Virtual Cells as Causal World Models: A Perspective on Evaluation

Tiffany J. Callahan
tiffany.callahan@sandboxquantum.com

Zane Beckwith
zane.beckwith@sandboxquantum.com

Thomas Merth
thomas.merth@sandboxquantum.com

Constantijn van der Poel
constantijn@sandboxquantum.com

Adam Lewis
adam.lewis@sandboxquantum.com

Pablo Lemos
pablo.lemos@sandboxquantum.com

SandboxAQ

Abstract

Evaluating virtual cells requires moving beyond predictive accuracy to assessing their ability to serve as causal world models of biology. Current benchmarks emphasize fit to observed data, rewarding pattern matching but rarely testing responses to interventions. We argue that building causal virtual cells demands a new evaluation paradigm based on metrics and benchmarks that assess intervention validity, counterfactual consistency, trajectory faithfulness, and mechanistic alignment. Our contribution is twofold: (1) a survey of recent approaches to virtual cell modeling, and (2) a taxonomy of causal evaluation metrics mapped to available perturbation datasets. By identifying gaps and proposing unified causal benchmarks, we position causal evaluation as the critical step toward making virtual cells reliable world models of biology.

1 Introduction

Modern biology sits at a crossroads: despite the wealth of data from complete genetic codes and vast single-cell atlases, such as the Human Cell Atlas [136] and scPerturb [123], our ability to predict cellular responses to drugs, mutations, or environmental change remains profoundly limited [174, 137]. The fundamental bottleneck isn't data volume, but a lack of models that capture how biological systems actually work [50, 97]. This gap motivates the vision of **AI virtual cells**: simulation-ready representations that reason about mechanisms, predict perturbation responses, and serve as in silico testbeds [20, 23, 116]. These models aspire to be biological world models, moving beyond simply reproducing observed data to answer critical "what if" and "how" questions. While recent biological Foundation Models (FMs), like GeneFormer [192] and scFoundation [57], show impressive predictive power, they often capture mere associations rather than causal mechanisms and are typically limited to a single biological layer.

The NeurIPS 2025 Workshop on AI Virtual Cells and Instruments: A New Era in Drug Discovery and Development (AI4D3 2025), San Diego, California, USA, 2025.

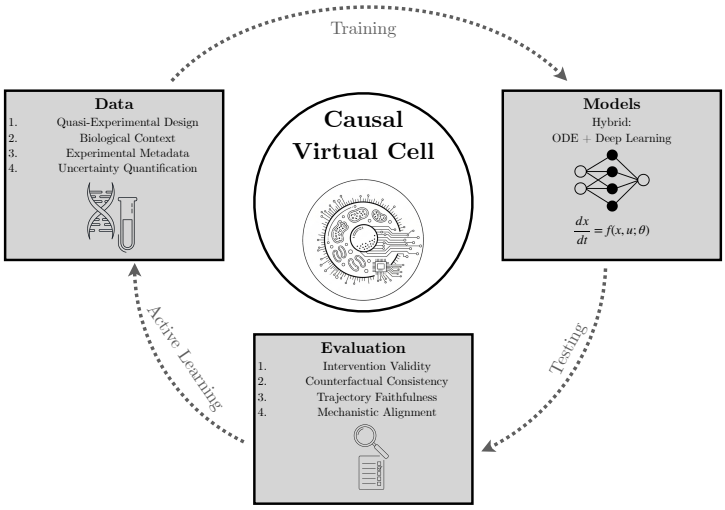

Figure 1: Summary of our proposed framework, which is described in Section 4.

Despite their variety, these models remain predictive rather than causal with evaluations centered on accuracy or likelihood rather than causal validity. Current benchmarks, often relying on scalar metrics like mean squared error (MSE) or coefficient of determination (R^2), reward pattern-matching but fail to test the model's response to direct interventions [129, 148]. Consequently, some advanced models struggle to generalize and even fail to outperform simple linear baselines on out-of-distribution perturbation tasks [137, 123]. Biology is inherently hierarchical—spanning the genome, transcriptome, and proteome—and disregarding this interdependence yields models fundamentally misaligned with the underlying reality [78, 69].

This raises a motivating question: *When does a predictive model of cells become a true world model, able to answer counterfactuals and generalize beyond its training distribution?* To achieve this requires the shift toward causal evaluation. Perturbation screens such as Perturb-seq [38] and Optical Pooled Screens [42] now generate the interventional data needed, but there is still no equivalent of ImageNet [35] or GLUE [172] for standardized causal assessment. Furthermore, because no dataset fully captures the multilayered complexity of the cell, **uncertainty** is an inherent property; evaluation must address not only whether a prediction is correct, but also how confident we should be in that prediction.

To this end, our contribution is twofold: (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. The taxonomy defines the axes of our proposed framework: **intervention validity, counterfactual consistency, trajectory faithfulness, and mechanistic alignment**, which are summarized in Figure 1. 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.

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 and 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

Predictive models primarily aim to reproduce observed cell states or transitions. These fall largely into three generative architectures:

(i) **Autoencoders and Conditionals** interpolate between cell states, including scGen [102], CPA [103], GEARS [138], scCade [117], and scPerb [162]. Variants like Biolord [131], CoupleVAE [176], and scVI [100] focus on enhancing disentangled latent representations.

(ii) **Generative Flow Models** apply techniques like GANs (MichiGAN [182]), flow-matching (CellFlow [79]), and optimal transport (CellOT [19]) to map cellular trajectories and generate disentangled single-cell data.

(iii) **Diffusion Models** [67, 152] have been adapted from image synthesis for single-cell imputation, denoising, and state simulation tasks.

Biological FMs build on the same generative principles but they distinguish themselves by pretraining scale and scope, enabling broader transferability across tasks. However, their evaluations remain primarily predictive.

- **Genomic/DNA** FMs are trained on DNA sequences to understand regulatory functions and predict genetic outcomes (e.g., Enformer [10], Geneformer [164], EVO2 [16]).
- **RNA** FMs learn sequence–structure relationships for tasks such as RNA structure/function prediction (RiNALMo [125], HydraRNA [92]), mRNA design (mRNA-FM [94]), and modification site detection (AIDO.RNA [193]).
- **Protein** FMs predict structures, attributes, and guide design (e.g., ProtGen [108], ESM-2 [96], AlphaFold 3 [1]).
- **Single-cell** FMs analyze omics data to model cellular states, useful for cell type annotation (scBERT [179], scGPT [33]) and perturbation prediction (scFoundation [57], CellFM [187]).
- **Multi-modal** FMs integrate layers, aiming to unify omic readouts (e.g., SCARF [98], LucaOne [60]) or capture cross-modality dynamics (scMultiSim [93], Xpressor [76]).

Despite the variety and scale, these models remain predictive rather than causal, with evaluations centered on accuracy or likelihood rather than causal validity.

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.** Observational atlases provide massive reference data. Examples include Tahoe-100M [188], Parse-PBMC [121], Tabula Sapiens [133], and the Human Cell Atlas [136]. Aggregation initiatives like CELLxGENE [132] and scBaseCount [181] combine hundreds of public datasets into harmonized resources.
- **Synthetic Data Generators.** These tools create controlled transcriptomic data for benchmarking predictive performance, often with known ground truth for association. Key examples include Splatter [186], SymSim [189], and scDesign3 [151].
- **Clinical and Phenotypic Data.** Predictive models often integrate macroscopic data to link cell states to disease, such as The Cancer Genome Atlas (TCGA) [165], UK Biobank [21], and EHR-derived datasets like the All of Us Research Program [5].

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.

Metrics. Predictive virtual cell models are typically evaluated using scalar metrics that quantify how well model outputs match observed data. Table 1 summarizes representative predictive evaluation methods, their objectives, and common datasets or implementations.

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 [177], scalar metrics on epigenome prediction often wash out signal and may encourage effects like "mode collapse." Rank-based interpretations

Table 1: Summary of predictive evaluation metrics.

Approach	Objective	Metrics
Sequence Modeling	Assess how well the predicted sequence distribution matches the ground truth; can serve as a proxy for gene essentiality [16]	Likelihood, log-probability; Kullback-Leibler (KL) divergence
Sequence Classification	Evaluate RNA FM classification of introns, exons, and splice variants by comparing predicted label distributions to true labels [25]	Cross-entropy (negative log-likelihood), accuracy, precision, recall, F1, Area Under the Receiver Operating Characteristic (AUROC)
Epigenome Prediction	Predict continuous expression or accessibility values and compare to experimental measurements [102, 103, 138, 117, 162]	MAE, MSE, R^2, cosine similarity; precision, recall, area under the precision–recall curve (AUPR)
Subcellular Localization	Compare predicted vs true spatial compartment labels using cluster-consistency measures on labeled 2D embeddings [56, 167, 155]	Adjusted Rand Index (ARI), Adjusted Mutual Information (AMI), label probability
Macroscopic Cell State Detection	Classify global cell properties (e.g., type, health/viability) against binary or logistic ground truth; mechanism of action detection follows the same scheme [16]	Recall, precision, F1, ROC-AUC; AUPR for class imbalance
Epigenome Delta Prediction	Predict perturbation-induced epigenetic deltas and compare against experimentally observed fold changes and differentially expressed genes (DEGs) [4, 117, 162]	(Log) fold-change, DEG overlap, directionality, Wilcoxon rank-sum, Top-k precision

(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 [102], CPA [103], GEARS [138]) 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 [113].

2.4 Limitations of Predictive Approaches

Predictive frameworks excel at interpolating and extrapolating trajectories but remain black boxes that lack mechanistic explanations [112]. They perform well on held-out data yet struggle to generalize to unseen perturbations or conditions [74, 163] and predict outcomes without testing causal guarantees or answering counterfactual questions [89].

Data and Context Limitations. These limits fundamentally reflect the data landscape: most resources are observational or transcriptome-only with few true interventions [135]; multi-omic and temporal datasets remain scarce [23]; and scRNA-seq yields only static snapshots, preventing necessary before–after comparisons [116]. Moreover, most datasets capture a single molecular layer, leaving genome-to-proteome mechanisms unevaluated, while the combinatorial complexity of perturbations demands coordinated community efforts [163].

Evaluation Misalignment and Uncertainty. Evaluation is likewise dominated by predictive metrics such as MSE, R^2, and log-likelihood, which capture correlations but not mechanisms [53]. Even perturbation benchmarks emphasize regression measures, which are insufficient for mechanistic alignment [116]. While newer metrics like uncertainty quantification (e.g., calibration error [180]), distributional similarity (e.g., MMD [55]), and rank-based evaluation (e.g., LogFC rank in PerturBench [123]) represent progress, they still treat predictions as point estimates. Uncertainty is, in fact, a crucial cross-cutting dimension: low-confidence predictions signal the need for more data or model refinement, shaping how validity, consistency, and mechanistic alignment are ultimately interpreted.

Predictive models capture correlations but not mechanisms. They perform well on familiar conditions yet struggle with unseen perturbations and causal questions. Without explicit tests of

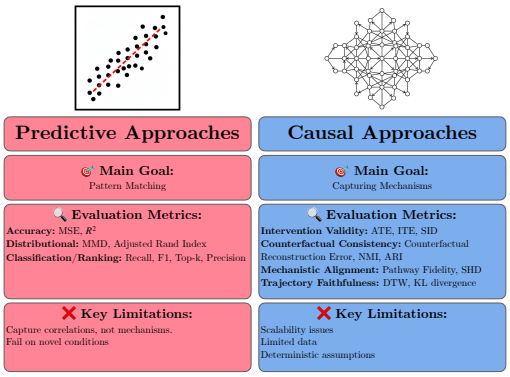

Figure 2: Comparison of predictive (Section 2) and causal (Section 3) approaches.

intervention validity or mechanistic grounding, predictive evaluation remains a measure of pattern matching rather than understanding.

3 Causal Methods

Compared to predictive methods that reproduce observed patterns, causal models aim to capture cause-and-effect relationships and are judged on whether they reproduce intervention outcomes or generate counterfactuals consistent with known mechanisms [185, 23, 122, 12, 114]. Figure 2 provides a visual comparison of both approaches. While interpretability and causality are related, they are not synonymous. **Interpretability** methods (e.g., attention mechanisms, SHAP, or feature attribution) help reveal associations or model heuristics, but they do not necessarily establish cause and effect relations. **Causal inference** extends interpretability by testing whether explanations remain valid under interventions or context shifts, providing stronger guarantees for biological understanding and decision-making [50, 112, 148]. Causal machine learning offers a path forward by treating perturbations as structured interventions and seeking mechanisms invariant across environments [51, 163].

Causality in biology can be defined in complementary ways: **(i) Mechanistic view:** emphasizes biochemical interactions and dynamical processes (e.g., MAPK phosphorylation cascades that link receptor activation to downstream gene expression) [163]. **(ii) Probabilistic view:** emphasizes conditional independences in observational data (e.g., ERK activation being independent of receptor status once Ras activity is accounted for) [51]. **(iii) Counterfactual view:** highlights potential outcomes under interventions (e.g., asking how a tumor cell's transcriptome would change if KRAS were knocked out versus left intact) [99].

3.1 Causal Models

Structural Causal Models (SCMs) represent variables as directed graphs with explicit rules for interventions via the do-operator [122, 135]. Dynamical causal models extend this using ordinary or stochastic differential equations (ODEs, SDEs) to describe how biological states evolve under perturbation [163]. These perspectives form the foundation for causal virtual cells: models that predict and explain cellular responses by explicitly capturing the mechanisms that remain invariant across conditions. We highlight four families of causal models relevant to virtual cells.

- **ODE-Based Models.** Use ODEs to describe biochemical networks [78, 6]. Classical examples include models in electrophysiology and metabolism [68, 157, 115, 31, 65, 61, 2]. Recent methods adapt to single-cell Gene Regulatory Network (GRN) inference and trajectories:
  - GRN/Trajectory Inference: SCODE [110], GRISLI [9], SINCERITIES [119].
  - RNA-Velocity Extensions: scVelo [14], UniTVelo [46], Velorama [150], DynaMO [84].

  - Structured Dynamics: GraphDynamo [190] and STORM [124] add graph or stochastic structure. Stochastic Differential Equations (SDEs) help model noise but raise identifiability challenges [82, 17, 126].
- **Hybrid Causal Deep Learning Models.** Address scalability by integrating neural networks with mechanistic constraints:
  - Differential Equations (DE) Hybrids: Neural ODEs [26], Latent ODEs [142], and Universal Differential Equations (UDEs) [134] parameterize dynamics or embed neural nets within structured equations.
  - Single-Cell Applications: DeepVelo [27], PerturbODE [95], and PHOENIX [70].
  - Constraint-Based Models: Knowledge-primed neural networks, including sparse MLPs and graph-informed architectures, constrain learning with pathway priors [44].
- **Graphical and Counterfactual Approaches.** Represent cellular dependencies as Directed Acyclic Graphs (DAGs) or SCMs:
  - Causal Discovery Algorithms: Constraint-based (Peter-Clark, Fast Causal Inference [153, 154]), score-based (GES [29]), and differentiable DAG learners (NOTEARS [191], DAG-GNN [183], GraN-DAG [87]).
  - Single-Cell Inference: CausalCell [174], LINEAGEOT [43], and CARDAMOM [184].
  - Invariance-Based Methods: ICP [128], Causal Dantzig [140], and anchor regression [141], which identify gene modules stable across environments.
- **Causal Perturbation Prediction Models.** Embed causal structure directly into predictive architectures for counterfactual simulation:
  - Variational/Generative Methods: scCausalVI [7] (disentangles heterogeneity from treatment effects) and CausCell [47] (SCMs with diffusion modeling).
  - Optimal Transport/Factor Graphs: CINEMA-OT [39] (separates confounding) and DCD-FG [101] (infers factor graphs with causal constraints).
  - Latent Space Alignment: GPO-VAE [11] (aligns VAE latent spaces with GRN priors) and GraphVCI [178] (predicts counterfactual responses on graphs).

3.2 Causal Data

Unlike predictive resources (Section 2.2), causal modeling requires data with explicit interventions, perturbations, or synthetic counterfactuals. These form the basis for testing whether models capture cause and effect relationships rather than correlations.

- **High-Throughput Single-Omic Perturbation Screens.** Provide the closest analogue to randomized controlled trials in cell biology [89]. CRISPR-based screens (e.g., Perturb-seq [38, 3] and X-Atlas [72]), and Optical Pooled Screens (OPS) [42] are the cornerstone of interventional single-cell data. These resources enable direct measurement of cellular responses to interventions, though they remain sparse and noisy.
- **Emerging Multi-Omic Perturbation Data.** For robust causal inference, single-modality measurements are often insufficient, as mechanisms span multiple regulatory layers. Emerging multi-omic data, including joint measurements of RNA and protein (perturbational CITE-seq [156, 58]) and chromatin accessibility (Perturb-ATAC [143]), will be crucial.

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.

Metrics. Causal metrics test counterfactual validity and mechanistic fidelity. Table 2 summarizes commonly used objectives and representative metrics across evaluation approaches.

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.** These use simulation frameworks such as GeneNetWeaver [147], SERGIO [36], and scDesign3 [151] to generate datasets with known causal graphs, enabling precise quantification of GRN recovery and counterfactual consistency.

Table 2: Summary of causal evaluation metrics.

Approach	Objective	Metrics
Intervention Validity	Reproduces observed outcomes under experimental interventions (e.g., CRISPR knockouts, drug perturbations)	**(i) Causal effect estimation**: Individual, Average, and Conditional Average Treatment Effects (ITE, ATE, CATE), log Fold-Change (LogFC) [66, 149, 175, 63] **(ii) Attribution:** regression coefficients to verify correct attribution to latent or confounding factors [75, 105, 148] **(iii) Distributional alignment:** Structural Intervention Distance (SID) [145, 127, 59], Maximum Mean Discrepancy (MMD) [55], energy distance [160], (ARI) [73]
Counterfactual Consistency	Biological plausibility and mechanistic grounding of counterfactual predictions, consistent with simulated and experimental causal effects	**(i) Reconstruction error:** Pearson correlation, Mean Squared Error (MSE), Normalized Mutual Information (NMI), ARI, marker gene preservation [48, 102] **(ii) Latent disentanglement:** clustering, silhouette indices for separability of causal factors [13, 47, 7] **(iii) Ground-truth agreement:** GeneNetWeaver, SynTReN, PerturBench; Sachs flow cytometry, Perturb-seq [147, 170, 177]
Mechanistic Alignment	Correspondence between inferred mechanisms and curated biological pathways and constraints	**(i) Pathway fidelity:** KEGG and Reactome overlap [77, 41] **(ii) Invariance tests:** stability across perturbations, modalities, and contexts (conditional independence checks, out-of-distribution generalization) [128, 62] **(iii) Graph similarity:** SID, Structural Hamming Distance (SHD)
Trajectory Faithfulness	Alignment between predicted and observed time-resolved responses, capturing the shape, timing, and magnitude of trajectories under perturbation	**(i) Trajectory similarity**: Dynamic Time Warping (DTW), KL divergence, optimal transport [34, 26] **(ii) Trend alignment:** Pearson correlation, MSE, RMSE [102, 7] **(iii) Structural consistency:** SID, SHD, graph recovery [128] **(iv) Benchmarks:** Perturb-seq, OPS, DREAM4, SynTReN [54, 170]
GRN Recovery	Recovery of causal and statistical structure in GRNs	**(i) Edge prediction:** AUROC, AUPR **(ii) Graph distance:** SHD, SID **(iii) Benchmarks:** DREAM4, GeneNetWeaver

- **Pathway Fidelity Tasks.** These evaluate whether models preserve mechanistic structure by testing predicted perturbation effects against curated biological pathways (e.g., KEGG [77], Reactome [41], BioModels [90]).
- **Invariance Tests.** This crucial strategy links causal reasoning to robustness. Invariance-based evaluation tests whether predictions remain stable across environments or cell contexts, using frameworks such as ICP [128] and anchor regression [141]. This directly addresses Domain Adaptation: both aim to identify mechanisms that remain stable across environments, linking causal reasoning with robustness objectives formalized in approaches such as invariant risk minimization [8].
- **Generalization Regimes and Domain Adaptation.** These tasks (e.g., unseen single perturbations, novel combinations, and temporal holdouts) require causal consistency [102, 148]. The pursuit of causal invariance in these regimes is intrinsically linked to Domain Adaptation. Both aim to identify mechanisms that remain stable across environments, thereby linking causal reasoning with robustness objectives formalized in approaches such as invariant risk minimization [8] and anchor regression [141].
- **Baselines and Ablations.** Causal models are compared against predictive-only baselines (e.g., scGen [102], CPA [104]) to test whether causal inductive biases improve counterfactual validity. Component ablations clarify which features drive causal performance [7, 47].

- **Standardized Benchmarks.** These enable systematic evaluation of interventional datasets. Perturbation benchmarks like PerturBench [177] and OP3 [159] provide standardized tasks, with OP3 emphasizing causal criteria. General causal benchmarks such as CausalBench [173] provide broader reference standards for evaluating causal inference methods across domains.

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 [20, 23, 88, 116].

Model Limitations and Assumptions. Many ODE-based and hybrid methods assume acyclicity or causal sufficiency [111, 174, 163], which restricts the modeling of feedback loops and hidden confounders. They also rely on idealized interventions and face unresolved parameter identifiability challenges [80]. Consequently, most approaches remain confined to small circuits, velocity-style embeddings, or low-dimensional summaries rather than the necessary genome-wide, multi-omic contexts [51, 99, 88].

Data Scarcity and Quality. Causal data availability remains a bottleneck [23]. Perturbation assays such as Perturb-seq and Optical Pooled Screens (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 [28].

Fragmentation and Deterministic Outcomes. Evaluation remains fragmented: current efforts emphasize GRN recovery, pathway fidelity, or counterfactual validation, but no unified taxonomy of causal metrics exists for virtual cells [20]. Critically, most evaluations treat outcomes as deterministic, even though biological systems 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.

Causal models introduce interpretability and mechanistic rigor but remain constrained by data scarcity, strong assumptions, and limited scalability [135]. Progress will depend on unifying benchmarks, propagating uncertainty, and coupling causal models with experimental feedback to move from proof-of-concepts to reliable virtual cells.

4 Proposed Framework

The ambition for virtual cells is to represent cellular machinery in mechanistic detail, ideally as systems of differential equations capturing causal interactions and dynamics [81, 6]. However, ODEs assume deterministic dynamics and face the "curse of dimensionality," making whole-cell simulation infeasible [171, 166]. Progress requires hybrids that combine mechanistic grounding with deep learning flexibility. Universal and neural ODEs [134, 26] integrate biological priors with neural architectures, while causal constraints, sparsity, and disentangled representations improve interpretability [18, 7]. Crucially, model design is inseparable from evaluation: benchmarks must test not only predictive accuracy but also causal validity [129, 148], ideally within a lab-in-the-loop paradigm where models are iteratively refined with experiments [45, 24]. These design principles are essential for the next generation of biological FMs. By incorporating interventional objectives and causal evaluation metrics during pretraining, such models could move beyond descriptive fit to learn mechanistic invariances across tissues, species, and modalities.

4.1 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, but most widely available datasets are observational [135]. 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 [139], paired sampling [144], and distributional methods like optimal transport [130] (exemplified by CINEMA-OT [39]) 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 [169, 146], RNA velocity [86, 14], dose–response designs [158], 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 [77], Reactome [41], STRING [161], and BioGRID [118], which provide pathway and interaction knowledge for fidelity tests. Meanwhile, ontologies such as GO [30] and Cell Ontology [37]) enable dataset alignment, and domain-specific language models like BioBERT [91] enrich metadata. (ii) *Synthetic data-based* tools such as GeneNetWeaver and DREAM [147, 109], SERGIO [36], DYNGEN [22], and scDesign3 [151] 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 [7], [47], [83], [58], and [100]. The following strategies help prepare datasets to provide this context: (i) *Metadata integration* on batch effects, protocols, and sample handling (GEO [40], ArrayExpress [120], CELLxGENE [132]) can stratify analyses; protocol-aware covariates improve comparability across assays (e.g. 10x vs. Smart-seq2) [64]. (ii) *Quality control and robustness*, such as UMIs, features, mitochondrial fraction, improve reliability [106], and invariance-based methods such as ICP [128] and anchor regression [141] 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 [71]) (vi) Simulation-based inference likelihood-free methods [32] quantify uncertainty in complex mechanistic models, with applications to stochastic gene expression [168], signaling dynamics [52], and single-cell electrophysiology [107]. 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.

4.2 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 [66] could be extended with calibration (e.g., expected calibration error or ECE [113], Brier score [49]), 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 [15] or KL divergence [85] 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 [77], Reactome [41]) 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. While data generation and model innovation remain crucial, evaluation defines whether progress is measurable and reproducible. Our focus on causal evaluation does not exclude these other challenges but provides the missing layer of accountability connecting them. 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

[1] 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.

[2] 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, pages 233–257. Elsevier, 2011.

[3] 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.

[4] 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`.

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

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

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

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

[9] 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.

[10] 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`.

[11] 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*, 2025.

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

[13] 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.

[14] 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.

[15] 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*, pages 359–370, 1994.

[16] 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*, pages 2025–02, 2025.

[17] 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.

[18] 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.

[19] 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.

[20] 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.

[21] 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.

[22] 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.

[23] 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.

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

[25] 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*, 2022. URL `https://api.semanticscholar.org/CorpusID:247922548`.

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

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

[28] 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.

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

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

[31] 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.

[32] 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.

[33] 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.

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

[35] 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*, pages 248–255. Ieee, 2009.

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

[37] 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.

[38] 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.

[39] 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.

[40] 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.

[41] 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.

[42] 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.

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

[44] 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.

[45] 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*, pages 2025–02, 2025.

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

[47] 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.

[48] 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.

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

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

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

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

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

[54] 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.

[55] 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.

[56] 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`.

[57] 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.

[58] 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.

[59] 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.

[60] 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*, pages 2024–05, 2024.

[61] 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.

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

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

[64] 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.

[65] 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.

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

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

[68] 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.

[69] 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.

[70] 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.

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

[72] 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*, pages 2025–06, 2025.

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

[74] 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.

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

[76] Jeremie Kalfon, Laura Cantini, and Gabriel Peyre. Towards foundation models that learn across biological scales. *bioRxiv*, pages 2025–05, 2025.

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

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

[79] 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*, pages 2025–04, 2025.

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

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

[82] 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.

[83] 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.

[84] 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.

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

[86] 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.

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

[88] 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.

[89] 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.

[90] 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.

[91] 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.

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

[93] 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*, pages 1–12, 2025.

[94] 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 2025. ISSN 1362-4962. doi: 10.1093/nar/gkaf044. URL `https://doi.org/10.1093/nar/gkaf044`.

[95] 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.

[96] 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.

[97] 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.

[98] 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*, pages 2025–04, 2025.

[99] 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.

[100] 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.

[101] 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.

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

[103] 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.

[104] 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.

[105] 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.

[106] 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.

[107] 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.

[108] 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.

[109] 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.

[110] 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.

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

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

[113] 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.

[114] 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.

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

[116] 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.

[117] 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)*, pages 441–444. IEEE, 2024.

[118] 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.

[119] 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.

[120] 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.

[121] 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/`.

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

[123] 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.

[124] 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.

[125] 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`.

[126] 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.

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

[128] 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.

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

[130] 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.

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

[132] 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.

[133] 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*, pages 2024–12, 2024.

[134] 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.

[135] 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.

[136] 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.

[137] 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.

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

[139] 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.

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

[141] 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.

[142] 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.

[143] 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.

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

[145] 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.

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

[147] 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.

[148] 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.

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

[150] 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.

[151] 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.

[152] 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.

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

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

[155] 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`.

[156] 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.

[157] 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.

[158] 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.

[159] 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.

[160] 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.

[161] 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.

[162] 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.

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

[164] 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.

[165] 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.

[166] 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.

[167] 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`.

[168] 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.

[169] 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.

[170] 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.

[171] 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.

[172] 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.

[173] 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)*, pages 143–151, 2024.

[174] 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.

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

[176] 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.

[177] 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.

[178] 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.

[179] 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.

[180] 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.

[181] 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*, pages 2025–02, 2025.

[182] 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.

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

[184] 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.

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

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

[187] 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.

[188] 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*, pages 2025–02, 2025.

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

[190] 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*, pages 2023–09, 2023.

[191] 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.

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

[193] 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*, pages 2024–11, 2024.