CAUSAL GRAPHS: A BLUEPRINT DRIVING DRUG DISCOVERY FROM SINGLE-CELL OMICS

Understanding causality in a cell, mapping causal graphs from gene regulations to cell phenotype inside a patient's microenvironment is a fundamental but daunting challenge. Pinpointing true causal drivers of disease pathology and drug action is vital for drug discovery. Traditional methods focus on statistical correlations. However, as Judea Pearl noted "Correlation is not causation", underscoring the need for explicit causal models.

SINGLE CELL DATA AS THE FOUNDATION FOR CAUSAL DISCOVERY

Causality at the cellular level is difficult to establish because biological systems are highly heterogeneous; a causal link may hold in one cell state, tissue, or patient but not another, and averaging mixed populations can mask or invert true effects. Single-cell omics counters this by capturing each cell's molecular state, enabling observation of gene-activity variation across thousands of cells and supporting inference of directionality. A few 100-plus-million-cell atlases including multiple species and cell lines have been released within the past year.

DILEMMA BETWEEN INTERVENTIONAL AND OBSERVATIONAL DATA

In AI terms, we can approach causal learning in two broad ways: (1) using interventional (perturbation) data to directly induce and observe effects, and (2) relying on purely observational data across multiple conditions with appropriate assumptions. Interventions break spurious correlations and yield explicit cause–effect samples, whereas observational data retain the native patient microenvironments.

Hard (known-target) interventions: Experiments that actively perturb a known factor (eg. CRISPR knockout of a specific gene) impose a well-defined do (X=x) operation. Each perturbed cells then provides a labelled cause-effect pair. Incorporating hard interventions present the most straight forward way to directly identify causal edges since the intervention breaks dependencies implied by noncausal paths.

Soft (unknown-targets) interventions: In other settings, the exact intervention targets are unknown or indirect (e.g. small molecule, stress etc.). These "soft" interventions perturb the system's mechanisms without specifying which variables are directly targeted. Recent theory shows that soft interventions are statistically indistinguishable from multi-environment observational data under certain assumptions [1]. In other words, though still more informative than single environment observational data, soft intervention almost offers no more identifying power than observational data gathered across independent environments.

One challenge of interventional studies is transportability, which is far from default. Because direct perturbations in patients are impossible, researchers use proxy systems (cell cultures, organoids, patient-derived cells), yet causal effects learned there does not transfer to patients natively. Bareinboim and Pearl formalize this— one must encode how the environments differ to know which causal effects can transfer. If the target environment is "drastically different", then "nothing can be learned" from the source experiments without additional assumptions. Additionally, large-scale hard-perturbation datasets remain expensive: total human perturb-seq scRNA-seq datasets top out at low tens of millions of cells— minimally one order of magnitude below observational data. Further scaling will depend on broad collaboration and open data sharing. Companies such as Xaira Therapeutics, Cellarity, and others have been actively building interventional database for novel mechanism discovery and target identification.

Observational Data: Majority of biological data is intrinsically observational, meaning no variables have been actively "do"-perturbed at the time of measurement. These data collected directly in the patient context inherently avoids cross-context mismatch and for causal discovery, there is no domain shift to correct for. However, while these observational data offer clues, traditionally they often recover the adjacency skeleton and remain ambiguous on direction of causal graphs unless strong assumptions are imposed. Recent theoretical development has significantly changed this perception. Pooling purely observational data from diverse environments/domains —different laboratories, tissues, or batches—creates a set of distinct distributions and can break more symmetry and providing additional edge orientation constraints for causal discovery. Under the Markov property, pooling these environments is formally equivalent to analyzing a single domain collection of soft interventions with unknown targets [1]. Though intrinsically, the approach does not demand larger sample counts, in practice more data would be needed to obtain statistical power.

Driven by these insights, Synlico employs AI models using causal-graph representations as the backbone to pinpoint disease-driving interactions, nominate novel targets, and design mechanism-guided therapies, integrating causal discovery and inference, generative modeling, and single-cell bioinformatics. To supply the required diversity and statistical power, we aggregate large-scale primary patient scRNA-seq datasets—more than 100 million cells from over 18 000 patients—as separate observational "domains". This pooled data enables mapping of context-specific gene–gene edges, and prediction of how perturbing a driver gene reverberates through downstream pathways.

DATA VS. AI MODELS: CO-ADAPTIVE SYNERGY

Biotech and AI circles debate whether progress hinges on richer single-cell data or stronger algorithms; in practice, they co-evolve—sound data prevents mis-inference, while better models extend what can be learned. Interventional datasets have long been the gold standard for causal identification, but recent theory has also elevated observational data with natural variation. Because each causal-AI framework rests on distinct assumptions, companies match their stacks to their data type—hard & soft interventions, or purely observational data.

OUTLOOK: TOWARDS CAUSAL MODELS FOR PRECISION BIOLOGY

The field still lacks a standardized single-cell perturbation–response atlas. Such shared benchmarks create a common objective function, and power leaderboard-style iteration that tightens the loop between design and validation in the community. Integrating multi-omics—especially spatial layers—promises to resolve causal ambiguities and move to spatial model. Yet, these are not automatically "better". The experiments tend to produce sparser data for each modality with additional computational complexity for integration. As these tools mature, causal graphs will steer experiment selection within closed-loop "predict-perturb-learn" cycles, shortening the path from hypothesis to mechanism. This is especially valuable for rare diseases: it transfers causal relations learned in larger populations, clarifies which mechanisms generalize, and lowers the number of scarce patient samples needed. Precision medicine will shift to directly manipulating the molecular levers of disease, enabling smaller, smarter trials and a more rational pipeline for targeted therapies design. ■