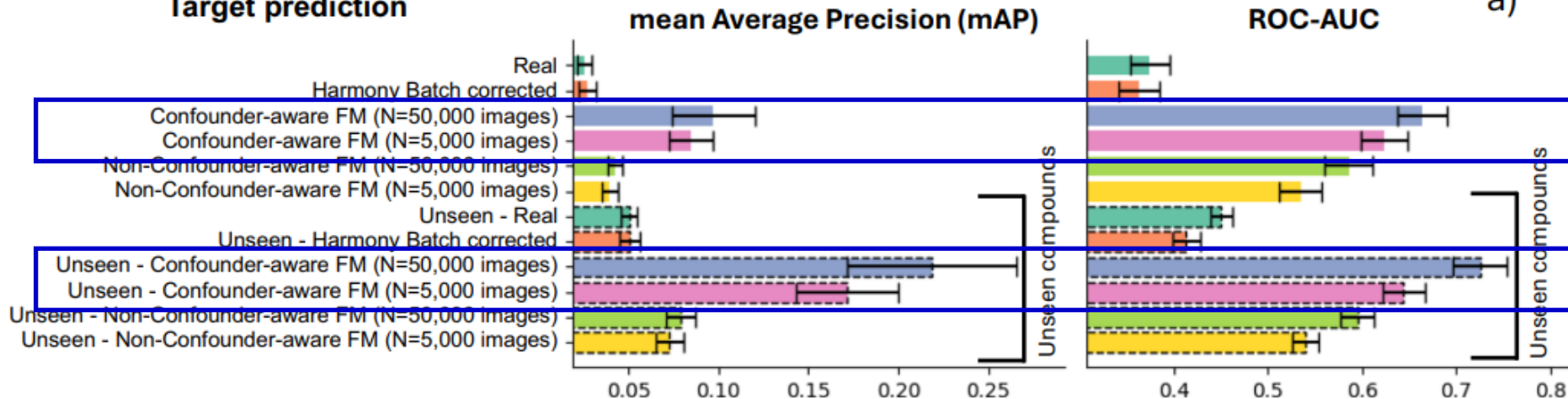




MoA prediction



Target prediction



b)

Causal AI team at Academy of Athens and Archimedes Unit, Athena RC

Nikos Spyrou, PhDc



Angelos Korakitis, MSc



Thanasis Triantafyllou, MSc

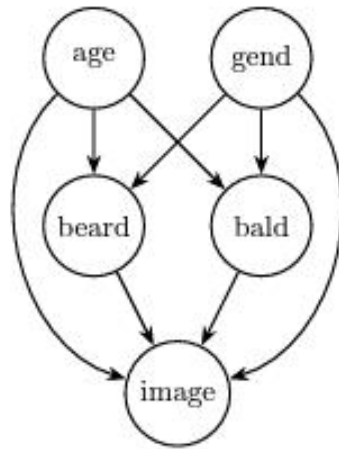


Thomas Melistas, PhDc



Exemplar 4 - Benchmarking counterfactual image generation

Defined DAG



(a) Causal graph

Factual



(b) Factual



(c) Causal

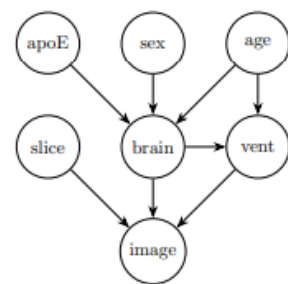


(d) Non-causal

Do(female) counterfactual

Do(young) counterfactual

Figure 1: (a) A plausible causal graph for human faces; (b) factual image (no intervention); (c) causal counterfactual images using the graph in (a) for interventions $do(Female)$ (upper panel) and $do(Young)$ (lower panel); (d) non-causal image editing. Observe that the latter is not realistic as it retains the beard and baldness, whereas they are omitted in (c).



(d) ADNI

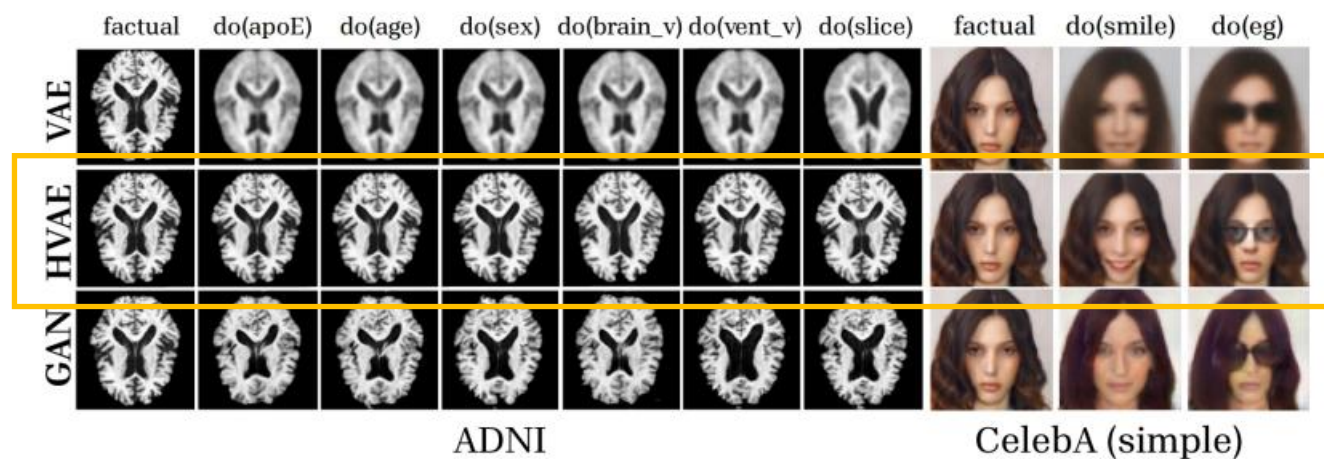
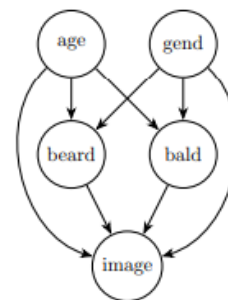


Figure 4: Qualitative evaluation of effectiveness for all datasets/graphs. From left to right across datasets: the leftmost image is the factual one and then each column shows the causal counterfactual image after intervening on a single attribute. v: volume; vent: ventricle; eg: eyeglasses.

Quantitative metrics for counterfactual evaluation

Table 3: Effectiveness on CelebA test set for both graphs.

Model	CelebA (simple)							
	Smiling (s) F1 ↑				Eyeglasses (e) F1 ↑			
	$do(s)$		$do(e)$		$do(s)$		$do(e)$	
VAE	0.897 _{0.02}		0.987 _{0.01}		0.938 _{0.05}		0.810 _{0.02}	
HVAE	0.998 _{0.01}		0.997 _{0.01}		0.883 _{0.06}		0.981 _{0.02}	
GAN	0.819 _{0.02}		0.873 _{0.01}		0.957 _{0.03}		0.891 _{0.01}	
Model	CelebA (complex)							
	Age (a) F1 ↑				Gender (g) F1 ↑			
	$do(a)$	$do(g)$	$do(br)$	$do(bl)$	$do(a)$	$do(g)$	$do(br)$	$do(bl)$
VAE	0.35 _{0.04}	0.782 _{0.02}	0.816 _{0.02}	0.819 _{0.02}	0.977 _{0.01}	0.909 _{0.02}	0.959 _{0.02}	0.973 _{0.01}
HVAE	0.654 _{0.01}	0.893 _{0.04}	0.908 _{0.03}	0.899 _{0.03}	0.988 _{0.02}	0.949 _{0.03}	0.994 _{0.01}	0.95 _{0.03}
GAN	0.413 _{0.04}	0.71 _{0.02}	0.818 _{0.02}	0.799 _{0.01}	0.952 _{0.01}	0.982 _{0.01}	0.92 _{0.01}	0.961 _{0.01}
Model	Beard (br) F1 ↑				Bald (bl) F1 ↑			
	$do(a)$	$do(g)$	$do(br)$	$do(bl)$	$do(a)$	$do(g)$	$do(br)$	$do(bl)$
VAE	0.944 _{0.01}	0.828 _{0.03}	0.296 _{0.05}	0.945 _{0.02}	0.023 _{0.03}	0.496 _{0.05}	0.045 _{0.04}	0.412 _{0.03}
HVAE	0.952 _{0.03}	0.951 _{0.03}	0.441 _{0.11}	0.916 _{0.04}	0.02 _{0.05}	0.86 _{0.05}	0.045 _{0.07}	0.611 _{0.04}
GAN	0.908 _{0.01}	0.838 _{0.02}	0.233 _{0.03}	0.907 _{0.01}	0.021 _{0.02}	0.82 _{0.02}	0.055 _{0.02}	0.492 _{0.02}

Table 4: Effectiveness on the ADNI test set. Note that the three direct causal parents to the image are presented: *Brain volume*, *Ventricular volume* and *Slice number*.

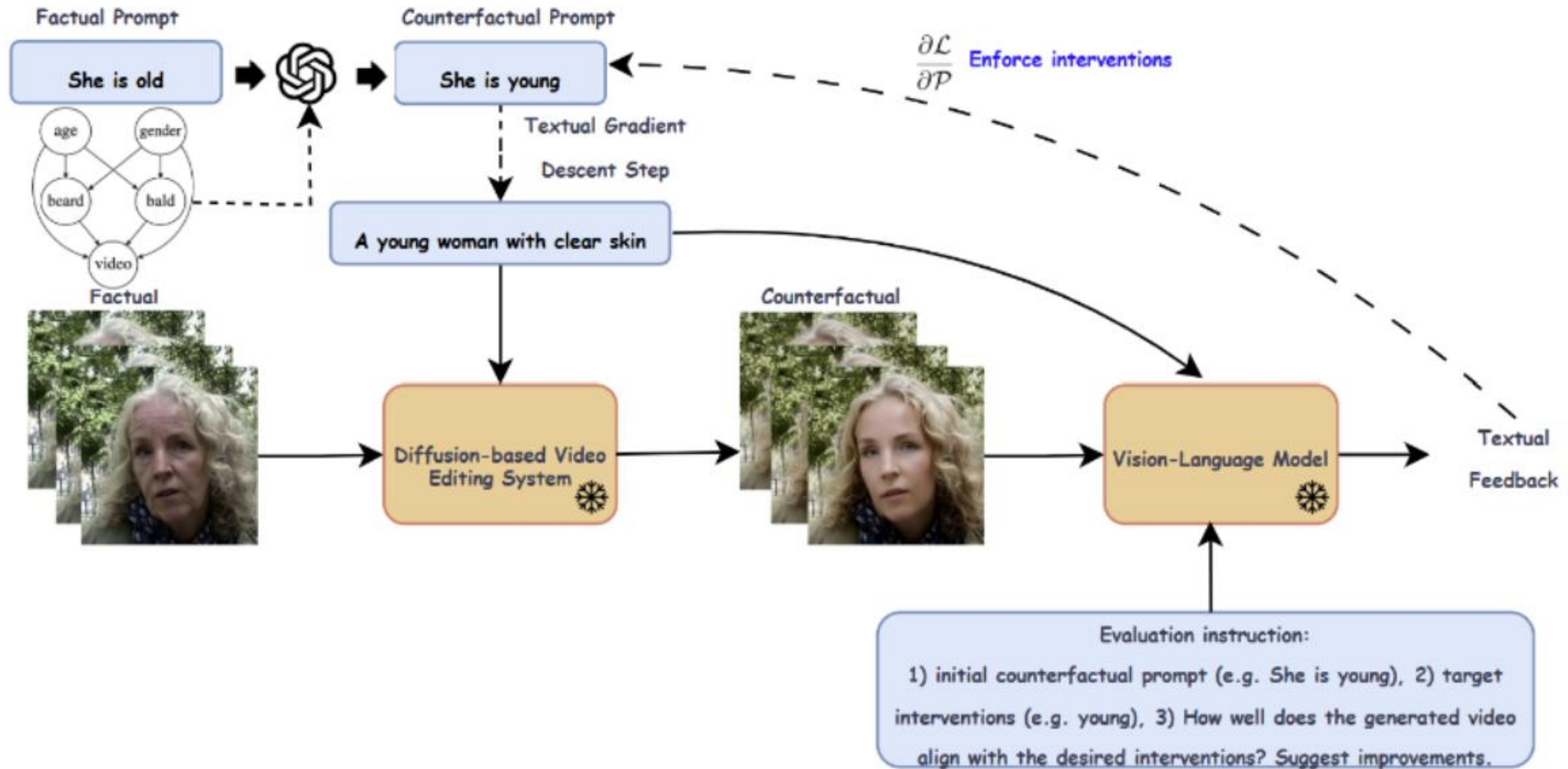
Model	Brain volume (b) MAE ↓			Ventricular volume (v) MAE ↓			Slice (s) F1 ↑		
	$do(b)$	$do(v)$	$do(s)$	$do(b)$	$do(v)$	$do(s)$	$do(b)$	$do(v)$	$do(s)$
VAE	0.17 _{0.03}	0.15 _{0.06}	0.15 _{0.06}	0.08 _{0.05}	0.20 _{0.04}	0.08 _{0.05}	0.52 _{0.15}	0.48 _{0.15}	0.46 _{0.10}
HVAE	0.09 _{0.03}	0.12 _{0.06}	0.13 _{0.06}	0.06 _{0.04}	0.04 _{0.01}	0.06 _{0.04}	0.38 _{0.15}	0.41 _{0.16}	0.41 _{0.11}
GAN	0.17 _{0.02}	0.16 _{0.07}	0.16 _{0.06}	0.12 _{0.02}	0.22 _{0.03}	0.12 _{0.03}	0.14 _{0.03}	0.16 _{0.03}	0.05 _{0.02}

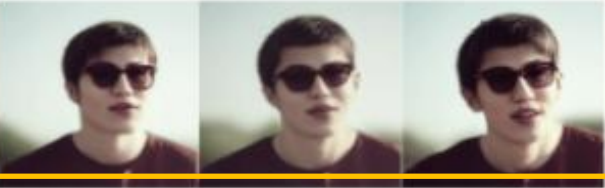
Table 5: Realism (FID) and Minimality (CLD) for MorphoMNIST, CelebA, ADNI.

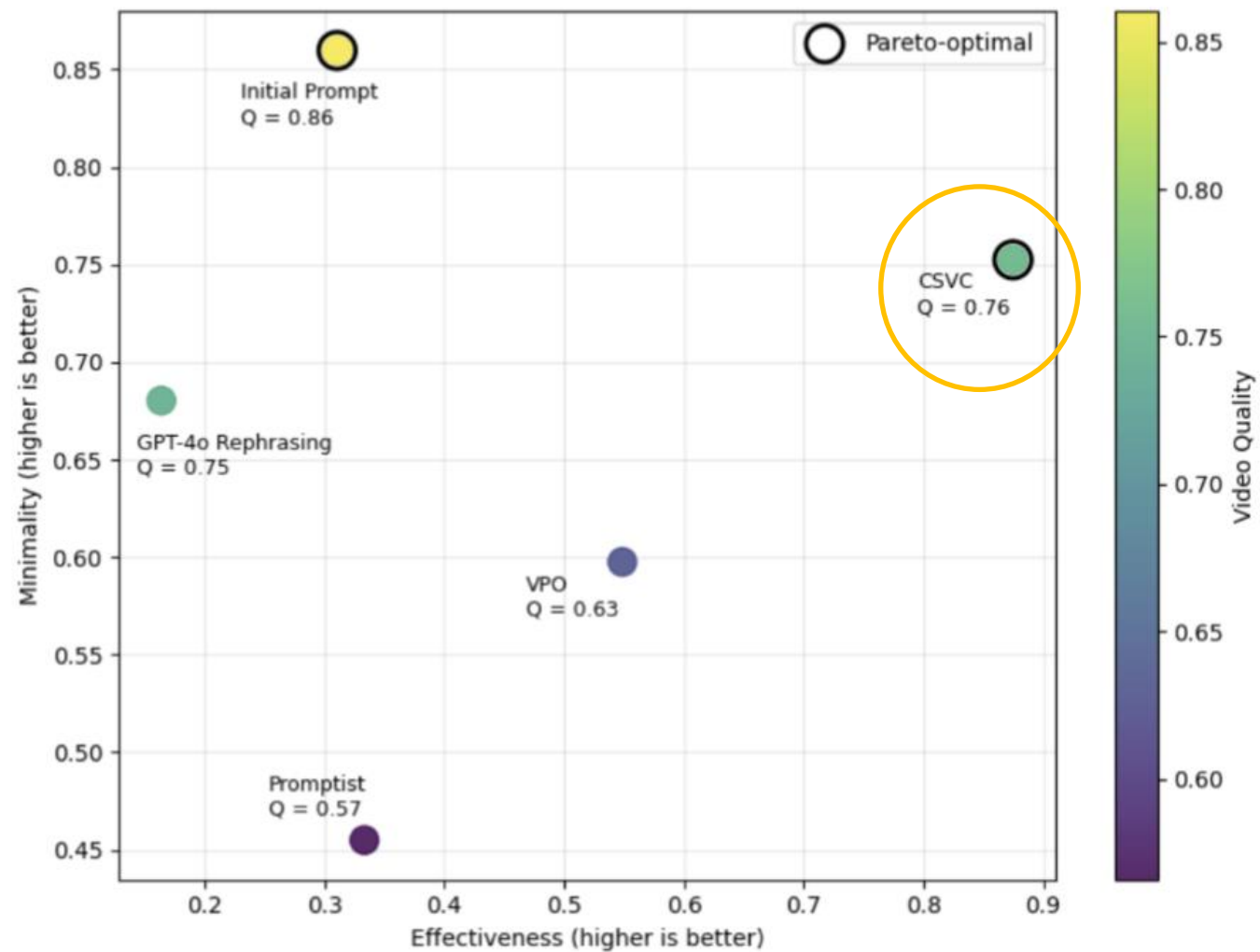
Model	MorphoMNIST		CelebA (simple/complex)		ADNI	
	FID ↓	CLD ↓	FID ↓	CLD ↓	FID ↓	CLD ↓
VAE	10.124	0.268	66.412/59.393	0.301/ 0.299	278.245	0.352
HVAE	5.362	0.272	22.047 /35.712	0.295 /0.305	74.696	0.347
GAN	35.568	0.286	31.560/ 27.861	0.38/0.304	113.749	0.353

Exemplar 5 - Counterfactual video generation via guided diffusion steering

Towards “causal” agents



	TokenFlow	Tune-A-Video	FLATTEN
Factual Video	<i>A man has a beard, he is young.</i> 	<i>This man is old, he has a beard.</i> 	<i>She is young</i> 
Initial Prompt	<i>A woman is young.</i> 	<i>A man is young.</i> 	<i>She is young, she has a beard.</i> 
VPO	<i>A young woman with a radiant complexion and bright eyes is seen wearing a casual outfit, such as a light sweater and jeans.</i> 	<i>A young man with a youthful and energetic appearance is seen in a vibrant outdoor setting under a bright sunny sky.</i> 	<i>A young woman with a well-groomed beard stands in a park under the soft glow of the afternoon sun.</i> 
CSVC	<i>A teenage girl who is clean-shaven.</i> 	<i>This man is a clean-shaved teenager.</i> 	<i>A young individual with a well-groomed beard.</i> 



Exemplar 6 – Identifiable estimation of causal concept effects under visual latent confounding

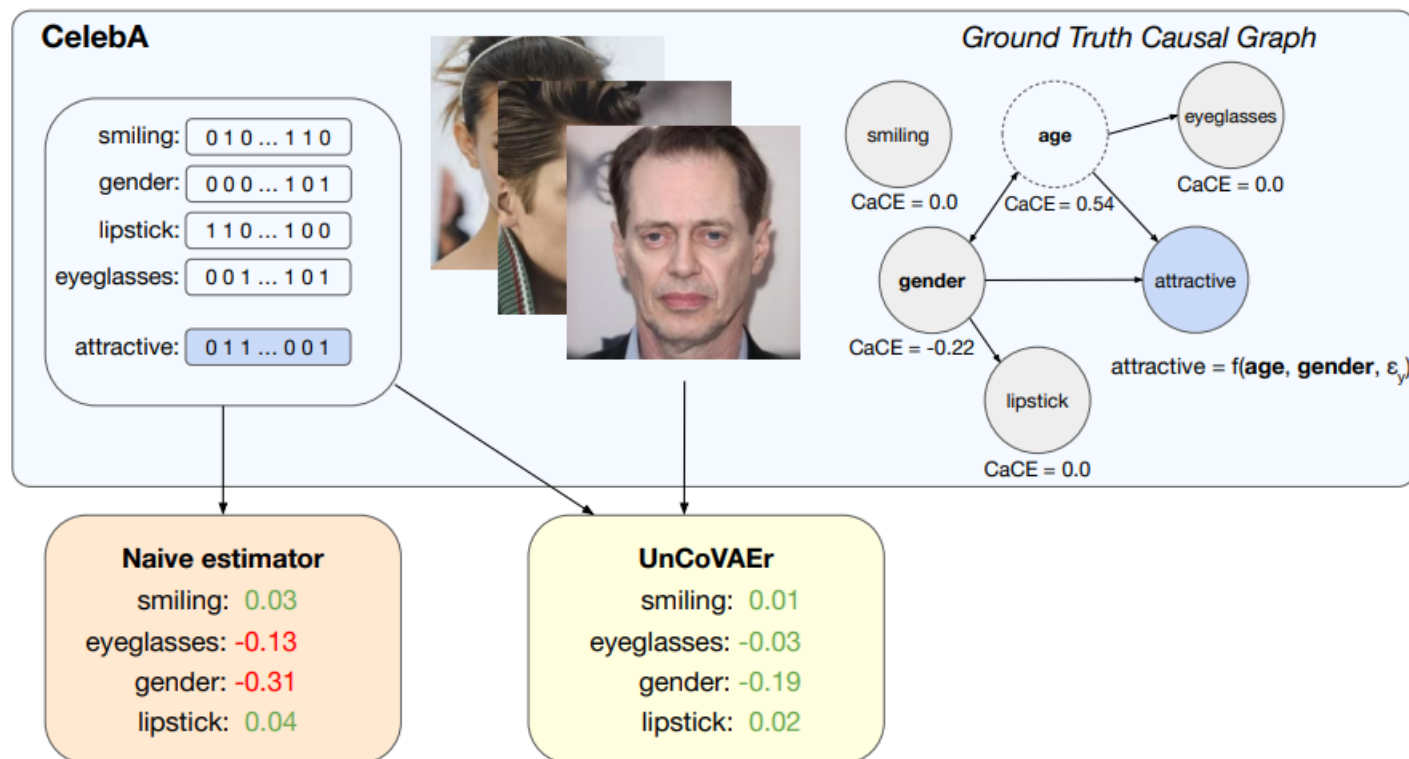


Figure 1: Latent confounders bias causal concept effect (CaCE) estimates. Here, *attractiveness* depends on *age* and *gender*, but *age* is unobserved (dashed node). A naive estimator yields biased CaCE due to open backdoor paths. UnCoVAEr exploits images to learn an *identifiable* latent representation Z that captures confounder-relevant information, substantially reducing bias.

Exemplar 7 – Causal modelling agent

Published as a conference paper at ICLR 2024

CAUSAL MODELLING AGENTS: CAUSAL GRAPH DISCOVERY THROUGH SYNERGISING METADATA- AND DATA-DRIVEN REASONING

**Ahmed Abdulaal^{1*} Adamos Hadjivasiliou¹ Nina Montaña-Brown¹ Tiantian He¹
Ayodeji Ijishakin¹ Ivana Drobnjak¹ Daniel C. Castro² Daniel C. Alexander¹**

¹Centre for Medical Image Computing, UCL, London, United Kingdom

²Microsoft Research, Cambridge

Method – neurosymbolic AI

1 A · Global amendment - no data.

GPT-4 gets the variable names, the graph from the previous iteration and - from the third iteration on - the memory text (see D).

2 B · Local amendment - a pairwise critique, still no data.

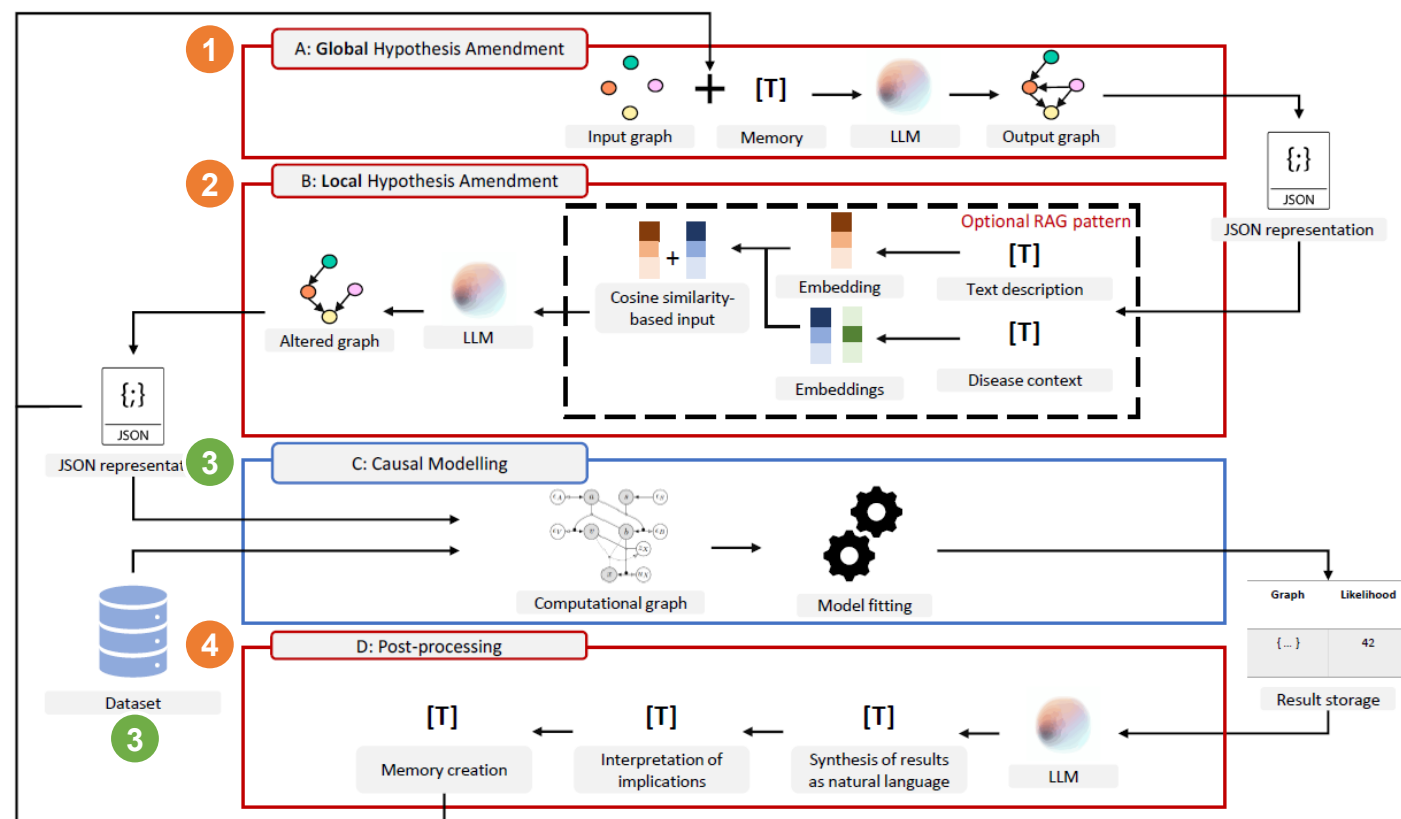
Per pair: keep / remove / reverse, or none / $u \rightarrow v$ / $v \rightarrow u$, reading from a previous iteration. Literature is an optional slot inside this step.

3 C · Causal modelling - where the data enter.

First, with the graph held fixed, one small network (a Normalizing Flow) per variable is trained on the real data, this is where **how strongly each parent acts on its child** is learned. Second, the trained model is asked how probable the observed data are under it, and that single number is **the graph's score**.

4 D · Post-processing - text only.

GPT-4 sees the graph difference and the change in likelihood, and writes the memory. It never sees an observation.



"Metadata" here = what GPT-4 already knows about a variable from its name - it lives in the weights, not in a file.
The data = 11 ADNI variables (age, sex, edu, APOE4, MoCA, brain and ventricular volume, P-tau, AV45, slice index) plus x, the MRI slice itself, which is a node in the same graph.

Results

Improved structural accuracy

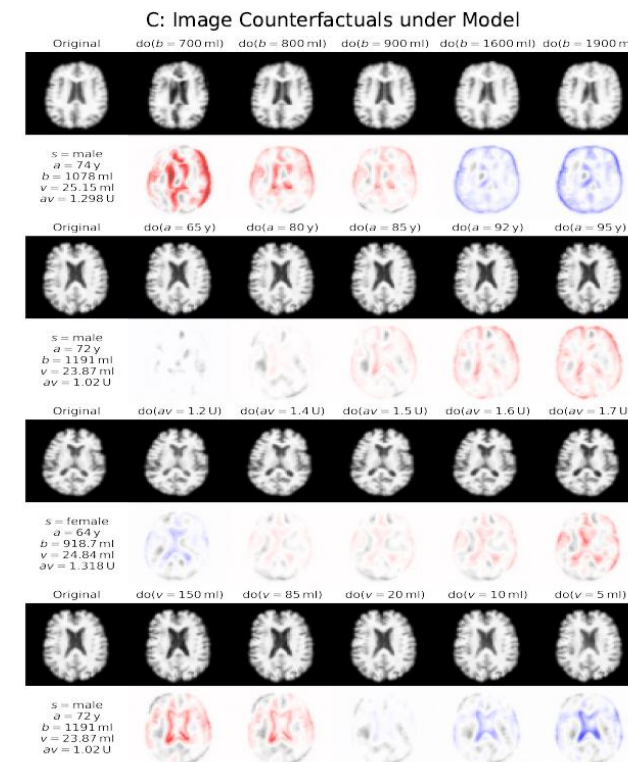
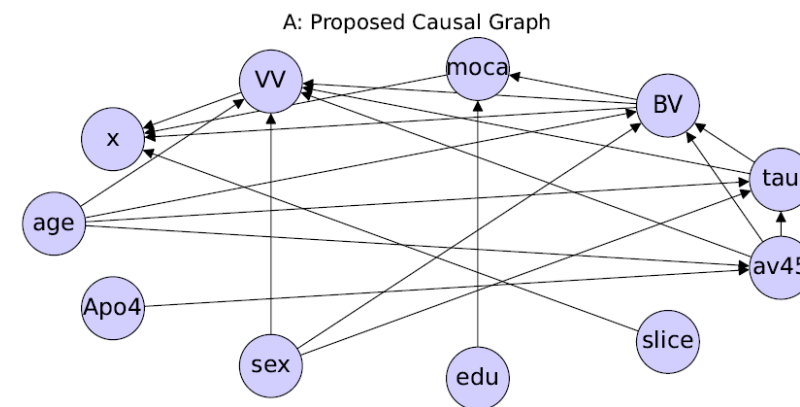
On all three benchmarks (Arctic sea ice, Alzheimer’s disease, Sangiovese vineyards) the CMA beat both families it was tested against: **data-driven discovery - NOTEARS, DAG-GNN, TCDF** - and **LLM-only discovery - GPT-3.5, GPT-4 prompted directly**.

Counterfactuals

The others output a graph and stop. The CMA outputs a trained causal model, so you can intervene on a real patient - do(age), do(brain volume) - and see the expected change in their own MRI. It is also the only one that handles tables and images in a single model.

A new hypothesis on real ADNI data

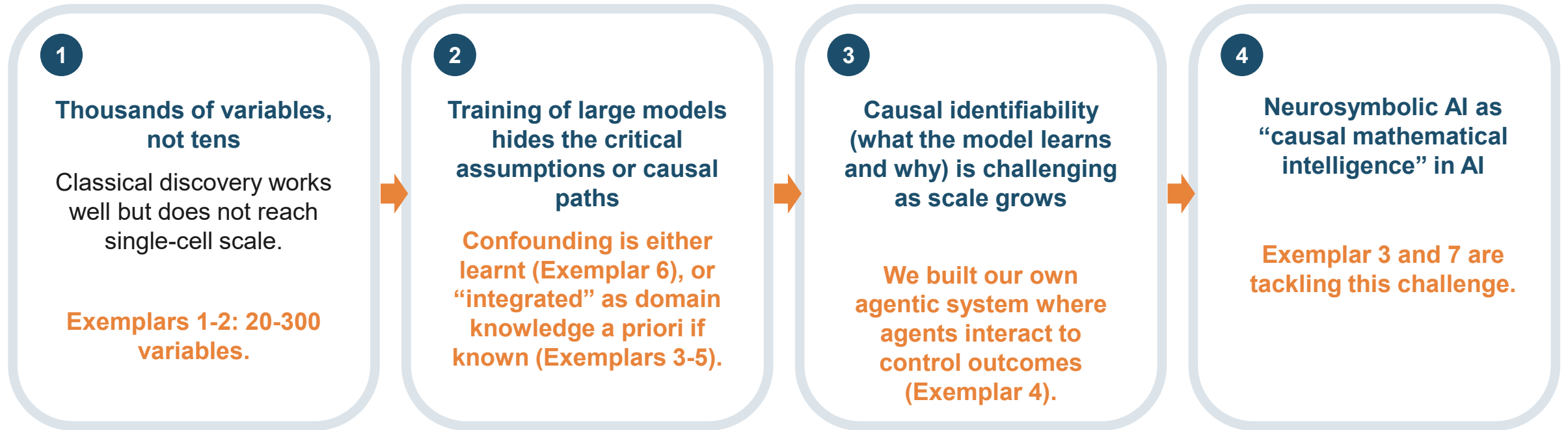
The CMA proposed **sex** → **tau**, drawn by none of the 5 experts; intervening on sex shifts P-tau significantly ($p < 0.001$). It had no literature access when it did so - the corroborating USP11 work is from 2022 and was found later by the authors.



Conclusions: the future of causal & generative AI

- By shifting from standard **"symmetric"** correlation-based AI to **directional causality**, we address fundamental challenges in generalization: deal with data shifts.
- **Causal Bayesian Networks** are the "reference standard" for structural discovery in tabulated data, with many open problems in terms of algorithmic development (out-of-distribution, data misspecifications & missingness, latent confounding).
- **Causal generative AI models** are powerful machines for producing faithful counterfactuals and many opportunities for ongoing research in terms of improving their foundations and applications (out-of-distribution, neurosymbolic AI, interpretability, identifiability).
- Recent breakthroughs demonstrate that **synergizing generative models with causal algorithms**, significantly outperform data specific algorithms on real, out-of-distribution and/ or misspecified data.

Key open challenges: causality meets large models



Take home message: the question is no longer how big the model is, but how to control the learning process as the model scales across data and problems.

Thank you

