VIRTUAL CELLS AS CAUSAL WORLD MODELS: A PERSPECTIVE ON EVALUATION

Anonymous authors
Paper under double-blind review

ABSTRACT

This perspective argues that evaluating AI virtual cells requires moving beyond predictive accuracy toward assessing their ability to function as causal world models of biology. Existing benchmarks emphasize fit to observed data, rewarding pattern matching but failing to test responses to interventions. We propose that trustworthy virtual cells require causal evaluation: metrics and benchmarks that assess intervention validity, counterfactual consistency, trajectory faithfulness, and mechanistic alignment. Our contribution is two-fold: (1) a survey of recent approaches to virtual cell modeling, and (2) a taxonomy of causal evaluation metrics mapped to available perturbation datasets and benchmarks. By outlining gaps and proposing unified causal benchmarks, we position causal evaluation as the key step toward making virtual cells reliable world models of biology.

1 INTRODUCTION

Modern biology sits at a crossroads: even with the complete genetic code and vast single-cell atlases such as the Human Cell Atlas (Regev et al., 2017), CELLxGENE (Program et al., 2025), Tahoe-100M (Zhang et al., 2025), and scPerturb (Peidli et al., 2024), our ability to predict cellular responses to drugs, mutations, or environmental change remains limited (Wen et al., 2023; Rood et al., 2024). The bottleneck is not lack of data, but models that fail to capture how biological systems actually work (Glocker et al., 2021; Listgarten, 2024). Recent biological Foundation Models (FMs) such as GeneFormer (Zheng & Gao, 2023), Evo2 (Brixi et al., 2025), scFoundation (Hao et al., 2024), and AIDO (Ellington et al., 2025) show impressive predictive power, but are often limited to a single biological layer and capture associations rather than causal mechanisms. *Despite their variety, these models remain predictive rather than causal, with evaluations centered on accuracy or likelihood rather than causal validity*. Some advanced models fail to outperform simple linear baselines (Rood et al., 2024; Peidli et al., 2024). Efforts like PerturBench (Peidli et al., 2024) have begun to standardize predictive benchmarking, but there is still no equivalent of ImageNet (Deng et al., 2009) or GLUE (Wang et al., 2018) for causal evaluation. Biology is hierarchical (i.e., genome, transcriptome, proteome, metabolome, phenome), and disregarding this interdependence yields models that are fundamentally misaligned with the underlying biology (Kitano, 2002; Hood & Flores, 2012).

This raises the motivating question: *When does a predictive model of cells become a true world model, able to answer counterfactuals and generalize beyond its training data?* Because no dataset fully captures the multilayered complexity of the cell, uncertainty is an inherent property of both biological systems and virtual models. Evaluation must therefore address not only whether a prediction is correct, but also how confident we should be in that prediction. To this end, our vision of AI virtual cells is simulation-ready representations that reason about mechanisms, predict perturbation responses, and serve as in silico testbeds (Bunne et al., 2024; Carr et al., 2024; Noutahi et al., 2025). These can be thought of as biological world models, not just reproducing observed data but answering "what if" and "how" questions.

Our contribution is two-fold: (1) a survey of recent approaches to virtual cell modeling, and (2) a taxonomy of causal evaluation metrics mapped to available perturbation datasets and benchmarks (Figure 1). We do not prescribe how to build causal virtual cells; rather, we argue that without principled evaluation, progress toward trustworthy, mechanistic virtual biology will remain directionless.

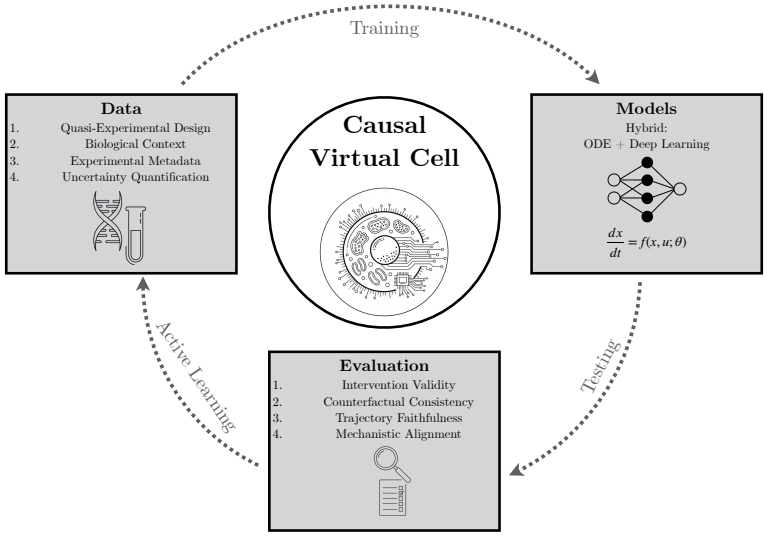

Figure 1: Summary of our proposed framework, which is described in Section 4.

2 Related Work: Predictive Approaches

Predictive approaches to virtual cell modeling aim to reproduce observed cell states or transitions rather than identify or test cause–effect relationships. In this section, we review predictive models, the data used to create them, and how they are evaluated, before outlining their key limitations.

2.1 Models

Autoencoder-based and conditional architectures such as scGen (Lotfollahi et al., 2019), CPA (Lotfollahi et al., 2021), GEARS (Roohani et al., 2024), scCade (Ou et al., 2024), and scPerb (Tang et al., 2024) interpolate from control to perturbed states, while models like Biolord (Piran et al., 2024), CoupleVAE (Wu et al., 2025), SAMS-VAE (Bereket & Karaletsos, 2023), scVI (Lopez et al., 2018), and CRADLE-VAE (Baek et al., 2025b) enhance latent representations or capture combinatorial and differential perturbations. Other architectures include MichiGAN, which applies GANs for disentangled single-cell generation (Yu & Welch, 2021), CellFlow, which uses flow-matching for phenotype modeling (Klein et al., 2025), and CellOT, which applies optimal transport to map cellular trajectories (Bunne et al., 2023). Beyond autoencoders and GANs, diffusion models (Ho et al., 2020; Song et al., 2020) have been adapted for imputation, denoising, and simulation tasks.

Biological FMs build on the same generative principles but distinguish themselves by scale and scope: they are pretrained on millions of cells or sequences and fine-tuned across diverse downstream tasks, enabling broader transferability. However, these evaluations remain predictive, emphasizing reconstruction accuracy or classification performance. **Genomic/DNA** FMs are trained on DNA sequences to understand regulatory functions and predict genetic outcomes. They enable variant effect prediction, with Enformer (Avsec et al.) and Geneformer (Theodoris et al., 2023) tackling regulatory variants, and EVO2 (Brixi et al., 2025) achieving breakthrough performance on noncoding pathogenicity. FMs learn RNA sequence–structure relationships for tasks such as RNA structure/function prediction (RiNALMo (Penić et al., 2025), HydraRNA (Li et al., 2025a)), mRNA design (mRNA-FM (Li et al., 2025c)), and RNA modification site detection (AIDO.RNA (Zou et al., 2024)). **Protein** FMs predict structures, attributes, and guide design. Beyond structure, they estimate stability and binding affinity, critical for therapeutic applications. Use cases include de novo protein design (ProtGen (Madani et al., 2023), ProtGPT2 (Ferruz et al., 2022)), structure prediction (ESM-2 (Lin et al., 2023), ESM-3 (Hayes et al., 2025), MSA-transformer (Rao et al., 2021), Boltz-2 (Passaro et al., 2025), AlphaFold 3 (Abramson et al., 2024)), and single-sequence property prediction. **Single-cell** FMs analyze omics data, often across modalities, to model cellular states. Applications include cell type annotation (scBERT (Yang et al., 2022), scGPT (Cui et al., 2024), CellFM (Zeng

et al., 2025)), batch correction (scPRINT (Kalfon et al., 2025b)), and perturbation response prediction (scFoundation (Hao et al., 2024), CellFM (Zeng et al., 2025)). **Multi-modal** FMs are emerging to unify layers. SCARF integrates scRNA-seq and scATAC-seq (Liu et al., 2025), LucaOne unifies DNA, RNA, and protein (He et al., 2024), and ChatNT frames genomic tasks as text-to-text (Richard et al., 2024). Simulation-aware models such as scMultiSim (Li et al., 2025b) and Xpressor (Kalfon et al., 2025a) capture cross-modality dynamics. These highlight progress beyond single layers, but integration remains challenging, with evaluation still dominated by predictive accuracy rather than causal benchmarks.

2.2 Data

The datasets highlighted here are widely used in virtual cell modeling, supporting training and evaluation of models that capture cell states or transitions without testing causal mechanisms. **Large-Scale Atlases.** Projects such as Tahoe-100M (Zhang et al., 2025) and Parse-PBMC (Parse Biosciences, 2023), now provide internally consistent datasets with millions to over one hundred million cells. The success of baseline-only efforts such as Tabula Sapiens (Quake & Consortium, 2024) and the Human Cell Atlas (Regev et al., 2017) highlights the importance of coordination across tissues and donors. Aggregation initiatives such as scBaseCount (Youngblut et al., 2025) and CELLxGENE (Program et al., 2025) have further created large harmonized resources by systematically combining hundreds of smaller public datasets. **Synthetic Data Generators.** Splatter (Zappia et al., 2017), SymSim (Zhang et al., 2019), and scDesign3 (Song et al., 2024) are increasingly used to generate controlled transcriptomic data for benchmarking predictive models. Predictive virtual cell models sometimes integrate **Clinical and Phenotypic Data**. The Cancer Genome Atlas (TCGA) (Tomczak et al., 2015), for example, has been used to link single-cell embeddings to tumor states (Tao et al., 2019; Chu et al., 2022), while UK Biobank (Bycroft et al., 2018) and EHR-derived datasets like the All of Us Research Program (All of Us Research Program Investigators, 2019) support predictive modeling of disease risk or treatment outcomes.

2.3 Evaluation

Evaluation in predictive virtual cell modeling relies on established metrics and strategies that measure how well models reproduce observed cell states or transitions. We organize these into metrics that assess predictive fit (e.g., sequence modeling, classification, perturbation response) and on strategies that give these metrics meaning through baseline comparisons and generalization tests.

2.3.1 Metrics

Predictive virtual cell models are typically evaluated using scalar metrics that quantify how well model outputs match observed data, and provide the basic measures of predictive fit. **Sequence modeling** metrics are used by Evo 2 as a proxy for evaluating how well the predicted biological sequence distribution matches the ground truth, as we all as a measure of gene essentiality (Brixi et al., 2025). **Sequence classification** metrics are computed by RNA FMs for distinguishing the introns vs. exons regions, splice variations, and splice variations (Chen et al., 2022a). Categorical cross entropy, for instance, can be used to assess the predicted output distribution in relation to labels from the set of class labels. **Epigenome prediction** is a common task which requires predicting expression values to compare to the ground truth expression values (Lotfollahi et al., 2019; 2021; Roohani et al., 2024; Ou et al., 2024; Tang et al., 2024). Mean absolute error, mean squared error (MSE), R^2, and cosine similarity are commonly used as metrics for regressing continuous expression values. Detection metrics are applied to the prediction of genetic interaction (Roohani et al., 2024). **Subcellular localization** evaluates predictions of spatial cell properties by comparing a set of predicted, labeled 2D Euclidean clusters to the ground-truth labeled cellular subcomponenets. Adjusted rand index (ARI) and adjusted mutual information are used to evaluate the SubCell (Gupta et al., 2024), and average probability of correct label is used to evaluate DeepProfiler (Tomkinson et al., 2024) and CellProfiler (Stirling et al., 2021). **Macroscopic cell state detection** (of cell type and cell health, for example) is also commonly used as a benchmark for virtual cell models (Brixi et al., 2025). This involves comparing the predicted per-state logistic labels, to the binary ground truth labels. Typical retrieval metrics are employed by (Brixi et al., 2025), such as recall, precision, F1, ROC-AUC, etc. Similarly, drug Mechanism-of-Action (MoA) is also used as detection benchmark in the same fashion. **Epigenome delta prediction** evaluates models on their ability to

predict perturbation outcomes. Here, epigenetic deltas are extracted from the model and compared against the ground truth deltas. Fold-change, log fold-change, DE overlap accuracy, directionality agreement, Wilcoxon rank-sum, and Top-k precision are commonly applied metrics in this setting (Adduri et al., 2025; Ou et al., 2024; Tang et al., 2024).

2.3.2 STRATEGIES

Evaluation strategies define how scalar metrics are applied to assess model capability and generalization. Metrics provide raw measures of predictive fit, while strategies organize them into benchmarks, baseline comparisons, and ablations that guide model selection and assess genuine progress. **Rank based metrics.** As noted by PerturBench (Wu et al., 2024), scalar metrics on epigenome prediction often wash out signal and may encourage effects like "mode collapse." Rank-based interpretations (e.g., Log-FC, cosine similarity) better capture differences and align with a common use of virtual cell models: ranking perturbations by effect size.

Calibration. Many virtual cell models (e.g., scGen (Lotfollahi et al., 2019), CPA (Lotfollahi et al., 2021), GEARS (Roohani et al., 2024)) are probabilistic, making calibration crucial. Measuring calibration helps weight predictions by uncertainty and build trust. Negative log-likelihood can be used for sequence metrics, while Expected Calibration Error (ECE) applies to classification tasks (Naeini et al., 2015).

2.4 LIMITATIONS OF PREDICTIVE APPROACHES

Predictive frameworks excel at interpolating and extrapolating trajectories but remain black boxes that lack mechanistic explanations (Moran & Aragam, 2025). They perform well on held-out data yet struggle to generalize to unseen perturbations or conditions (Jiao et al., 2024; Tejada-Lapuerta et al., 2025) and they predict outcomes without testing causal guarantees or answering counterfactual questions (Laubach et al., 2021). These limits reflect the data: most resources are observational or transcriptome-only with few true interventions (Rawal et al., 2025); multi-omic and temporal datasets remain scarce (Carr et al., 2024); and scRNA-seq yields only static snapshots, preventing before–after comparisons (Noutahi et al., 2025). Most datasets capture a single layer, leaving genome-to-proteome mechanisms unevaluated, while the combinatorial complexity of perturbations demands coordinated community efforts (Tejada-Lapuerta et al., 2025).

Evaluation is likewise dominated by predictive metrics such as MSE, R^2, and log-likelihood, which capture correlations but not mechanisms (Goshisht, 2024). Even perturbation benchmarks emphasize regression measures, insufficient for mechanistic alignment (Noutahi et al., 2025). Newer metrics, uncertainty quantification (e.g., calibration error (Yao et al., 2019)), distributional similarity (e.g., MMD (Gretton et al., 2012)), and rank-based evaluation (e.g., LogFC rank in PerturBench (Peidli et al., 2024)) are progress but still treat predictions as point estimates. In practice, uncertainty guides how results are used: low-confidence predictions signal the need for more data or refinement, while high-confidence results provide greater justification for moving forward. Uncertainty is therefore a cross-cutting dimension of evaluation, shaping how validity, consistency, and mechanistic alignment are interpreted.

3 CAUSAL METHODS

Compared to predictive methods that reproduce observed patterns, causal models aim to capture cause–effect relationships and are judged on whether they reproduce intervention outcomes or generate counterfactuals consistent with known mechanisms (Zanga et al., 2022; Carr et al., 2024; Pearl, 2012; Bareinboim & Pearl, 2016; Niu et al., 2024). See Figure 2 for a visual comparison of predictive and causal approaches. Causal machine learning offers a path forward by treating perturbations as structured interventions and seeking mechanisms invariant across environments (Glymour et al., 2019; Tejada-Lapuerta et al., 2025). Causality in biology can be defined in complementary ways. (i) A *mechanistic* view emphasizes biochemical interactions and dynamical processes (e.g., MAPK phosphorylation cascades that link receptor activation to downstream gene expression) (Tejada-Lapuerta et al., 2025). (ii) A *probabilistic* view emphasizes conditional independences in observational data (e.g., ERK activation being independent of receptor status once Ras activity is accounted for) (Glymour et al., 2019). (iii) A *counterfactual* view highlights potential outcomes under

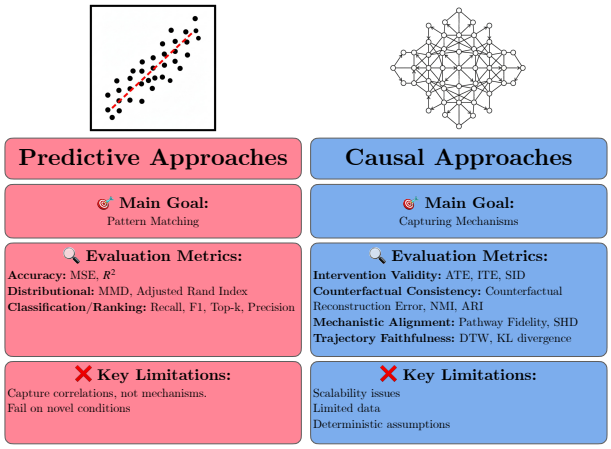

Figure 2: Comparison of predictive (Section 2) and causal (Section 3) approaches.

interventions (e.g., asking how a tumor cell's transcriptome would change if KRAS were knocked out versus left intact) (Lobentanzer et al., 2024).

3.1 Causal Models

Structural Causal Models (SCMs) represent variables as directed graphs with explicit rules for interventions via the do-operator (Pearl, 2012; Rawal et al., 2025). Dynamical causal models extend this framework, using ordinary or stochastic differential equations to describe how biological states evolve under perturbation (Tejada-Lapuerta et al., 2025). Together, these perspectives form the foundation for causal virtual cells: models that not only predict cellular responses but also explain them in terms of mechanisms that remain invariant across conditions.

We highlight four broad families of causal models relevant to virtual cells. Early causal models in systems biology and pharmacology used **ODE-Based Models.** Early causal models used ODEs to describe biochemical networks (Kitano, 2002; Alon, 2019), with classical examples in electrophysiology and metabolism (Hodgkin & Huxley, 1952; Strassberg & DeFelice, 1993; Noble, 1960; Courtemanche et al., 1998; Higgins, 1964; Heinrich & Rapoport, 1974; Adamczyk et al., 2011). Recent methods adapt to single-cell gene regulatory networks (GRN) inference and trajectories: SCODE (Matsumoto et al., 2017), GRISLI (Aubin-Frankowski & Vert, 2020), SINCERITIES (Papili Gao et al., 2018); RNA-velocity extensions include scVelo (Bergen et al., 2020), UniTVelo (Gao et al., 2022), Velorama (Singh et al., 2024), DynaMO (Kuang et al., 2018). GraphDynamo (Zhang et al., 2023) and STORM (Peng et al., 2024) add graph/stochastic structure. SDEs help with noise but raise scalability/identifiability challenges (Komorowski et al., 2011; Browning et al., 2020; Persson et al., 2022). **Hybrid Causal Deep Learning Models** address the scalability limits of mechanistic models by integrating neural networks. Neural ODEs (Chen et al., 2018) flexibly parameterize dynamics, Latent ODEs (Rubanova et al., 2019) extend this to hidden states, and Universal Differential Equations (UDEs) (Rackauckas et al., 2020) embed neural nets within ODEs to learn unknown processes while preserving structure. In single-cell biology, DeepVelo (Chen et al., 2022b) extends RNA velocity with neural ODEs, PerturbODE (Lin et al., 2025) models perturbation dynamics, and PHOENIX (Hossain et al., 2024) integrates lineage information. Knowledge-primed neural networks, including sparse MLPs and graph-informed architectures, constrain learning with pathway priors (Fortelny & Bock, 2020). **Graphical and Counterfactual approaches** represent cellular dependencies as Directed Acyclic Graphs (DAGs) or SCMs. Causal discovery methods aim to recover such graphs: constraint-based algorithms such as Peter-Clark and Fast Causal Inference (Spirtes & Glymour, 1991; Spirtes et al., 2000), score-based approaches like GES (Chickering, 2002), and differentiable DAG learners including NOTEARS (Zheng et al., 2018), DAG-GNN (Yu et al., 2019), and GraN-DAG (Lachapelle et al., 2019). Applications to single-cell data remain early but are grow-

ing: CausalCell (Wen et al., 2023) integrates multiple strategies for GRN inference, LINEAGEOT (Forrow & Schiebinger, 2021) combines lineage tracing with optimal transport, and CARDAMOM (Yuan & Duren, 2025) applies a Bayesian SCM-inspired framework. Invariance-based methods, including ICP (Peters et al., 2016), Causal Dantzig (Rothenhäusler et al., 2019), and anchor regression (Rothenhäusler et al., 2021), identify gene modules stable across environments. **Causal Perturbation Prediction models** embed causal structure into predictive architectures, enabling counterfactual simulation and enforcing invariance. scCausalVI (An et al., 2025) disentangles baseline heterogeneity from treatment effects using a variational inference framework guided by SCM principles, while CausCell (Gao et al., 2025) combines SCMs with diffusion-based generative modeling to generate counterfactual single-cell states. CINEMA-OT (Dong et al., 2023) leverages independent component analysis and optimal transport to separate confounding from treatment effects. Other methods extend this paradigm: GPO-VAE (Baek et al., 2025a) aligns VAE latent spaces with GRN priors, GraphVCI (Wu et al., 2022) predicts counterfactual responses on graphs, and DCD-FG (Lopez et al., 2022) infers factor graphs with causal constraints.

3.2 Causal Data

Unlike predictive resources in Section 2.2, which are mostly observational or baseline-only, causal modeling requires datasets with explicit interventions, perturbations, or synthetic counterfactuals. These form the basis for testing whether models capture cause–effect relationships rather than correlations. Perturbation datasets provide the closest analogue to randomized controlled trials in cell biology (Laubach et al., 2021). High-throughput CRISPR-based screens such as Perturb-seq (Dixit et al., 2016; Adamson et al., 2016) and its large-scale extensions (Replogle et al., 2022), as well as Optical Pooled Screens (OPS) (Feldman et al., 2019), have become cornerstones of interventional single-cell data. Recent large-scale initiatives such as X-Atlas (Huang et al., 2025) extend Perturb-seq to tens of millions of cells, providing a reference-scale atlas of genetic perturbations that could serve as a benchmark for causal modeling. These datasets enable direct measurement of how cellular systems respond to interventions, though they remain sparse, noisy, and limited to subsets of possible perturbations. For causal inference, single-modality measurements (e.g., transcriptomes) are often insufficient, as mechanisms span multiple regulatory layers. Emerging perturbational datasets incorporate multi-omic readouts, including joint measurements of RNA and protein (perturbational CITE-seq (Stoeckius et al., 2017; Hao et al., 2021)), and chromatin accessibility (Perturb-ATAC (Rubin et al., 2019)).

3.3 Evaluation

Evaluation of causal virtual cells requires metrics and strategies that assess whether models capture underlying mechanisms, respect known biological pathways, and generalize to unseen interventions.

3.3.1 Metrics

Intervention Validity measures whether the model reproduces observed outcomes under experimental interventions (e.g., CRISPR knockouts, drug perturbations). This can be tested through effect size and causal effect estimation, including (Individual, Average, Conditional Average) Treatment Effects (ITE, ATE, CATE; respectively), and population-level summaries such as log Fold-Change (LogFC) (Hill, 2011; Shalit et al., 2017; Winship & Morgan, 1999; Hernán & Robins, 2006). Attribution accuracy and regression coefficients further quantify whether responses are correctly attributed to latent or confounding factors (Johansson et al., 2016; Louizos et al., 2017; Schölkopf et al., 2021). Distributional alignment metrics complement these by comparing predicted and observed interventional outcomes: Structural Intervention Distance (SID, which measures discrepancies in interventional distributions (Sachs et al., 2005; Peters & Bühlmann, 2013)) (Hauser & Bühlmann, 2012), Maximum Mean Discrepancy (MMD) (Gretton et al., 2012), energy distance (Székely & Rizzo, 2013), and cluster-preservation indices (e.g., ARI (Hubert & Arabie, 1985)) assess how well causal structure and response space are preserved. **Counterfactual Consistency** quantifies whether counterfactual predictions are biologically plausible, mechanistically grounded, and consistent with both simulated and experimental causal effects. Evaluation involves: (i) *counterfactual reconstruction error*, which compares predicted states against observed perturbation responses, using metrics such as Pearson correlation, MSE, Normalized Mutual Information (NMI), ARI, and marker gene preservation (Gayoso et al., 2022; Lotfollahi et al., 2019); (ii) *latent disentan-*

glement scores, to assess how causal factors are separated in latent space, and are quantified via clustering and silhouette-based indices (Bengio et al., 2019; Gao et al., 2025; An et al., 2025); and (iii) *agreement with ground-truth*, benchmarked against (semi)synthetic datasets such as GeneNetWeaver (Schaffter et al., 2011), SynTReN (Van den Bulcke et al., 2006), PerturBench (Wu et al., 2024), and real-world intervention datasets (e.g., Sachs flow cytometry, Perturb-seq). **Trajectory Faithfulness** measures alignment between predicted and observed time-resolved responses. Perturb-seq, OPS, and synthetic benchmarks such as DREAM4 (Greenfield et al., 2010) and SynTReN (Van den Bulcke et al., 2006) provide the experimental and simulated foundations for evaluating trajectory faithfulness. Evaluation metrics include: (i) *trajectory similarity*, using Dynamic Time Warping (DTW), KL divergence of state distributions, and optimal transport to compare predicted versus experimental temporal profiles (Cuturi, 2013; Chen et al., 2018); (ii) *trend alignment*, where Pearson correlation, MSE, and RMSE quantify concordance between predicted and observed expression dynamics, including treatment effects (Lotfollahi et al., 2019; An et al., 2025); and (iii) *structural consistency*, such as SID and causal graph recovery scores assess whether perturbation trajectories follow known pathways (Peters et al., 2016). **Mechanistic Alignment** quantifies overlap between known pathways and mechanistic constraints. Evaluation includes: (i) *pathway fidelity scores*, measure overlap between inferred interactions and curated databases such as KEGG (Kanehisa, 2002) and Reactome (Fabregat et al., 2018), and assess whether models recover literature-supported mechanisms; (ii) *invariance tests*, evaluate the stability of causal predictions across perturbations, using conditional independence checks, out-of-distribution generalization, and robustness to modality or context shifts (Peters et al., 2016; Heinze-Deml et al., 2018); and (iii) *causal graph similarity*, using metrics like SID and Structural Hamming Distance (SHD). **GRN Recovery** tests whether models recover both the statistical associations and causal intervention structure underlying biological regulatory graphs like GRNs. Standard measures include: (i) *edge prediction accuracy*, with AUROC and AUPR quantifying discrimination between true and false regulatory edges across thresholds; (ii) *graph distance metrics*, such as SHD and SID; and (iii) *benchmark datasets*, including DREAM4 challenges (Greenfield et al., 2010) and GeneNetWeaver (Schaffter et al., 2011) simulations, which provide community standards for comparing GRN inference methods.

3.3.2 Strategies

Causal evaluation strategies define how metrics are applied to probe causal validity. They specify the setups, tasks, and comparisons that reveal whether models generalize beyond observed data. **Synthetic ground-truth tests** enable precise quantification of GRN recovery and counterfactual consistency. Simulation frameworks such as GeneNetWeaver (Schaffter et al., 2011), SERGIO (Dibaeinia & Sinha, 2020), and scDesign3 (Song et al., 2024) generate datasets with known causal graphs. **Pathway fidelity tasks** evaluate whether models preserve mechanistic structure by testing predicted perturbation effects against curated pathways (e.g., KEGG (Kanehisa, 2002), Reactome (Fabregat et al., 2018), BioModels (Le Novere et al., 2006)). **Invariance-based evaluation** tests whether predictions remain stable across environments or cell contexts, using causal discovery frameworks such as ICP (Peters et al., 2016) and anchor regression (Rothenhäusler et al., 2021). **Generalization regimes** are examined under unseen single perturbations, novel combinations, and temporal holdouts. These tasks parallel predictive benchmarks but require causal consistency rather than fit alone (Arjovsky et al., 2019; Lotfollahi et al., 2019; Schölkopf et al., 2021; An et al., 2025). For **baselines and ablations**, causal models are compared against predictive-only baselines (e.g., scGen (Lotfollahi et al., 2019), CPA (Lotfollahi et al., 2023)), to test whether causal inductive biases improve counterfactual validity. Component ablations (e.g., removing causal regularizers, pathway priors) clarify which features drive causal performance (An et al., 2025; Gao et al., 2025). **Perturbation benchmarks** enable systematic evaluation of interventional datasets (e.g., Perturb-seq, OPS). Frameworks such as PerturBench (Wu et al., 2024) and OP3 (Szałata et al., 2024) provide standardized tasks for perturbation response prediction, with OP3 emphasizing causal evaluation criteria such as intervention validity and counterfactual prediction. **General causal benchmarks** include broader efforts such as CausalBench (Wang, 2024), which provide reference standards for evaluating causal inference methods across domains, including perturbation modeling.

3.4 Current Limitations of Causal Approaches

Causal models for virtual cells provide interpretability and mechanistic grounding but remain limited by strong assumptions and scalability issues (Bunne et al., 2024; Carr et al., 2024; Lan et al.,

2025; Noutahi et al., 2025). Many ODE-based and hybrid methods assume acyclicity or causal sufficiency (Michoel & Zhang, 2023; Wen et al., 2023; Tejada-Lapuerta et al., 2025), restricting feedback loops and hidden confounders. They also rely on idealized interventions and face unresolved parameter identifiability challenges (Klipp & Liebermeister, 2006). Consequently, most approaches remain confined to small circuits, velocity-style embeddings, or low-dimensional summaries rather than genome-wide, multi-omic contexts (Glymour et al., 2019; Lobentanzer et al., 2024; Lan et al., 2025). Causal data availability remains a bottleneck (Carr et al., 2024). Perturbation assays such as Perturb-seq and OPS expand access to interventional data but are sparse, noisy, and context-biased. Ground-truth causal graphs are rare, temporal measurements limited, and destructive assays like scRNA-seq prevent before–after comparisons. Synthetic benchmarks help but cannot fully capture biological complexity or generalize to real systems (Cheng et al., 2022). Evaluation remains fragmented: efforts emphasize GRN recovery, pathway fidelity, or counterfactual validation, but no unified taxonomy of causal metrics exists for virtual cells (Bunne et al., 2024). Most evaluations also treat outcomes as deterministic, even though biological systems and perturbational data are inherently uncertain. Noisy interventions, incomplete priors, and hidden confounders require models and metrics to propagate uncertainty; otherwise, causal models risk overstating confidence in fragile or context-specific findings. Overall, causal approaches remain proof-of-concept; without standardized datasets, metrics, and benchmarks, virtual cells cannot yet reliably test mechanisms over correlations (Rawal et al., 2025).

4 Proposed Framework

4.1 Mechanistic Approaches to Model Design

The ambition for virtual cells is to represent cellular machinery in mechanistic detail, ideally as systems of differential equations capturing causal interactions and dynamics (Klipp et al., 2005; Alon, 2019). ODEs assume deterministic dynamics and face the "curse of dimensionality," making whole-cell simulation infeasible (Waltemath et al., 2011; Tomita et al., 1999). Extensions to SDEs capture intrinsic noise and uncertainty, essential for models that must quantify confidence as well as mean behavior. Progress will require hybrids that combine mechanistic grounding with deep learning flexibility. Universal and neural ODEs (Rackauckas et al., 2020; Chen et al., 2018) integrate biological priors with neural architectures, while causal constraints, sparsity, and disentangled representations improve interpretability (Brunton et al., 2016; An et al., 2025). Crucially, model design is inseparable from evaluation: benchmarks must test not only predictive accuracy but also causal validity (Peters et al., 2017; Schölkopf et al., 2021), ideally within a lab-in-the-loop paradigm where models are iteratively refined with experiments (Frey et al., 2025; Chandak et al., 2023).

4.2 Causal Evaluation from Established Data

In an ideal setting, causal evaluation would use multi-omic interventional time-series data with matched controls and rich context. A fundamental challenge is that most widely available datasets are observational, whereas causal inference requires interventional data (Rawal et al., 2025). Below, we propose four improvements to leverage existing data.

Quasi-Experimental Design can strengthen existing observational resources with matched controls to approximate causal contrasts. Propensity score matching (Rosenbaum & Rubin, 1983), paired sampling (Rubin, 1974), and distributional methods like optimal transport (Peyré et al., 2019) (exemplified by CINEMA-OT (Dong et al., 2023)) illustrate how confounders can be separated from perturbation effects to reconstruct counterfactual states. The goal is not full causal identification, but extending robust statistical tools to high-dimensional single-cell settings. Furthermore, most assays capture only static snapshots, so obtaining temporal anchors and allowing evaluating trajectory faithfulness requires proxies such as pseudotime (Trapnell et al., 2014; Saelens et al., 2019), RNA velocity (La Manno et al., 2018; Bergen et al., 2020), dose–response designs (Subramanian et al., 2017), and repeated sampling.

Biological Context Enhancement (in the absence of large-scale multi-omic interventional datasets) can capture interdependencies across molecular layers. The following strategies offer partial solutions: (i) *Structured priors*, such as KEGG (Kanehisa, 2002), Reactome (Fabregat et al., 2018), STRING (Szklarczyk et al., 2021), and BioGRID (Oughtred et al., 2019), which provide pathway

and interaction knowledge for fidelity tests. Meanwhile, ontologies such as GO (Consortium, 2019) and Cell Ontology (Diehl et al., 2016)) enable dataset alignment, and domain-specific LMs like BioBERT (Lee et al., 2020) enrich metadata. (ii) *Synthetic data-based* tools such as GeneNetWeaver and DREAM (Schaffter et al., 2011; Marbach et al., 2012), SERGIO (Dibaeinia & Sinha, 2020), DYNGEN (Cannoodt et al., 2021), and scDesign3 (Song et al., 2024) simulate perturbations and multi-omic readouts, providing ground truth for benchmarking.

Experimental Metadata helps discriminate between experimental variation and true biological signal. Examples of models that explicitly take these variations into account can be found in (An et al., 2025), (Gao et al., 2025), (Korsunsky et al., 2019), (Hao et al., 2021), and (Lopez et al., 2018). The following strategies help prepare datasets to provide this context: (i) *Metadata integration* on batch effects, protocols, and sample handling (GEO (Edgar et al., 2002), ArrayExpress (Parkinson et al., 2009), CELLxGENE (Program et al., 2025)) can stratify analyses; protocol-aware covariates improve comparability across assays (e.g. 10x vs. Smart-seq2) (Hicks et al., 2018). (ii) *Quality control and robustness*, such as UMIs, features, mitochondrial fraction, improve reliability (Luecken & Theis, 2019), and invariance-based methods such as ICP (Peters et al., 2016) and anchor regression (Rothenhäusler et al., 2021) test whether relationships remain stable across conditions.

Uncertainty Quantification (UQ) is essential to distinguish true signals from noise. While UQ *alone* does not yield causal models, it improves robustness in data-sparse regimes and guides experiment design. Approaches include: (i) Bayesian inference, (ii) Gaussian processes (iii) Ensembles and resampling (iv) Calibration (v) Information-theoretic scores (e.g. entropy, mutual information (BALD), and sensitivity indices (Houlsby et al., 2011)) (vi) Simulation-based inference (SBI) likelihood-free methods (Cranmer et al., 2020) quantify uncertainty in complex mechanistic models, with applications to stochastic gene expression (Toni et al., 2009), signaling dynamics (Golightly & Wilkinson, 2005), and single-cell electrophysiology (Lueckmann et al., 2017). Together, these methods enable virtual cells to attach explicit confidence to hypotheses, prioritize robust discoveries, and guide experimental validation in a lab-in-the-loop paradigm.

4.3 Uncertainty-Aware Causal Evaluation

A critical step is to adapt existing metrics to be uncertainty-aware, bridging current practice with the needs of causal virtual cells. For **intervention validity**, measures such as effect size correlation, treatment effect error, or distributional distances (Hill, 2011) could be extended with calibration (e.g., ECE (Naeini et al., 2015), Brier score (Glenn et al., 1950)), variance-aware distances, or likelihood-based comparisons of full distributions. For **counterfactual consistency**, where outcomes are unobservable, models should indicate high uncertainty for far out-of-distribution queries rather than overconfident predictions. For **trajectory faithfulness**, metrics such as DTW (Berndt & Clifford, 1994) or KL divergence (Kullback & Leibler, 1951) assume precise trajectories, but destructive assays prevent true before/after comparisons; evaluation should propagate error over time and flag uncertain regions in dose–response or developmental dynamics. For **mechanistic alignment**, pathway fidelity scores and graph distances like SHD and SID are deterministic; uncertainty-aware versions would weight edges by confidence, assigning higher certainty to well-established interactions (KEGG (Kanehisa, 2002), Reactome (Fabregat et al., 2018)) and lower to novel ones.

5 Discussion & Conclusion

Causal evaluation is the critical test of whether AI virtual cells can evolve from predictive simulators into trustworthy world models of biology. We outlined a taxonomy of causal metrics, emphasizing uncertainty as a cross-cutting principle. Standardized benchmarks that integrate interventions, trajectories, multi-omic context, and uncertainty are essential for robustness, interpretability, and translational impact. Without them, virtual cells remain unproven; with them, they can become reliable engines for discovery and therapeutic innovation. Embedding uncertainty at every level ensures evaluation asks not only 'was the prediction correct?' but also 'how certain should we be, and what should we do next?', providing the foundation for virtual cells that are not just predictive, but trustworthy and actionable.

REFERENCES

Josh Abramson, Jonas Adler, Jack Dunger, Richard Evans, Tim Green, Alexander Pritzel, Olaf Ronneberger, Lindsay Willmore, Andrew J Ballard, Joshua Bambrick, et al. Accurate structure prediction of biomolecular interactions with alphafold 3. *Nature*, 630(8016):493–500, 2024.

Malgorzata Adamczyk, Karen van Eunen, Barbara M Bakker, and Hans V Westerhoff. Enzyme kinetics for systems biology: When, why and how. In *Methods in enzymology*, volume 500, pp. 233–257. Elsevier, 2011.

Britt Adamson, Thomas M Norman, Marco Jost, Min Y Cho, James K Nuñez, Yuwen Chen, Jacqueline E Villalta, Luke A Gilbert, Max A Horlbeck, Marco Y Hein, et al. A multiplexed single-cell crispr screening platform enables systematic dissection of the unfolded protein response. *Cell*, 167(7):1867–1882, 2016.

Abhinav Adduri, Dhruv Gautam, Beatrice Bevilacqua, Alishba Imran, Rohan Shah, Mohsen Naghipourfar, Noam B. Teyssier, Rajesh Ilango, Sanjay Nagaraj, Mingze Dong, Chiara Ricci-Tam, Christopher Carpenter, Vishvak Subramanyam, Aidan Winters, Sravya Tirukkovular, Jeremy Sullivan, Brian S. Plosky, Basak Eraslan, Nicholas D. Youngblut, Jure Leskovec, Luke A. Gilbert, Silvana Konermann, Patrick D. Hsu, Alexander Dobin, Dave P. Burke, Hani Goodarzi, and Yusuf H. Roohani. Predicting cellular responses to perturbation across diverse contexts with state. *bioRxiv*, 2025. URL https://api.semanticscholar.org/CorpusID:279620354.

All of Us Research Program Investigators. The "all of us" research program. *New England Journal of Medicine*, 381(7):668–676, 2019.

Uri Alon. *An introduction to systems biology: design principles of biological circuits*. Chapman and Hall/CRC, 2019.

Shaokun An, Jae-Won Cho, Kai Cao, Jiankang Xiong, Martin Hemberg, and Lin Wan. sccausalvi disentangles single-cell perturbation responses with causality-aware generative model. *bioRxiv*, pp. 2025–02, 2025.

Martin Arjovsky, Léon Bottou, Ishaan Gulrajani, and David Lopez-Paz. Invariant risk minimization. *arXiv preprint arXiv:1907.02893*, 2019.

Pierre-Cyril Aubin-Frankowski and Jean-Philippe Vert. Gene regulation inference from single-cell rna-seq data with linear differential equations and velocity inference. *Bioinformatics*, 36(18): 4774–4780, 2020.

Žiga Avsec, Vikram Agarwal, Daniel Visentin, Joseph R. Ledsam, Agnieszka Grabska-Barwinska, Kyle R. Taylor, Yannis Assael, John Jumper, Pushmeet Kohli, and David R. Kelley. Effective gene expression prediction from sequence by integrating long-range interactions. 18(10):1196–1203. ISSN 1548-7105. doi: 10.1038/s41592-021-01252-x. URL https://doi.org/10.1038/s41592-021-01252-x.

Seungheun Baek, Soyon Park, Yan Ting Chok, Mogan Gim, and Jaewoo Kang. Gpo-vae: Modeling explainable gene perturbation responses utilizing grn-aligned parameter optimization. *arXiv preprint arXiv:2501.18973*, 2025a.

Seungheun Baek, Soyon Park, Yan Ting Chok, Junhyun Lee, Jueon Park, Mogan Gim, and Jaewoo Kang. Cradle-vae: Enhancing single-cell gene perturbation modeling with counterfactual reasoning-based artifact disentanglement. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 15445–15452, 2025b.

Elias Bareinboim and Judea Pearl. Causal inference and the data-fusion problem. *Proceedings of the National Academy of Sciences*, 113(27):7345–7352, 2016.

Yoshua Bengio, Tristan Deleu, Nasim Rahaman, Rosemary Ke, Sébastien Lachapelle, Olexa Bilaniuk, Anirudh Goyal, and Christopher Pal. A meta-transfer objective for learning to disentangle causal mechanisms. *arXiv preprint arXiv:1901.10912*, 2019.

Michael Bereket and Theofanis Karaletsos. Modelling cellular perturbations with the sparse additive mechanism shift variational autoencoder. *Advances in Neural Information Processing Systems*, 36:1–12, 2023.

Volker Bergen, Marius Lange, Stefan Peidli, F Alexander Wolf, and Fabian J Theis. Generalizing rna velocity to transient cell states through dynamical modeling. *Nature biotechnology*, 38(12): 1408–1414, 2020.

Donald J Berndt and James Clifford. Using dynamic time warping to find patterns in time series. In *Proceedings of the 3rd international conference on knowledge discovery and data mining*, pp. 359–370, 1994.

Garyk Brixi, Matthew G Durrant, Jerome Ku, Michael Poli, Greg Brockman, Daniel Chang, Gabriel A Gonzalez, Samuel H King, David B Li, Aditi T Merchant, et al. Genome modeling and design across all domains of life with evo 2. *BioRxiv*, pp. 2025–02, 2025.

Alexander P Browning, David J Warne, Kevin Burrage, Ruth E Baker, and Matthew J Simpson. Identifiability analysis for stochastic differential equation models in systems biology. *Journal of the Royal Society Interface*, 17(173):20200652, 2020.

Steven L Brunton, Joshua L Proctor, and J Nathan Kutz. Discovering governing equations from data by sparse identification of nonlinear dynamical systems. *Proceedings of the national academy of sciences*, 113(15):3932–3937, 2016.

Charlotte Bunne, Stefan G Stark, Gabriele Gut, Jacobo Sarabia Del Castillo, Mitch Levesque, Kjong-Van Lehmann, Lucas Pelkmans, Andreas Krause, and Gunnar Rätsch. Learning single-cell perturbation responses using neural optimal transport. *Nature methods*, 20(11):1759–1768, 2023.

Charlotte Bunne, Yusuf Roohani, Yanay Rosen, Ankit Gupta, Xikun Zhang, Marcel Roed, Theo Alexandrov, Mohammed AlQuraishi, Patricia Brennan, Daniel B Burkhardt, et al. How to build the virtual cell with artificial intelligence: Priorities and opportunities. *Cell*, 187(25):7045–7063, 2024.

Clare Bycroft, Colin Freeman, Desislava Petkova, Gavin Band, Lloyd T Elliott, Kevin Sharp, Allan Motyer, Damjan Vukcevic, Olivier Delaneau, Jared O'Connell, et al. The uk biobank resource with deep phenotyping and genomic data. *Nature*, 562(7726):203–209, 2018.

Robrecht Cannoodt, Wouter Saelens, Louise Deconinck, and Yvan Saeys. Spearheading future omics analyses using dyngen, a multi-modal simulator of single cells. *Nature communications*, 12(1):3942, 2021.

Ambrose Carr, Jonah Cool, Theofanis Karaletsos, Donghui Li, Alan R Lowe, Stephani Otte, and Sandra L Schmid. Ai: A transformative opportunity in cell biology. *Molecular Biology of the Cell*, 35(12):pe4, 2024.

Payal Chandak, Kexin Huang, and Marinka Zitnik. Building a knowledge graph to enable precision medicine. *Scientific Data*, 10(1):67, 2023.

Jiayang Chen, Zhihang Hu, Siqi Sun, Qingxiong Tan, Yixuan Wang, Qinze Yu, Licheng Zong, Liang Hong, Jin Xiao, Irwin King, and Yu Li. Interpretable rna foundation model from unannotated data for highly accurate rna structure and function predictions. *bioRxiv*, 2022a. URL `https://api.semanticscholar.org/CorpusID:247922548`.

Ricky TQ Chen, Yulia Rubanova, Jesse Bettencourt, and David K Duvenaud. Neural ordinary differential equations. *Advances in neural information processing systems*, 31, 2018.

Zhanlin Chen, William C King, Aheyon Hwang, Mark Gerstein, and Jing Zhang. Deepvelo: Single-cell transcriptomic deep velocity field learning with neural ordinary differential equations. *Science advances*, 8(48):eabq3745, 2022b.

Lu Cheng, Ruocheng Guo, Raha Moraffah, Paras Sheth, K Selçuk Candan, and Huan Liu. Evaluation methods and measures for causal learning algorithms. *IEEE Transactions on Artificial Intelligence*, 3(6):924–943, 2022.

David Maxwell Chickering. Optimal structure identification with greedy search. *Journal of machine learning research*, 3(Nov):507–554, 2002.

Tinyi Chu, Zhong Wang, Dana Pe'er, and Charles G Danko. Cell type and gene expression deconvolution with bayesprism enables bayesian integrative analysis across bulk and single-cell rna sequencing in oncology. *Nature cancer*, 3(4):505–517, 2022.

Gene Ontology Consortium. The gene ontology resource: 20 years and still going strong. *Nucleic acids research*, 47(D1):D330–D338, 2019.

Marc Courtemanche, Rafael J Ramirez, and Stanley Nattel. Ionic mechanisms underlying human atrial action potential properties: insights from a mathematical model. *American Journal of Physiology-Heart and Circulatory Physiology*, 275(1):H301–H321, 1998.

Kyle Cranmer, Johann Brehmer, and Gilles Louppe. The frontier of simulation-based inference. *Proceedings of the National Academy of Sciences*, 117(48):30055–30062, 2020.

Haotian Cui, Chloe Wang, Hassaan Maan, Kuan Pang, Fengning Luo, Nan Duan, and Bo Wang. scgpt: toward building a foundation model for single-cell multi-omics using generative ai. *Nature methods*, 21(8):1470–1480, 2024.

Marco Cuturi. Sinkhorn distances: Lightspeed computation of optimal transport. *Advances in neural information processing systems*, 26, 2013.

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pp. 248–255. Ieee, 2009.

Payam Dibaeinia and Saurabh Sinha. Sergio: a single-cell expression simulator guided by gene regulatory networks. *Cell systems*, 11(3):252–271, 2020.

Alexander D Diehl, Terrence F Meehan, Yvonne M Bradford, Matthew H Brush, Wasila M Dahdul, David S Dougall, Yongqun He, David Osumi-Sutherland, Alan Ruttenberg, Sirarat Sarntivijai, et al. The cell ontology 2016: enhanced content, modularization, and ontology interoperability. *Journal of biomedical semantics*, 7(1):44, 2016.

Atray Dixit, Oren Parnas, Biyu Li, Jenny Chen, Charles P Fulco, Livnat Jerby-Arnon, Nemanja D Marjanovic, Danielle Dionne, Tyler Burks, Raktima Raychowdhury, et al. Perturb-seq: dissecting molecular circuits with scalable single-cell rna profiling of pooled genetic screens. *cell*, 167(7): 1853–1866, 2016.

Mingze Dong, Bao Wang, Jessica Wei, Antonio H de O. Fonseca, Curtis J Perry, Alexander Frey, Feriel Ouerghi, Ellen F Foxman, Jeffrey J Ishizuka, Rahul M Dhodapkar, et al. Causal identification of single-cell experimental perturbation effects with cinema-ot. *Nature methods*, 20(11): 1769–1779, 2023.

Ron Edgar, Michael Domrachev, and Alex E Lash. Gene expression omnibus: Ncbi gene expression and hybridization array data repository. *Nucleic acids research*, 30(1):207–210, 2002.

Caleb N Ellington, Dian Li, Shuxian Zou, Elijah Cole, Ning Sun, Sohan Addagudi, Le Song, and Eric P Xing. Rapid and reproducible multimodal biological foundation model development with aido. modelgenerator. *bioRxiv*, pp. 2025–06, 2025.

Antonio Fabregat, Steven Jupe, Lisa Matthews, Konstantinos Sidiropoulos, Marc Gillespie, Phani Garapati, Robin Haw, Bijay Jassal, Florian Korninger, Bruce May, et al. The reactome pathway knowledgebase. *Nucleic acids research*, 46(D1):D649–D655, 2018.

David Feldman, Avtar Singh, Jonathan L Schmid-Burgk, Rebecca J Carlson, Anja Mezger, Anthony J Garrity, Feng Zhang, and Paul C Blainey. Optical pooled screens in human cells. *Cell*, 179(3):787–799, 2019.

Noelia Ferruz, Steffen Schmidt, and Birte Höcker. Protgpt2 is a deep unsupervised language model for protein design. *Nature communications*, 13(1):4348, 2022.

Aden Forrow and Geoffrey Schiebinger. Lineageot is a unified framework for lineage tracing and trajectory inference. *Nature communications*, 12(1):4940, 2021.

Nikolaus Fortelny and Christoph Bock. Knowledge-primed neural networks enable biologically interpretable deep learning on single-cell sequencing data. *Genome biology*, 21(1):190, 2020.

Nathan C Frey, Isidro Hötzel, Samuel D Stanton, Ryan Kelly, Robert G Alberstein, Emily Makowski, Karolis Martinkus, Daniel Berenberg, Jack Bevers III, Tyler Bryson, et al. Lab-in-the-loop therapeutic antibody design with deep learning. *bioRxiv*, pp. 2025–02, 2025.

Mingze Gao, Chen Qiao, and Yuanhua Huang. Unitvelo: temporally unified rna velocity reinforces single-cell trajectory inference. *Nature Communications*, 13(1):6586, 2022.

Yicheng Gao, Kejing Dong, Caihua Shan, Dongsheng Li, and Qi Liu. Causal disentanglement for single-cell representations and controllable counterfactual generation. *Nature Communications*, 16(1):6775, 2025.

Adam Gayoso, Romain Lopez, Galen Xing, Pierre Boyeau, Valeh Valiollah Pour Amiri, Justin Hong, Katherine Wu, Michael Jayasuriya, Edouard Mehlman, Maxime Langevin, et al. A python library for probabilistic analysis of single-cell omics data. *Nature biotechnology*, 40(2):163–166, 2022.

W Brier Glenn et al. Verification of forecasts expressed in terms of probability. *Monthly weather review*, 78(1):1–3, 1950.

Ben Glocker, Mirco Musolesi, Jonathan Richens, and Caroline Uhler. Causality in digital medicine. *Nature Communications*, 12(1), 2021.

Clark Glymour, Kun Zhang, and Peter Spirtes. Review of causal discovery methods based on graphical models. *Frontiers in genetics*, 10:524, 2019.

Andrew Golightly and Darren J Wilkinson. Bayesian inference for stochastic kinetic models using a diffusion approximation. *Biometrics*, 61(3):781–788, 2005.

Manoj Kumar Goshisht. Machine learning and deep learning in synthetic biology: Key architectures, applications, and challenges. *ACS omega*, 9(9):9921–9945, 2024.

Alex Greenfield, Aviv Madar, Harry Ostrer, and Richard Bonneau. Dream4: Combining genetic and dynamic information to identify biological networks and dynamical models. *PloS one*, 5(10): e13397, 2010.

Arthur Gretton, Karsten M Borgwardt, Malte J Rasch, Bernhard Schölkopf, and Alexander Smola. A kernel two-sample test. *The journal of machine learning research*, 13(1):723–773, 2012.

Ankit Gupta, Zoe Wefers, Konstantin Kahnert, Jan Niklas Hansen, William D. Leineweber, Anthony J. Cesnik, Dan Lu, Ulrika Axelsson, Frederic Ballllosera Navarro, Theofanis Karaletsos, and Emma Lundberg. Subcell: Vision foundation models for microscopy capture single-cell biology. *bioRxiv*, 2024. URL https://api.semanticscholar.org/CorpusID:274610971.

Minsheng Hao, Jing Gong, Xin Zeng, Chiming Liu, Yucheng Guo, Xingyi Cheng, Taifeng Wang, Jianzhu Ma, Xuegong Zhang, and Le Song. Large-scale foundation model on single-cell transcriptomics. *Nature methods*, 21(8):1481–1491, 2024.

Yuhan Hao, Stephanie Hao, Erica Andersen-Nissen, William M Mauck, Shiwei Zheng, Andrew Butler, Maddie J Lee, Aaron J Wilk, Charlotte Darby, Michael Zager, et al. Integrated analysis of multimodal single-cell data. *Cell*, 184(13):3573–3587, 2021.

Alain Hauser and Peter Bühlmann. Characterization and greedy learning of interventional markov equivalence classes of directed acyclic graphs. *The Journal of Machine Learning Research*, 13 (1):2409–2464, 2012.

Thomas Hayes, Roshan Rao, Halil Akin, Nicholas J Sofroniew, Deniz Oktay, Zeming Lin, Robert Verkuil, Vincent Q Tran, Jonathan Deaton, Marius Wiggert, et al. Simulating 500 million years of evolution with a language model. *Science*, 387(6736):850–858, 2025.

Yong He, Pan Fang, Yongtao Shan, Yuanfei Pan, Yanhong Wei, Yichang Chen, Yihao Chen, Yi Liu, Zhenyu Zeng, Zhan Zhou, et al. Generalized biological foundation model with unified nucleic acid and protein language. *BioRxiv*, pp. 2024–05, 2024.

Reinhart Heinrich and Tom A Rapoport. A linear steady-state treatment of enzymatic chains: general properties, control and effector strength. *European journal of biochemistry*, 42(1):89–95, 1974.

Christina Heinze-Deml, Jonas Peters, and Nicolai Meinshausen. Invariant causal prediction for nonlinear models. *Journal of Causal Inference*, 6(2):20170016, 2018.

Miguel A Hernán and James M Robins. Estimating causal effects from epidemiological data. *Journal of Epidemiology & Community Health*, 60(7):578–586, 2006.

Stephanie C Hicks, F William Townes, Mingxiang Teng, and Rafael A Irizarry. Missing data and technical variability in single-cell rna-sequencing experiments. *Biostatistics*, 19(4):562–578, 2018.

Joseph Higgins. A chemical mechanism for oscillation of glycolytic intermediates in yeast cells. *Proceedings of the National Academy of Sciences*, 51(6):989–994, 1964.

Jennifer L Hill. Bayesian nonparametric modeling for causal inference. *Journal of Computational and Graphical Statistics*, 20(1):217–240, 2011.

Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020.

Alan L Hodgkin and Andrew F Huxley. A quantitative description of membrane current and its application to conduction and excitation in nerve. *The Journal of physiology*, 117(4):500, 1952.

Leroy Hood and Mauricio Flores. A personal view on systems medicine and the emergence of proactive p4 medicine: predictive, preventive, personalized and participatory. *New biotechnology*, 29(6):613–624, 2012.

Intekhab Hossain, Viola Fanfani, Jonas Fischer, John Quackenbush, and Rebekka Burkholz. Biologically informed neuralodes for genome-wide regulatory dynamics. *Genome Biology*, 25(1): 127, 2024.

Neil Houlsby, Ferenc Huszár, Zoubin Ghahramani, and Máté Lengyel. Bayesian active learning for classification and preference learning. *arXiv preprint arXiv:1112.5745*, 2011.

Ann C Huang, Tsung-Han S Hsieh, Jiang Zhu, Jackson Michuda, Ashton Teng, Soohong Kim, Elizabeth M Rumsey, Sharon K Lam, Ikenna Anigbogu, Philip Wright, et al. X-atlas/orion: Genome-wide perturb-seq datasets via a scalable fix-cryopreserve platform for training dose-dependent biological foundation models. *bioRxiv*, pp. 2025–06, 2025.

Lawrence Hubert and Phipps Arabie. Comparing partitions. *Journal of classification*, 2(1):193–218, 1985.

Licheng Jiao, Yuhan Wang, Xu Liu, Lingling Li, Fang Liu, Wenping Ma, Yuwei Guo, Puhua Chen, Shuyuan Yang, and Biao Hou. Causal inference meets deep learning: A comprehensive survey. *Research*, 7:0467, 2024.

Fredrik Johansson, Uri Shalit, and David Sontag. Learning representations for counterfactual inference. In *International conference on machine learning*, pp. 3020–3029. PMLR, 2016.

Jeremie Kalfon, Laura Cantini, and Gabriel Peyre. Towards foundation models that learn across biological scales. *bioRxiv*, pp. 2025–05, 2025a.

Jérémie Kalfon, Jules Samaran, Gabriel Peyré, and Laura Cantini. scprint: pre-training on 50 million cells allows robust gene network predictions. *Nature Communications*, 16(1):3607, 2025b.

Minoru Kanehisa. The kegg database. In *'In silico'simulation of biological processes: Novartis Foundation Symposium 247*, volume 247, pp. 91–103. Wiley Online Library, 2002.

Hiroaki Kitano. Systems biology: a brief overview. *science*, 295(5560):1662–1664, 2002.

Dominik Klein, Jonas Simon Fleck, Daniil Bobrovskiy, Lea Zimmermann, Sören Becker, Alessandro Palma, Leander Dony, Alejandro Tejada-Lapuerta, Guillaume Huguet, Hsiu-Chuan Lin, et al. Cellflow enables generative single-cell phenotype modeling with flow matching. *bioRxiv*, pp. 2025–04, 2025.

Edda Klipp and Wolfram Liebermeister. Mathematical modeling of intracellular signaling pathways. *BMC neuroscience*, 7(Suppl 1):S10, 2006.

Edda Klipp, Ralf Herwig, Axel Kowald, Christoph Wierling, and Hans Lehrach. *Systems biology in practice: concepts, implementation and application*. John Wiley & Sons, 2005.

Michał Komorowski, Maria J Costa, David A Rand, and Michael PH Stumpf. Sensitivity, robustness, and identifiability in stochastic chemical kinetics models. *Proceedings of the National Academy of Sciences*, 108(21):8645–8650, 2011.

Ilya Korsunsky, Nghia Millard, Jean Fan, Kamil Slowikowski, Fan Zhang, Kevin Wei, Yuriy Baglaenko, Michael Brenner, Po-ru Loh, and Soumya Raychaudhuri. Fast, sensitive and accurate integration of single-cell data with harmony. *Nature methods*, 16(12):1289–1296, 2019.

Zheng Kuang, Zhicheng Ji, Jef D Boeke, and Hongkai Ji. Dynamic motif occupancy (dynamo) analysis identifies transcription factors and their binding sites driving dynamic biological processes. *Nucleic acids research*, 46(1):e2–e2, 2018.

Solomon Kullback and Richard A Leibler. On information and sufficiency. *The annals of mathematical statistics*, 22(1):79–86, 1951.

Gioele La Manno, Ruslan Soldatov, Amit Zeisel, Emelie Braun, Hannah Hochgerner, Viktor Petukhov, Katja Lidschreiber, Maria E Kastriti, Peter Lönnerberg, Alessandro Furlan, et al. Rna velocity of single cells. *Nature*, 560(7719):494–498, 2018.

Sébastien Lachapelle, Philippe Brouillard, Tristan Deleu, and Simon Lacoste-Julien. Gradient-based neural dag learning. *arXiv preprint arXiv:1906.02226*, 2019.

Yunduo Lan, Sung-Young Shin, and Lan K Nguyen. From shallow to deep: the evolution of machine learning and mechanistic model integration in cancer research. *Current Opinion in Systems Biology*, 40:100541, 2025.

Zachary M Laubach, Eleanor J Murray, Kim L Hoke, Rebecca J Safran, and Wei Perng. A biologist's guide to model selection and causal inference. *Proceedings of the Royal Society B*, 288(1943): 20202815, 2021.

Nicolas Le Novere, Benjamin Bornstein, Alexander Broicher, Melanie Courtot, Marco Donizelli, Harish Dharuri, Lu Li, Herbert Sauro, Maria Schilstra, Bruce Shapiro, et al. Biomodels database: a free, centralized database of curated, published, quantitative kinetic models of biochemical and cellular systems. *Nucleic acids research*, 34(suppl_1):D689–D691, 2006.

Jinhyuk Lee, Wonjin Yoon, Sungdong Kim, Donghyeon Kim, Sunkyu Kim, Chan Ho So, and Jaewoo Kang. Biobert: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*, 36(4):1234–1240, 2020.

Guipeng Li, Feifei Jiang, Junhao Zhu, Huanhuan Cui, Zefeng Wang, and Wei Chen. Hydrarna: a hybrid architecture based full-length rna language model. *bioRxiv*, pp. 2025–03, 2025a.

Hechen Li, Ziqi Zhang, Michael Squires, Xi Chen, and Xiuwei Zhang. scmultisim: simulation of single-cell multi-omics and spatial data guided by gene regulatory networks and cell–cell interactions. *Nature Methods*, pp. 1–12, 2025b.

Sizhen Li, Shahriar Noroozizadeh, Saeed Moayedpour, Lorenzo Kogler-A nele, Zexin Xue, Dinghai Zheng, Fernando Ulloa Montoya, Vikram Agarwa l, Ziv Bar-Joseph, and Sven Jager. mrna-lm: full-length integrated slm for mrna analysis. *Nucleic Acids Research*, 53(3):gkaf044, 02 2025c. ISSN 1362-4962. doi: 10.1093/nar/gkaf044. URL https://doi.org/10.1093/nar/gkaf044.

Zaikang Lin, Sei Chang, Aaron Zweig, Minseo Kang, Elham Azizi, and David A Knowles. Interpretable neural odes for gene regulatory network discovery under perturbations. *arXiv preprint arXiv:2501.02409*, 2025.

Zeming Lin, Halil Akin, Roshan Rao, Brian Hie, Zhongkai Zhu, Wenting Lu, Nikita Smetanin, Robert Verkuil, Ori Kabeli, Yaniv Shmueli, et al. Evolutionary-scale prediction of atomic-level protein structure with a language model. *Science*, 379(6637):1123–1130, 2023.

Jennifer Listgarten. The perpetual motion machine of ai-generated data and the distraction of chatgpt as a 'scientist'. *nature biotechnology*, 42(3):371–373, 2024.

Guole Liu, Yongbing Zhao, Yingying Zhao, Tianyu Wang, Quanyou Cai, Xiaotao Wang, Ziyi Wen, Lihui Lin, Ge Yang, and Jiekai Chen. Scarf: Single cell atac-seq and rna-seq foundation model. *bioRxiv*, pp. 2025–04, 2025.

Sebastian Lobentanzer, Pablo Rodriguez-Mier, Stefan Bauer, and Julio Saez-Rodriguez. Molecular causality in the advent of foundation models. *Molecular Systems Biology*, 20(8):848–858, 2024.

Romain Lopez, Jeffrey Regier, Michael B Cole, Michael I Jordan, and Nir Yosef. Deep generative modeling for single-cell transcriptomics. *Nature methods*, 15(12):1053–1058, 2018.

Romain Lopez, Jan-Christian Hütter, Jonathan Pritchard, and Aviv Regev. Large-scale differentiable causal discovery of factor graphs. *Advances in Neural Information Processing Systems*, 35: 19290–19303, 2022.

Mohammad Lotfollahi, F Alexander Wolf, and Fabian J Theis. scgen predicts single-cell perturbation responses. *Nature methods*, 16(8):715–721, 2019.

Mohammad Lotfollahi, Anna Klimovskaia Susmelj, Carlo De Donno, Yuge Ji, Ignacio L Ibarra, F Alexander Wolf, Nafissa Yakubova, Fabian J Theis, and David Lopez-Paz. Compositional perturbation autoencoder for single-cell response modeling. *BioRxiv*, 2021.

Mohammad Lotfollahi, Anna Klimovskaia Susmelj, Carlo De Donno, Leon Hetzel, Yuge Ji, Ignacio L Ibarra, Sanjay R Srivatsan, Mohsen Naghipourfar, Riza M Daza, Beth Martin, et al. Predicting cellular responses to complex perturbations in high-throughput screens. *Molecular systems biology*, 19(6):e11517, 2023.

Christos Louizos, Uri Shalit, Joris M Mooij, David Sontag, Richard Zemel, and Max Welling. Causal effect inference with deep latent-variable models. *Advances in neural information processing systems*, 30, 2017.

Malte D Luecken and Fabian J Theis. Current best practices in single-cell rna-seq analysis: a tutorial. *Molecular systems biology*, 15(6):e8746, 2019.

Jan-Matthis Lueckmann, Pedro J Goncalves, Giacomo Bassetto, Kaan Öcal, Marcel Nonnenmacher, and Jakob H Macke. Flexible statistical inference for mechanistic models of neural dynamics. *Advances in neural information processing systems*, 30, 2017.

Ali Madani, Ben Krause, Eric R Greene, Subu Subramanian, Benjamin P Mohr, James M Holton, Jose Luis Olmos Jr, Caiming Xiong, Zachary Z Sun, Richard Socher, et al. Large language models generate functional protein sequences across diverse families. *Nature biotechnology*, 41 (8):1099–1106, 2023.

Daniel Marbach, James C Costello, Robert Küffner, Nicole M Vega, Robert J Prill, Diogo M Camacho, Kyle R Allison, Manolis Kellis, James J Collins, et al. Wisdom of crowds for robust gene network inference. *Nature methods*, 9(8):796–804, 2012.

Hirotaka Matsumoto, Hisanori Kiryu, Chikara Furusawa, Minoru SH Ko, Shigeru BH Ko, Norio Gouda, Tetsutaro Hayashi, and Itoshi Nikaido. Scode: an efficient regulatory network inference algorithm from single-cell rna-seq during differentiation. *Bioinformatics*, 33(15):2314–2321, 2017.

Tom Michoel and Jitao David Zhang. Causal inference in drug discovery and development. *Drug discovery today*, 28(10):103737, 2023.

Gemma E Moran and Bryon Aragam. Towards interpretable deep generative models via causal representation learning. *arXiv preprint arXiv:2504.11609*, 2025.

Mahdi Pakdaman Naeini, Gregory Cooper, and Milos Hauskrecht. Obtaining well calibrated probabilities using bayesian binning. In *Proceedings of the AAAI conference on artificial intelligence*, volume 29, 2015.

Wenjin Niu, Zijun Gao, Liyan Song, and Lingbo Li. Comprehensive review and empirical evaluation of causal discovery algorithms for numerical data. *arXiv preprint arXiv:2407.13054*, 2024.

Denis Noble. Cardiac action and pacemaker potentials based on the hodgkin-huxley equations. *Nature*, 188(4749):495–497, 1960.

Emmanuel Noutahi, Jason Hartford, Prudencio Tossou, Shawn Whitfield, Alisandra K Denton, Cas Wognum, Kristina Ulicna, Michael Craig, Jonathan Hsu, Michael Cuccarese, et al. Virtual cells: Predict, explain, discover. *arXiv preprint arXiv:2505.14613*, 2025.

Jingfeng Ou, Jiawei Li, Zhiliang Xia, Shurui Dai, Yulian Ding, Yan Guo, Limin Jiang, and Jijun Tang. sccade: A superior tool for predicting perturbation responses in single-cell gene expression using contrastive learning and attention mechanisms. In *2024 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, pp. 441–444. IEEE, 2024.

Rose Oughtred, Chris Stark, Bobby-Joe Breitkreutz, Jennifer Rust, Lorrie Boucher, Christie Chang, Nadine Kolas, Lara O'Donnell, Genie Leung, Rochelle McAdam, et al. The biogrid interaction database: 2019 update. *Nucleic acids research*, 47(D1):D529–D541, 2019.

Nan Papili Gao, SM Minhaz Ud-Dean, Olivier Gandrillon, and Rudiyanto Gunawan. Sincerities: inferring gene regulatory networks from time-stamped single cell transcriptional expression profiles. *Bioinformatics*, 34(2):258–266, 2018.

Helen Parkinson, Misha Kapushesky, Nikolay Kolesnikov, Gabriella Rustici, Mohammad Shojatalab, Niran Abeygunawardena, Hugo Berube, Miroslaw Dylag, Ibrahim Emam, Anna Farne, et al. Arrayexpress update—from an archive of functional genomics experiments to the atlas of gene expression. *Nucleic acids research*, 37(suppl_1):D868–D872, 2009.

Parse Biosciences. 10 million human pbmcs in a single experiment, 2023. URL https://www.parsebiosciences.com/datasets/10-million-human-pbmcs-in-a-single-experiment/.

Saro Passaro, Gabriele Corso, Jeremy Wohlwend, Mateo Reveiz, Stephan Thaler, Vignesh Ram Somnath, Noah Getz, Tally Portnoi, Julien Roy, Hannes Stark, et al. Boltz-2: Towards accurate and efficient binding affinity prediction. *BioRxiv*, pp. 2025–06, 2025.

Judea Pearl. The do-calculus revisited. *arXiv preprint arXiv:1210.4852*, 2012.

Stefan Peidli, Tessa D Green, Ciyue Shen, Torsten Gross, Joseph Min, Samuele Garda, Bo Yuan, Linus J Schumacher, Jake P Taylor-King, Debora S Marks, et al. scperturb: harmonized single-cell perturbation data. *Nature Methods*, 21(3):531–540, 2024.

Qiangwei Peng, Xiaojie Qiu, and Tiejun Li. Storm: Incorporating transient stochastic dynamics to infer the rna velocity with metabolic labeling information. *PLOS Computational Biology*, 20(11): e1012606, 2024.

Rafael Josip Penić, Tin Vlašić, Roland G. Huber, Yue Wan, and Mile Šikić. Rinalmo: general-purpose rna language models can generalize well on structure prediction tasks. *Nature Communications*, 16(1), July 2025. ISSN 2041-1723. doi: 10.1038/s41467-025-60872-5. URL http://dx.doi.org/10.1038/s41467-025-60872-5.

Sebastian Persson, Niek Welkenhuysen, Sviatlana Shashkova, Samuel Wiqvist, Patrick Reith, Gregor W Schmidt, Umberto Picchini, and Marija Cvijovic. Scalable and flexible inference framework for stochastic dynamic single-cell models. *PLoS computational biology*, 18(5):e1010082, 2022.

Jonas Peters and Peter Bühlmann. Structural intervention distance (sid) for evaluating causal graphs. *arXiv preprint arXiv:1306.1043*, 2013.

Jonas Peters, Peter Bühlmann, and Nicolai Meinshausen. Causal inference by using invariant prediction: identification and confidence intervals. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 78(5):947–1012, 2016.

Jonas Peters, Dominik Janzing, and Bernhard Schölkopf. *Elements of causal inference: foundations and learning algorithms*. The MIT press, 2017.

Gabriel Peyré, Marco Cuturi, et al. Computational optimal transport: With applications to data science. *Foundations and Trends® in Machine Learning*, 11(5-6):355–607, 2019.

Zoe Piran, Niv Cohen, Yedid Hoshen, and Mor Nitzan. Disentanglement of single-cell data with biolord. *Nature Biotechnology*, 42(11):1678–1683, 2024.

CZI Cell Science Program, Shibla Abdulla, Brian Aevermann, Pedro Assis, Seve Badajoz, Sidney M Bell, Emanuele Bezzi, Batuhan Cakir, Jim Chaffer, Signe Chambers, et al. Cz cellxgene discover: a single-cell data platform for scalable exploration, analysis and modeling of aggregated data. *Nucleic acids research*, 53(D1):D886–D900, 2025.

Stephen R Quake and Tabula Sapiens Consortium. Tabula sapiens reveals transcription factor expression, senescence effects, and sex-specific features in cell types from 28 human organs and tissues. *bioRxiv*, pp. 2024–12, 2024.

Christopher Rackauckas, Yingbo Ma, Julius Martensen, Collin Warner, Kirill Zubov, Rohit Supekar, Dominic Skinner, Ali Ramadhan, and Alan Edelman. Universal differential equations for scientific machine learning. *arXiv preprint arXiv:2001.04385*, 2020.

Roshan M Rao, Jason Liu, Robert Verkuil, Joshua Meier, John Canny, Pieter Abbeel, Tom Sercu, and Alexander Rives. Msa transformer. In *International conference on machine learning*, pp. 8844–8856. PMLR, 2021.

Atul Rawal, Adrienne Raglin, Danda B Rawat, Brian M Sadler, and James McCoy. Causality for trustworthy artificial intelligence: status, challenges and perspectives. *ACM Computing Surveys*, 57(6):1–30, 2025.

Aviv Regev, Sarah A Teichmann, Eric S Lander, Ido Amit, Christophe Benoist, Ewan Birney, Bernd Bodenmiller, Peter Campbell, Piero Carninci, Menna Clatworthy, et al. The human cell atlas. *elife*, 6:e27041, 2017.

Joseph M Replogle, Reuben A Saunders, Angela N Pogson, Jeffrey A Hussmann, Alexander Lenail, Alina Guna, Lauren Mascibroda, Eric J Wagner, Karen Adelman, Gila Lithwick-Yanai, et al. Mapping information-rich genotype-phenotype landscapes with genome-scale perturb-seq. *Cell*, 185(14):2559–2575, 2022.

Guillaume Richard, Bernardo P de Almeida, Hugo Dalla-Torre, Christopher Blum, Lorenz Hexemer, Priyanka Pandey, Stefan Laurent, Marie Lopez, Alexandre Laterre, Maren Lang, et al. Chatnt: A multimodal conversational agent for dna, rna and protein tasks. *bioRxiv*, pp. 2024–04, 2024.

Jennifer E Rood, Anna Hupalowska, and Aviv Regev. Toward a foundation model of causal cell and tissue biology with a perturbation cell and tissue atlas. *Cell*, 187(17):4520–4545, 2024.

Yusuf Roohani, Kexin Huang, and Jure Leskovec. Predicting transcriptional outcomes of novel multigene perturbations with gears. *Nature Biotechnology*, 42(6):927–935, 2024.

Paul R Rosenbaum and Donald B Rubin. The central role of the propensity score in observational studies for causal effects. *Biometrika*, 70(1):41–55, 1983.

Dominik Rothenhäusler, Peter Bühlmann, and Nicolai Meinshausen. Causal dantzig. *The Annals of Statistics*, 47(3):1688–1722, 2019.

Dominik Rothenhäusler, Nicolai Meinshausen, Peter Bühlmann, and Jonas Peters. Anchor regression: Heterogeneous data meet causality. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 83(2):215–246, 2021.

Yulia Rubanova, Ricky TQ Chen, and David K Duvenaud. Latent ordinary differential equations for irregularly-sampled time series. *Advances in neural information processing systems*, 32, 2019.

Adam J Rubin, Kevin R Parker, Ansuman T Satpathy, Yanyan Qi, Beijing Wu, Alvin J Ong, Maxwell R Mumbach, Andrew L Ji, Daniel S Kim, Seung Woo Cho, et al. Coupled single-cell crispr screening and epigenomic profiling reveals causal gene regulatory networks. *Cell*, 176 (1):361–376, 2019.

Donald B Rubin. Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of educational Psychology*, 66(5):688, 1974.

Karen Sachs, Omar Perez, Dana Pe'er, Douglas A Lauffenburger, and Garry P Nolan. Causal protein-signaling networks derived from multiparameter single-cell data. *Science*, 308(5721): 523–529, 2005.

Wouter Saelens, Robrecht Cannoodt, Helena Todorov, and Yvan Saeys. A comparison of single-cell trajectory inference methods. *Nature biotechnology*, 37(5):547–554, 2019.

Thomas Schaffter, Daniel Marbach, and Dario Floreano. Genenetweaver: in silico benchmark generation and performance profiling of network inference methods. *Bioinformatics*, 27(16):2263–2270, 2011.

Bernhard Schölkopf, Francesco Locatello, Stefan Bauer, Nan Rosemary Ke, Nal Kalchbrenner, Anirudh Goyal, and Yoshua Bengio. Toward causal representation learning. *Proceedings of the IEEE*, 109(5):612–634, 2021.

Uri Shalit, Fredrik D Johansson, and David Sontag. Estimating individual treatment effect: generalization bounds and algorithms. In *International conference on machine learning*, pp. 3076–3085. PMLR, 2017.

Rohit Singh, Alexander P Wu, Anish Mudide, and Bonnie Berger. Causal gene regulatory analysis with rna velocity reveals an interplay between slow and fast transcription factors. *Cell systems*, 15(5):462–474, 2024.

Dongyuan Song, Qingyang Wang, Guanao Yan, Tianyang Liu, Tianyi Sun, and Jingyi Jessica Li. scdesign3 generates realistic in silico data for multimodal single-cell and spatial omics. *Nature Biotechnology*, 42(2):247–252, 2024.

Yang Song, Jascha Sohl-Dickstein, Diederik P Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. Score-based generative modeling through stochastic differential equations. *arXiv preprint arXiv:2011.13456*, 2020.

Peter Spirtes and Clark Glymour. An algorithm for fast recovery of sparse causal graphs. *Social science computer review*, 9(1):62–72, 1991.

Peter Spirtes, Clark N Glymour, and Richard Scheines. *Causation, prediction, and search*. MIT press, 2000.

David R. Stirling, Madison J. Swain-Bowden, Alice M. Lucas, Anne E Carpenter, Beth A. Cimini, and Allen Goodman. Cellprofiler 4: improvements in speed, utility and usability. *BMC Bioinformatics*, 22, 2021. URL `https://api.semanticscholar.org/CorpusID:235718146`.

Marlon Stoeckius, Christoph Hafemeister, William Stephenson, Brian Houck-Loomis, Pratip K Chattopadhyay, Harold Swerdlow, Rahul Satija, and Peter Smibert. Simultaneous epitope and transcriptome measurement in single cells. *Nature methods*, 14(9):865–868, 2017.

Adam F Strassberg and Louis J DeFelice. Limitations of the hodgkin-huxley formalism: Effects of single channel kinetics on transmembrane voltage dynamics. *Neural computation*, 5(6):843–855, 1993.

Aravind Subramanian, Rajiv Narayan, Steven M Corsello, David D Peck, Ted E Natoli, Xiaodong Lu, Joshua Gould, John F Davis, Andrew A Tubelli, Jacob K Asiedu, et al. A next generation connectivity map: L1000 platform and the first 1,000,000 profiles. *Cell*, 171(6):1437–1452, 2017.

Artur Szałata, Andrew Benz, Robrecht Cannoodt, Mauricio Cortes, Jason Fong, Sunil Kuppasani, Richard Lieberman, Tianyu Liu, Javier A Mas-Rosario, Rico Meinl, et al. A benchmark for prediction of transcriptomic responses to chemical perturbations across cell types. *Advances in Neural Information Processing Systems*, 37:20566–20616, 2024.

Gábor J Székely and Maria L Rizzo. Energy statistics: A class of statistics based on distances. *Journal of statistical planning and inference*, 143(8):1249–1272, 2013.

Damian Szklarczyk, Annika L Gable, Katerina C Nastou, David Lyon, Rebecca Kirsch, Sampo Pyysalo, Nadezhda T Doncheva, Marc Legeay, Tao Fang, Peer Bork, et al. The string database in 2021: customizable protein–protein networks, and functional characterization of user-uploaded gene/measurement sets. *Nucleic acids research*, 49(D1):D605–D612, 2021.

Zijia Tang, Minghao Zhou, Kai Zhang, and Qianqian Song. Scperb: Predict single-cell perturbation via style transfer-based variational autoencoder. *Journal of Advanced Research*, 2024.

Yifeng Tao, Chunhui Cai, William W Cohen, and Xinghua Lu. From genome to phenome: Predicting multiple cancer phenotypes based on somatic genomic alterations via the genomic impact transformer. In *PACIFIC SYMPOSIUM ON BIOCOMPUTING 2020*, pp. 79–90. World Scientific, 2019.

Alejandro Tejada-Lapuerta, Paul Bertin, Stefan Bauer, Hananeh Aliee, Yoshua Bengio, and Fabian J Theis. Causal machine learning for single-cell genomics. *Nature Genetics*, pp. 1–12, 2025.

Christina V Theodoris, Ling Xiao, Anant Chopra, Mark D Chaffin, Zeina R Al Sayed, Matthew C Hill, Helene Mantineo, Elizabeth M Brydon, Zexian Zeng, X Shirley Liu, et al. Transfer learning enables predictions in network biology. *Nature*, 618(7965):616–624, 2023.

Katarzyna Tomczak, Patrycja Czerwińska, and Maciej Wiznerowicz. Review the cancer genome atlas (tcga): an immeasurable source of knowledge. *Contemporary Oncology/Współczesna Onkologia*, 2015(1):68–77, 2015.

Masaru Tomita, Kenta Hashimoto, Koichi Takahashi, Thomas Simon Shimizu, Yuri Matsuzaki, Fumihiko Miyoshi, Kanako Saito, Sakura Tanida, Katsuyuki Yugi, J Craig Venter, et al. E-cell: software environment for whole-cell simulation. *Bioinformatics (Oxford, England)*, 15(1):72–84, 1999.

Jenna Tomkinson, Roshan Kern, Cameron Mattson, and Gregory P. Way. Toward generalizable phenotype prediction from single-cell morphology representations. *bioRxiv*, 2024. URL https://api.semanticscholar.org/CorpusID:268417539.

Tina Toni, David Welch, Natalja Strelkowa, Andreas Ipsen, and Michael PH Stumpf. Approximate bayesian computation scheme for parameter inference and model selection in dynamical systems. *Journal of the Royal Society Interface*, 6(31):187–202, 2009.

Cole Trapnell, Davide Cacchiarelli, Jonna Grimsby, Prapti Pokharel, Shuqiang Li, Michael Morse, Niall J Lennon, Kenneth J Livak, Tarjei S Mikkelsen, and John L Rinn. The dynamics and regulators of cell fate decisions are revealed by pseudotemporal ordering of single cells. *Nature biotechnology*, 32(4):381–386, 2014.

Tim Van den Bulcke, Koenraad Van Leemput, Bart Naudts, Piet van Remortel, Hongwu Ma, Alain Verschoren, Bart De Moor, and Kathleen Marchal. Syntren: a generator of synthetic gene expression data for design and analysis of structure learning algorithms. *BMC bioinformatics*, 7(1):43, 2006.

Dagmar Waltemath, Richard Adams, Frank T Bergmann, Michael Hucka, Fedor Kolpakov, Andrew K Miller, Ion I Moraru, David Nickerson, Sven Sahle, Jacky L Snoep, et al. Reproducible computational biology experiments with sed-ml-the simulation experiment description markup language. *BMC systems biology*, 5(1):198, 2011.

Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R Bowman. Glue: A multi-task benchmark and analysis platform for natural language understanding. *arXiv preprint arXiv:1804.07461*, 2018.

Zeyu Wang. Causalbench: A comprehensive benchmark for evaluating causal reasoning capabilities of large language models. In *Proceedings of the 10th SIGHAN Workshop on Chinese Language Processing (SIGHAN-10)*, pp. 143–151, 2024.

Yujian Wen, Jielong Huang, Shuhui Guo, Yehezqel Elyahu, Alon Monsonego, Hai Zhang, Yanqing Ding, and Hao Zhu. Applying causal discovery to single-cell analyses using causalcell. *Elife*, 12: e81464, 2023.

Christopher Winship and Stephen L Morgan. The estimation of causal effects from observational data. *Annual review of sociology*, 25(1):659–706, 1999.

Yahao Wu, Jing Liu, Yanni Xiao, Shuqin Zhang, and Limin Li. Couplevae: coupled variational autoencoders for predicting perturbational single-cell rna sequencing data. *Briefings in Bioinformatics*, 26(2), 2025.

Yan Wu, Esther Wershof, Sebastian M Schmon, Marcel Nassar, Błażej Osiński, Ridvan Eksi, Zichao Yan, Rory Stark, Kun Zhang, and Thore Graepel. Perturbench: Benchmarking machine learning models for cellular perturbation analysis. *arXiv preprint arXiv:2408.10609*, 2024.

Yulun Wu, Robert A Barton, Zichen Wang, Vassilis N Ioannidis, Carlo De Donno, Layne C Price, Luis F Voloch, and George Karypis. Predicting cellular responses with variational causal inference and refined relational information. *arXiv preprint arXiv:2210.00116*, 2022.

Fan Yang, Wenchuan Wang, Fang Wang, Yuan Fang, Duyu Tang, Junzhou Huang, Hui Lu, and Jianhua Yao. scbert as a large-scale pretrained deep language model for cell type annotation of single-cell rna-seq data. *Nature Machine Intelligence*, 4(10):852–866, 2022.

Jiayu Yao, Weiwei Pan, Soumya Ghosh, and Finale Doshi-Velez. Quality of uncertainty quantification for bayesian neural network inference. *arXiv preprint arXiv:1906.09686*, 2019.

Nicholas D Youngblut, Christopher Carpenter, Jaanak Prashar, Chiara Ricci-Tam, Rajesh Ilango, Noam Teyssier, Silvana Konermann, Patrick D Hsu, Alexander Dobin, David P Burke, et al. scbasecount: an ai agent-curated, uniformly processed, and continually expanding single cell data repository. *bioRxiv*, pp. 2025–02, 2025.

Hengshi Yu and Joshua D Welch. Michigan: sampling from disentangled representations of single-cell data using generative adversarial networks. *Genome biology*, 22(1):158, 2021.

Yue Yu, Jie Chen, Tian Gao, and Mo Yu. Dag-gnn: Dag structure learning with graph neural networks. In *International conference on machine learning*, pp. 7154–7163. PMLR, 2019.

Qiuyue Yuan and Zhana Duren. Inferring gene regulatory networks from single-cell multiome data using atlas-scale external data. *Nature Biotechnology*, 43(2):247–257, 2025.

Alessio Zanga, Elif Ozkirimli, and Fabio Stella. A survey on causal discovery: theory and practice. *International Journal of Approximate Reasoning*, 151:101–129, 2022.

Luke Zappia, Belinda Phipson, and Alicia Oshlack. Splatter: simulation of single-cell rna sequencing data. *Genome biology*, 18(1):174, 2017.

Yuansong Zeng, Jiancong Xie, Ningyuan Shangguan, Zhuoyi Wei, Wenbing Li, Yun Su, Shuangyu Yang, Chengyang Zhang, Jinbo Zhang, Nan Fang, et al. Cellfm: a large-scale foundation model pre-trained on transcriptomics of 100 million human cells. *Nature Communications*, 16(1):4679, 2025.

Jesse Zhang, Airol A Ubas, Richard de Borja, Valentine Svensson, Nicole Thomas, Neha Thakar, Ian Lai, Aidan Winters, Umair Khan, Matthew G Jones, et al. Tahoe-100m: A giga-scale single-cell perturbation atlas for context-dependent gene function and cellular modeling. *BioRxiv*, pp. 2025–02, 2025.

Xiuwei Zhang, Chenling Xu, and Nir Yosef. Simulating multiple faceted variability in single cell rna sequencing. *Nature communications*, 10(1):2611, 2019.

Yan Zhang, Xiaojie Qiu, Ke Ni, Jonathan Weissman, Ivet Bahar, and Jianhua Xing. Graph-dynamo: Learning stochastic cellular state transition dynamics from single cell data. *BioRxiv*, pp. 2023–09, 2023.

Xun Zheng, Bryon Aragam, Pradeep K Ravikumar, and Eric P Xing. Dags with no tears: Continuous optimization for structure learning. *Advances in neural information processing systems*, 31, 2018.

Yuxuan Zheng and George F Gao. Geneformer: a deep learning model for exploring gene networks. *Science China Life Sciences*, 66(12):2952–2954, 2023.

Shuxian Zou, Tianhua Tao, Sazan Mahbub, Caleb N Ellington, Robin Algayres, Dian Li, Yonghao Zhuang, Hongyi Wang, Le Song, and Eric P Xing. A large-scale foundation model for rna function and structure prediction. *bioRxiv*, pp. 2024–11, 2024.