nature genetics

Perspective

https://doi.org/10.1038/s41588-025-02124-2

Causal machine learning for single-cell genomics

Received: 4 December 2023

Accepted: 10 February 2025

Published online: 31 March 2025

Check for updates

Alejandro Tejada-Lapuerta[1,2,10], Paul Bertin[3,4,10], Stefan Bauer[2,5,6], Hananeh Aliee[7] ✉, Yoshua Bengio[3,4,8] ✉ & Fabian J. Theis[1,2,9] ✉

Advances in single-cell '-omics' allow unprecedented insights into the transcriptional profiles of individual cells and, when combined with large-scale perturbation screens, enable measuring of the effect of targeted perturbations on the whole transcriptome. These advances provide an opportunity to better understand the causative role of genes in complex biological processes. In this Perspective, we delineate the application of causal machine learning to single-cell genomics and its associated challenges. We first present the causal model that is most commonly applied to single-cell biology and then identify and discuss potential approaches to three open problems: the lack of generalization of models to novel experimental conditions, the complexity of interpreting learned models, and the difficulty of learning cell dynamics.

Cells are the basic unit of life, with their functions and identities shaped by complex physical and biochemical processes. Understanding causal relationships within cellular processes is essential for revealing the intricate biological mechanisms that drive cellular behaviors such as proliferation, differentiation and apoptosis, and for identifying the associated signaling molecules, genetic mutations or environmental stimuli, as well as for aiding the development of targeted therapies for diseases such as cancer, neurodegenerative disorders and metabolic syndrome.

Advances in molecular profiling at single-cell resolution have provided an unprecedented view of cellular processes. Machine learning has begun to be applied to single-cell genomics, with crucial contributions, such as dimensionality reduction[1,2] (used mainly for visualization purposes), data integration[3] (to construct cell atlases), trajectory inference[4] (for inferring cell fate) and transfer of model predictions across modalities[5]. However, such methods have only provided limited insights into the underlying biological processes and do not lead to improved predictions of experimental outcomes.

The majority of the machine learning methods applied to single-cell genomics have their foundations in noncausal statistical learning, which leverages patterns within a unique data distribution. However, when the experimental conditions change, the data distribution changes accordingly; so previously identified patterns may not be relevant anymore, and noncausal statistical learning may fail to generalize[6] (Box 1). Causal machine learning (depicted in Fig. 1a) aims to achieve good predictions under novel conditions by discovering the biological mechanisms that correspond to (a chain of) biochemical interactions through which one biological quantity affects another. This is to be opposed to spurious correlations (which do not reflect underlying biochemical interactions) that are captured via statistical learning and typically lead to poor generalization in changed conditions (Box 1). Most biological mechanisms are expected to remain unchanged even when experimental conditions vary (akin to the immutability of the rules of physics) in contexts such as small-molecule or CRISPR perturbations (as highlighted in Fig. 1b), which constitute the focus of this Perspective. In some cases, however, such as changes in temperature and pressure, most mechanisms can be directly affected, and the dependency of causal mechanisms on such factors should be learned. More broadly, mechanisms are subject to change if any aspect of the experimental conditions, whether technical or biological, changes too drastically.

[1]Institute of Computational Biology, Helmholtz Munich, Munich, Germany. [2]School of Computing, Information and Technology, Technical University of Munich, Munich, Germany. [3]Mila, the Quebec AI Institute, Montreal, Quebec, Canada. [4]Université de Montréal, Montreal, Quebec, Canada. [5]Helmholtz Munich, Munich, Germany. [6]Munich Center for Machine Learning (MCML), Munich, Germany. [7]Wellcome Sanger Institute, Hinxton, UK. [8]Learning in Machines and Brains Program, Canadian Institute for Advanced Research, Toronto, Ontario, Canada. [9]TUM School of Life Sciences Weihenstephan, Technical University of Munich, Munich, Germany. [10]These authors contributed equally: Alejandro Tejada-Lapuerta, Paul Bertin. ✉e-mail: ha10@sanger.ac.uk; yoshua.bengio@mila.quebec; fabian.theis@helmholtz-munich.de

https://doi.org/10.1038/s41588-025-02124-2

BOX 1

Definitions and key concepts

Causality

- **Biological causality** aims to find mechanistic explanations, relying on a (cascade of) physical interactions and chemical reactions between biological entities inside the cell, that lead a cause to have a particular effect.
- **Probabilistic causality** aims to find statistical association between variables (typically gene expression levels) that remain unchanged unless they are directly intervened on.

These two notions of causality are aligned well in the context of learning causal models from single-cell data, as biochemical interactions (through which one biological quantity affects another) are reflected in robust statistical associations.

Causal model
A causal model is able to generate an entire family of distributions; each distribution corresponds to a different environment (for example, experimental condition). It is usually a pair (g, h) composed of:

- **A graph-based model** g, which encodes explicit directed relationships between causal variables (associated with the nodes of the graph) and can generate observations in a given environment.
- **An interventional model** h that modifies the graph-based model such that it can generate samples from different environments. Given an intervention i and the current (unintervened) parameters of the graph-based model, h generates intervened parameters such that g generates observations under intervention i. Typically, h modifies the adjacency matrix of the graph-based model and removes some edges.

Environment
The environment is defined by all the characteristics of the cell population on which the experiment is performed, as well as the characteristics of the experimental protocol, including exposure to biological perturbations.

Intervention
An intervention is any action performed by the interventional model over the graph-based model.

SCM[19,82]
A type of graph-based model in which the value of each variable is generated through a so-called structural assignment, which takes the value of its parents as input:

$$X_i := f_i(\mathrm{PA}_i, U_i), (i = 1, \ldots, n),$$

where f_i is a deterministic function, PA_i is the set of causal parents of node X_i in the graph and U_i is a noise variable that represents variability within the population of cells. The members of the set of noise variables $U_1, \ldots, U_n$ are jointly independent. Most importantly, the graph (with nodes $X_1, \ldots, X_n$ and edges going from parents PA_i to X_i for all i) is required to be acyclic.

Causal kinetic model[69]
A type of graph-based model in which structural assignments govern the temporal evolution of causal variables via an ODE or a stochastic differential equation (to account for the intrinsic stochasticity of biology).

Generalization
Ability of a model to make accurate predictions in new, previously unseen environments.

Latent variable
Variable that is not directly observed but inferred from other observed variables. It can capture biological quantities such as pathway activation or aspects of the acquisition protocol.

A latent variable is considered causal if it plays a role similar to that of observed causal variables within the causal model: namely, they depend on and influence other causal variables (whether latent or observed) and the mechanisms among them are preserved across environments.

Conditional independence testing
Variables A and B are considered conditionally independent given C if $P(A|B,C)=P(A|C)$.

Granger causality
A statistical hypothesis test that aims to determine whether a time series is useful for forecasting another.

GRN
A GRN is a graph representing interactions among genes (and sometimes other molecular regulators) that govern the expression levels of mRNA and proteins, which in turn determine the functions of the cell. The links in the network reflect cascades of biochemical interactions involved in gene regulation mechanisms (for example, transcriptional regulation, post-transcriptional modifications).

This mechanistic definition is stricter than the correlative approaches that some available GRNs may rely on.

ODE
An ODE models how a variable changes over time in relation to others, often used in biology to describe temporal dynamics of gene expression or protein concentrations based on causal relationships:

$$dX = f(\mathrm{PA}_X, t)dt \text{ or } \frac{dX}{dt} = f(\mathrm{PA}_X, t),$$

where PA_X represents the causal parents of the variable X.

Stochastic differential equation
A stochastic differential equation extends an ODE by adding a noise term, capturing random fluctuations inherent in biological systems, making it ideal for modeling stochastic behaviors in single-cell systems:

$$dX = f(\mathrm{PA}_X, t)dt + \sigma(X, t)d\varepsilon,$$

where PA_X represents the causal parents of the variable X and $\sigma(X,t)d\varepsilon$ is the stochastic term.

To have an impact in cell biology, causal machine learning must be adapted to the specificities of the biological systems being modeled and the data modalities being used. Moreover, causal inference methods can fail when some of their core independence assumptions (such as stable unit treatment value assumption[7] and no hidden confounding[8]) are violated or when the model is misspecified (for example, assuming complete observability of the system or assuming linear regulatory mechanisms while real ones are known to be nonlinear). The violation of assumptions might lead to wrong predictions in unseen conditions and provide inaccurate insights. Uncertainty estimation can help avoid highly confident but wrong predictions but might still lead to inaccurate predictions when core assumptions are violated and not relaxed.

In genomics, there has long been interest in discovering interactions between genes to provide mechanistic explanations of biological processes, often summarized through module networks that group genes (that is, modules) that function together and for which expression is tightly correlated[9] or through gene regulatory networks (GRNs) (Box 1) that contain directed connections from regulator to regulated genes. Furthermore, mechanistic and dynamical approaches from systems biology, traditionally applied to small-scale data such as results from western blotting and quantitative PCR, are now being adapted to large-scale genomic data[10]. Inference approaches range from the use of conditional independence testing[11,12] to detect pairs of genes that are in direct interaction with one another to Granger causality[13] (Box 1) used for the analysis of time series, as well as methods that directly try to predict a graph from experimental data using black box approaches[14]. Large research efforts are aimed at improving GRN inference[15] via heuristics using multimodal data[16,17] (for example, restricting the set of possible edges in the GRN on the basis of the accessibility of transcription factor binding sites measured by single-cell ATAC-seq (assay for transposase-accessible chromatin using sequencing) and prior knowledge, such as known transcription factor binding sites. Validating inferred GRNs has been a major challenge[15], particularly in human cells, in which the true GRN remains mostly unknown and is highly context dependent. Other organisms, such as *Escherichia coli*, are better understood, and databases for GRNs exist[18] but are still noisy and incomplete.

The increasing availability of perturbational data may enhance the applicability of causal approaches to transcriptomics. This Perspective aims to identify and analyze open problems in the field, as well as to place them into perspective with ongoing research directions. After providing some background on causal inference techniques in genetics and transcriptomics, we present the default causal model that underlies most current causal approaches to single-cell biology. We then discuss three open problems, namely, the lack of generalization to novel experimental conditions, the complexity of interpreting learned models and the difficulty of learning cell dynamics.

Causality for transcriptomics

Single-cell resolution offers an unprecedented view of how biological processes unfold at the cellular level. It provides fine-grained detail of cellular heterogeneity, allowing the discovery of distinct mechanisms operating within different cell types. This level of resolution enables the identification and analysis of rare cell populations and specific responses to perturbations that might be overlooked in bulk analyses. However, this resolution comes with the trade-off of noisier observations and biases such as technical dropout.

A single-cell experiment typically involves a population of cells belonging to a certain environment. Here, the notion of environment represents the characteristics of the cell population (for example, cell type information), as well as information on the experimental protocol, such as exposure to a biological perturbation or the devices used to perform the experiment. This definition of the environment corresponds to the terminology used in the causality community and is broader than the notion of the extracellular environment used in cell biology.

Causal models can usually be broken down into two components (Box 1). The first component models biological mechanisms (usually through a causal graph), while the second component models the way these mechanisms are affected by biological perturbations, typically an edge-removal operation (that is, the perturbation is assumed to remove some specific interactions that are no longer represented by edges in the causal graph), or how mechanisms change across environments.

Causal graph of a cell

Within a cell, biological mechanisms can be portrayed as a causal graph with nodes denoting gene expression levels and edges indicating causal relationships between genes. An edge is directed from a 'parent' to a 'child' node and means that the gene expression level of the child node depends on the expression level of the parent node. This representation, called a structural causal model (SCM)[18,19] (Box 1), defines nodes as causal variables and the functions that govern the expression of a gene in relation to its parents as causal mechanisms. In the context of cell biology, causal mechanisms correspond to biological mechanisms, typically transcriptional regulation (Fig. 2a). Most causal approaches to single-cell genomics have been based on this model and the assumptions it entails[12,18–22]. Note that causal models do not always rely on an explicit graph. In some cases, the graph is not explicitly constructed but can be recovered through the model's internal dependencies[23].

The default SCM model has several limitations. First, the causal graph must be acyclic to be able to generate synthetic cell observations: the gene expression levels of the root nodes (that is, nodes with no incoming edges) are sampled first, followed by the expression levels of their direct children (conditionally on the value of the parents), and so on, until all nodes have been sampled. This contrasts with real GRNs, in which cyclic interactions are commonly found as part of regulatory motifs, such as autoregulation or feedback loops[24]. Additionally, the default SCM lacks a temporal dimension, limiting its ability to capture the dynamic aspect of transcription regulation. Finally, this model only accounts for gene expression levels, but, in reality, numerous other variables, such as the levels of transcription factors and the proportions of different splicing variants for a given gene, also play a role in transcriptional regulation.

Biological perturbations as causal interventions

A biological perturbation refers to a disturbance or alteration in the normal functioning of biological systems, often induced experimentally to study the system's response and understand underlying mechanisms. In an SCM, the impact of a perturbation can be replicated in the causal graph through a so-called intervention by manipulating the specific variables or causal mechanisms targeted in the experiment. The commonly assumed perfect intervention (illustrated in Fig. 2b) removes the dependency of the intervened causal variable on its causal parents. It is set to zero to signify a complete loss of function in the targeted transcript, independent of the value of its regulators.

In practice, the applicability of the perfect intervention assumption is limited. Evidence indicates that CRISPR knockouts have off-target effects[24–28] and may fail to edit the genome. Methods have been proposed to identify cells within a dataset that have not been affected by the knockout[29]. Similarly, drug perturbations cannot usually be approximated by perfect interventions. This is due to the fact that their mechanisms of action are not always known, and, in numerous instances, the drug does not directly affect transcription regulation mechanisms.

The exact nature of biological perturbations largely remains unknown, and different modeling choices can be relied on, which we refer to as interventional models, to successfully represent the impact of perturbations on the causal graph. In some contexts (for example,

https://doi.org/10.1038/s41588-025-02124-2

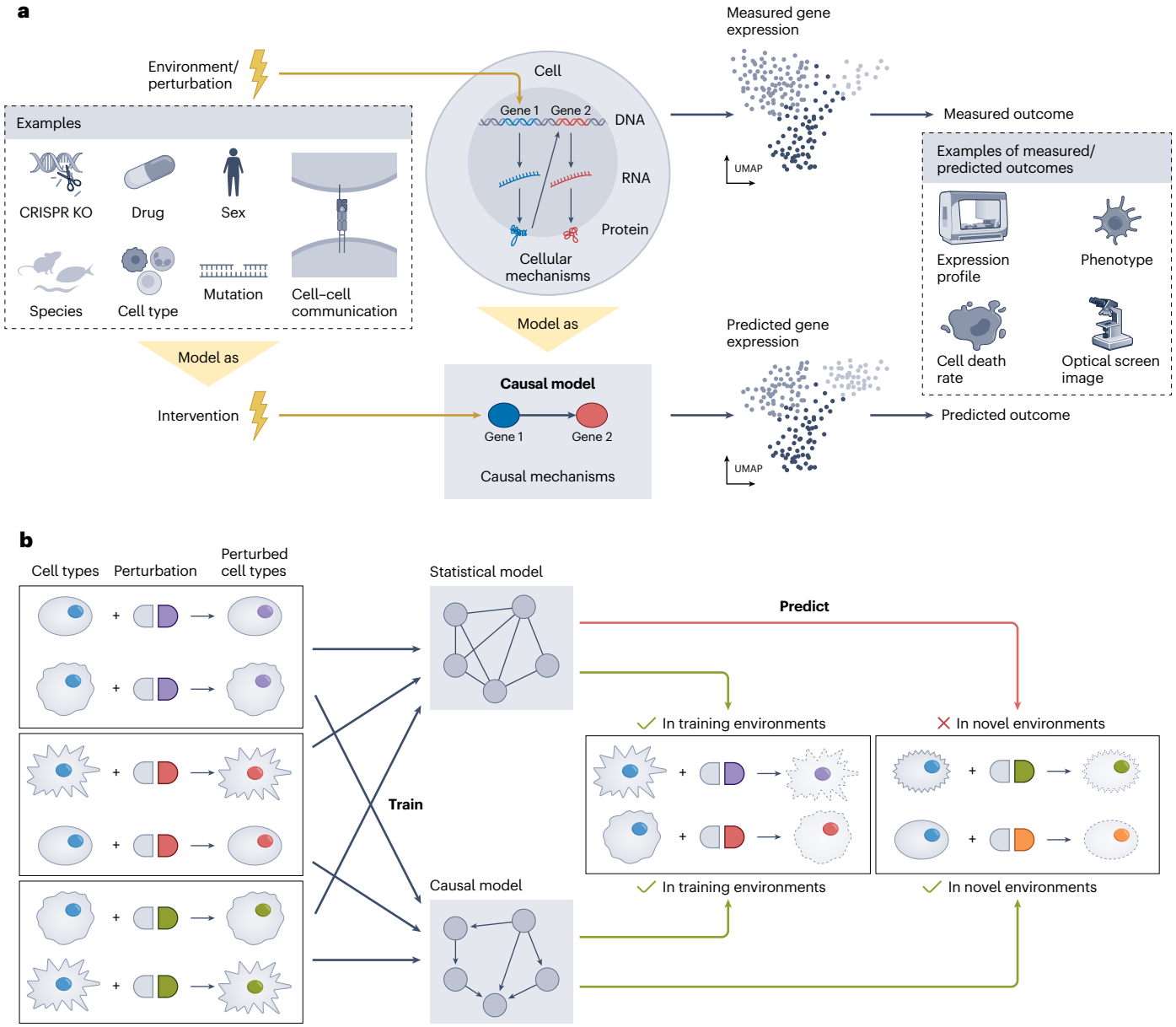

Fig. 1 | Causal modeling of the cell. a, Biological perturbations or environmental changes affect cellular mechanisms, leading to altered gene expression and other measurable outcomes. These alterations in cellular mechanisms can be simulated through interventions within a causal model to predict gene expression and outcomes under different conditions. KO, knockout; UMAP, uniform manifold projection and approximation. **b**, Using a dataset of observations from various cell types affected by different biological perturbations, statistical learning methods (shown at the top) can predict outcomes for cell types and perturbations present in the training data. However, they often fail to accurately predict outcomes for novel perturbations or novel cell types, a challenge tackled by causal learning methods (shown at the bottom).

small-molecule perturbations), it might be more appropriate to assume that changes in cell state correspond to a new state within a static graph (except for the directly altered mechanism). In other contexts (for example, new cell types), it may be more suitable to assume that many mechanisms on the graph, and possibly the graph's structure itself, have changed. Different approaches can be considered: from interventional models that remove or modify the parameters of more than one edge at a time (which can help to model off-target effects, as depicted in Fig. 2c) to interventional models that take uncertainty into account or create new edges within the graph. This means that each type of intervention requires a modeling framework that resembles as closely as possible the true biological perturbation.

Below, we identify and discuss three open problems associated with the use of causal models in single-cell genomics.

Predicting the outcome of novel experimental conditions

One of the grand challenges of computational biology is to develop models that can predict the outcome of novel experimental conditions, for instance, predicting the effect of a disease on unseen cell states or the effect of an unseen drug[5,30,31]. Below, we discuss the importance of having access to high-quality data from diverse environments to train a causal model that effectively translates knowledge to novel environments and how such models can be used for the prospective acquisition of informative perturbations.

Importance of the diversity of observed perturbations

Advanced interventional models such as the one depicted in Fig. 2c have parameters that should be learned on the basis of data across

https://doi.org/10.1038/s41588-025-02124-2

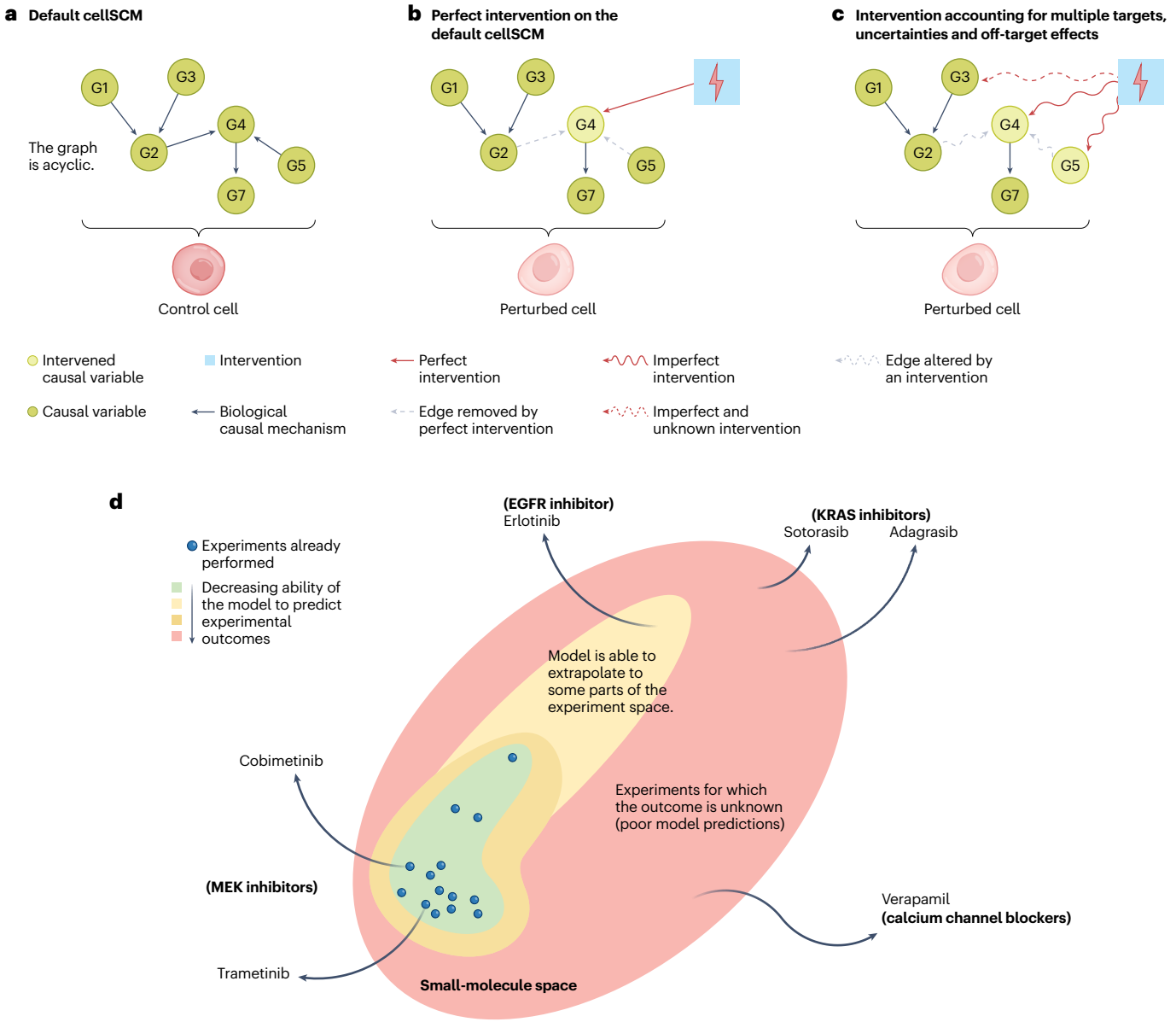

Fig. 2 | Overview of the default SCM model and strategies for learning generalizable causal models of cells. a, Modeling of a cell through a model termed the default SCM (cellSCM). Edges represent causal relationships between genes, corresponding to transcriptional regulation. In the absence of any intervention on the cell, the model generates expression profiles associated with control cells. Here, genes G1 and G3 are causal parents of G2, meaning that the value of G2 is determined on the basis of the values of G1 and G3. Cyclic interactions are not allowed in the graph, which prevents, for instance, G2 being a causal parent of G1 or G3. **b**, Perturbations in an SCM are naively treated as perfect interventions, which completely abolishes the dependence of targeted variables on their regulators. Therefore, the expression level of gene G4 is directly set by the intervention, as it no longer depends on the expression of G2 and G5. **c**, A more accurate means to model biological perturbations allows several targets to account for uncertainties in the interventional targets, as well as for an incomplete removal of regulator dependency. **d**, Schematic overview of model capabilities across different regions of the experimental space, using small-molecule effect prediction as an example. A model trained primarily with MEK inhibitors may be able to extrapolate predictions to compounds that act over similar pathways, such as EGFR inhibitors. However, it may struggle with accurately predicting the effects of molecules acting through different pathways, such as KRAS inhibitors or calcium channel blockers, for which responses differ greatly from those previously observed.

multiple environments and can only be applied in contexts in which data are available for a sufficient amount of conditions. Ultimately, the right choice of interventional model is the one that yields the most accurate predictions for unseen perturbations, which can be assessed by reserving several conditions for the test set. This highlights the need for a diverse set of observed perturbations to accurately train the interventional component of the causal model.

Recent breakthroughs in experimental methodologies, such as Perturb-seq[32], have facilitated the generation of large single-cell CRISPR screen datasets[32–34], encompassing perturbations across tens to thousands of genes. However, in most cases, the vast majority of the available perturbations are associated with only a small number of cells, limiting the reliability of estimates of perturbational effects.

Additionally, single-cell data often contain technical noise and measurement errors entangled with the biological signals[23], leading to so-called batch effects. Disentangling these technical variations from biological signals is, therefore, crucial for causal models, often trained on diverse conditions from multiple batches. Extensive literature is

https://doi.org/10.1038/s41588-025-02124-2

BOX 2

Toward understanding the complexity of causal modeling in cellular biology

Evaluating the full complexity of causal modeling in single-cell genomics is challenging. However, we can provide some estimations based on the complexity of interactions of the proteome and the transcriptome to illustrate the challenge of the task.

Proteomic interactions

The human genome comprises approximately $N \approx 20{,}000$ protein-coding genes. As a first approximation, we can consider that there are similarly N unique proteins in the human proteome. With that simplifying assumption and considering only pairwise interactions, the total number of possible interactions is on the order of 10^8:

$$\text{Possible proteomic interactions} = \frac{N}{2} \approx 10^8.$$

Proteoform-specific interactions

The value above is an oversimplification and does not account for the various factors that increase the complexity of the human proteome, such as alternative splicing, post-translational modifications, translation errors and coding SNPs.

Recent studies suggest that the total number of unique mRNA transcripts generated by protein-coding genes, accounting for alternative splicing and coding SNPs, is approximately 150,000 (ref. 85). Of these, just around 90,000 are protein coding, while the noncoding isoforms possess modulation functions[86]. This provides a more realistic baseline for proteome diversity, even before considering post-translational modifications.

While the total number of proteoforms is still an open question, studies suggest that the total number of proteoforms in a given cell can be approximated. By making assumptions, such as only half of the human genome being expressed in each cell type, the total number of proteoforms in a given cell is estimated to be around $N \approx 1{,}000{,}000$ (ref. 87). Thus, the total number of possible protein–protein interactions increases to the order of 10^{11}:

$$\text{Possible proteoform-specific interactions} = \frac{N}{2} \approx 10^{11}.$$

To contextualize this result, current state-of-the-art large language models operate with a number of parameters that are also on a similar order of complexity[88]. Therefore, whole-cell models with a constant number of parameters per pairwise proteoform interaction are within reach of current computational capabilities.

Beyond pairwise interactions

Proteins do not only interact in a pairwise manner but also form multiprotein complexes. Using the same estimated number of proteoforms as above ($N \approx 1{,}000{,}000$), we estimate that the number of triple proteoform interactions is on the order of 10^{17}. It is worth noting that protein complexes often consist of more than three distinct proteins, increasing this value substantially.

Multiple types of interactions

The above calculation corresponds only to the estimated number of protein–protein interactions within a single cell. It does not encompass the complexity of noncoding mRNA transcripts or the RNA–DNA interactome, from which studies identify over 40 million contacts[88,89]. Cell–cell interactions should also be incorporated. As a point of reference, there are estimations of more than 100,000 ligand–receptor interactions[90].

Biological perturbations

Assuming CRISPR as the main experimental tool to validate causal interactions, there are approximately 10^4 unique single-gene knockouts. Multigene perturbations are most likely necessary to effectively learn a causal model, as biological mechanisms are known to often be redundant (that is, a similar function is encoded by more than one gene[91]). There are ~10^8 double-gene knockouts and ~10^{12} triple-gene knockouts.

Complexity can increase further considering that different parts of a gene's regulatory sequence can be targeted using different CRISPR guides, as happens in nature and is reflected in our genomic variability, provoking even more complex effects.

Takeaway

The complexity of causal modeling in biological cells is enormous, and we have not even considered temporal and spatial dynamics or cell–cell interactions. With such a complex system in which there are redundant mechanisms, learning and validating causal relationships is a technical and experimental challenge.

available for how to account for technical covariates and integrate different datasets[3,35–37], which can be leveraged to build training sets for causal models when required. However, these methods can potentially remove important biological signals. Experimental replicates, in which variability is expected to be purely technical, are vital for calibrating and validating these approaches. Some types of technical noise can be reduced by standardizing experimental protocols between laboratories, but others, such as uncontrolled genetic mutations in cell lines grown in different laboratories or the stochastic nature of cell differentiation, are more challenging to manage.

In summary, diverse, high-quality perturbational data are required when attempting to model biological perturbations in a realistic and physiologically relevant manner. Currently, the number of cells used per perturbation, as well as the number of available perturbations and technical noise have been limiting factors that hinder the effective development and training of generalizable causal models. It is also important to note that most experimental methods capture proxy measurements rather than direct observations of biological processes. For instance, RNA sequencing measures RNA abundance as a proxy for transcriptional activity but does not fully capture transcription rates or RNA processing. Similarly, ATAC-seq indicates chromatin accessibility, not active gene transcription. These limitations highlight challenges in bridging experimental observations with biological ground truth for causal modeling.

Machine learning-driven experimental design

One of the key applications of causal approaches is to propose testable hypotheses for experimental validation. The huge complexity of cell biology (Box 2) leads to an overwhelming number of potential experiments. Hence, it is essential to choose the most promising and

informative experiments to perform. Causal models can be used to suggest the design of future experiments through their ability to predict the outcome of yet untested conditions. Experimental results can then be incorporated into the available data to improve the model's predictions before new recommendations are generated again in an iterative fashion. This is known as machine learning-driven experimental design[38,39].

It is common to have access to only a limited set of previously acquired conditions. Models trained on such data may struggle to generalize to considerably different experimental conditions (Fig. 2d). Understanding when model predictions can be trusted is therefore extremely important. To this end, a common approach is to design models that provide, for a given input, a distribution of predictions instead of a single prediction. If this predictive distribution spreads across a wide range of values, uncertainty is considered to be high. Several approaches exist for estimating uncertainty[40–43]. They rely on inferring a distribution over the parameters of the models that are consistent with the data or rely on directly predicting the error made by the model. In causal modeling, this extends to obtain probabilities over graph structures[39,44,45].

The next step involves leveraging model predictions and uncertainty to guide the design of future experiments. The goal is to design strategies that can reduce uncertainty in a minimal number of experiments, a process known as active learning, or maximize some properties (for example, a phenotype, proportion of a given cell type), known as sequential model optimization or Bayesian optimization[46]. Recommendations for future experiments are made on the basis of their predicted outcome and informativeness derived from uncertainty estimation. These sequential approaches have been adapted to the context of causal models[47,48] to recommend informative interventions.

Sequential approaches have shown great promise in various areas of science, including molecular property prediction[49,50] and material design[51]. However, their application to cell-based assays, in which batch effects are relatively large[3], remains challenging. To the best of our knowledge, suggesting drug combinations is the only context in which sequential model optimization has been applied prospectively and quantitatively validated in human cells[52,53]. Adapting and scaling up these methods to, for instance, CRISPR knockout recommendation for GRN inference is an open challenge.

Learning interpretable models

Another great challenge of computational biology is to be able to derive biological insights from models. We consider a model, or part thereof, interpretable when the operations it performs can be associated with known processes (for example, transcriptomic regulation) and the values it computes can be associated with known and measurable biological quantities (for example, concentration of a specific molecule). Such an interpretable view can help biologists to extract meaningful insights from the model, propose new experiments and advance our understanding of cellular systems.

In practice, genome-scale models contain numerous interacting variables, making it difficult for a human to grasp as a whole, and it might be easier to analyze subparts of the model separately. For such an analysis to be meaningful, however, causal interactions need to be sparse and lead to reasonably independent clusters (corresponding to pathways or gene modules) that can be interpreted separately. Interestingly, interpretability often aligns with model faithfulness (that is, causal mechanisms accurately reflect biochemical interactions). Indeed, most entities within a cell directly interact with only a limited number of other entities, as for instance seen in the extreme sparsity of known protein–protein interaction networks[54], making faithful models usually easy to interpret. However, when broad influences occur (for example, environmental stress such as an increase in pressure or temperature), interactions may be more widespread, limiting interpretability.

Conversely, existing knowledge can be incorporated into an interpretable model to constrain the model's operations to known molecular interactions. However, incorporating prior knowledge can limit discovery potential. This is in sharp contrast with a non-interpretable model that would be provided with some representation of the prior knowledge as input. In addition, latent variables (Box 1), while useful for capturing unobserved biological factors, can introduce interpretability challenges if they lack direct biological correspondence.

Below, we discuss the opportunities and challenges associated with the incorporation of prior knowledge into interpretable models as well how introducing latent variables can impact the interpretability of causal models.

Incorporation of prior knowledge

Extensive research in cell biology has provided a vast amount of prior knowledge, including binding motifs and interaction databases, that can be incorporated into models. For instance, groups of genes known to function together, such as gene programs or biological pathways, are cataloged in databases[54–56]. By incorporating these known relationships into the causal learning framework, we can guide the model toward biologically plausible solutions. For example, if we know that a specific protein acts upstream of another in a signaling cascade, we can introduce this constraint into the model to make it more representative of known biology and so potentially improve its performance in novel scenarios. This represents a direct way of incorporating prior knowledge into causal models: a constraint or prior over edges of the causal graph. Prior knowledge incorporation may ease the task of complete causal graph discovery by providing a good preliminary estimate or by turning the task into a partial graph discovery problem[44].

Assessing the quality of available prior knowledge and figuring out the extent to which it can improve model performance is an ongoing challenge[57]. While high-quality prior knowledge can enhance causal models, incorporating flawed information can bias the model and impede its effectiveness. There is a need for methods that tackle potential biases in existing databases. For example, some proteins are well studied, while others are less so. Methods that can account for missing links between understudied proteins could help mitigate this bias. These predicted links could in turn be validated and used to refine existing databases based on empirical evidence.

Moreover, a more systematic reflection on how various types of prior knowledge should be incorporated is needed. For instance, protein–protein interaction networks should be incorporated differently than GRNs, as they represent distinct biological phenomena. Another challenge is the extraction and encoding of relevant metadata and information about the experimental protocol from the scientific publications that released the datasets.

A combination of using data-driven learning and prior biological knowledge holds substantial promise for building more robust and interpretable causal models of cellular processes.

Latent causal variables and their challenges for interpretability

Latent variables can represent any aspect of the generative process and are thus a broad concept. They are considered causal if they play a role similar to that of observed causal variables within the causal model: namely, they depend on and influence other causal variables (whether latent or observed) and the mechanisms among them are preserved across environments. They are expected to capture biological quantities involved in cellular mechanisms but are not directly observed. An example of noncausal latent variables is those designed to capture true gene expression levels: these variables are linked to observed gene counts via a negative binomial distribution[58], which accounts for technical dropout artifacts. Such artifacts reflect aspects of the data acquisition protocol rather than the underlying biological system. Such noncausal latent variables can help distinguish between technique-specific signatures and true biological properties.

https://doi.org/10.1038/s41588-025-02124-2

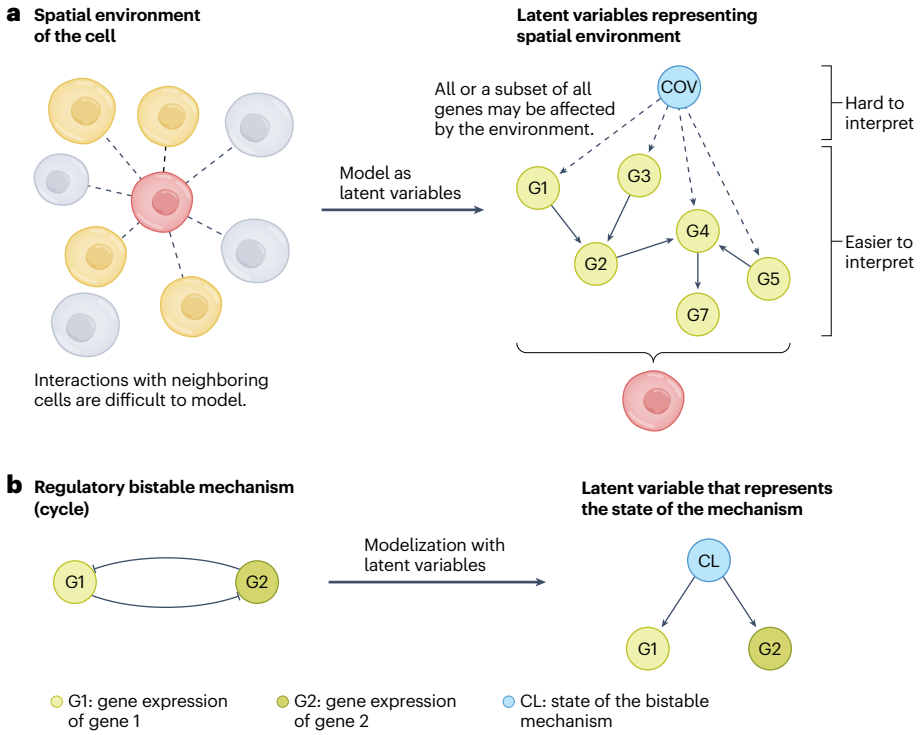

Fig. 3 | Latent causal variables help to model complex processes. a, Complex cellular processes that are difficult to model, such as the effect of the spatial environment, can be captured by latent variables and included in the causal graph as additional causal variables (shown on the right). However, these learned latent variables might be difficult to interpret. **b**, Including latent variables in the causal model may resolve algorithmic limitations such as the existence of cyclic processes in the causal graph. Here, genes 1 and 2 participate in a bistable mechanism and downregulate each other. This allows two stable states in which only one gene (either G1 or G2) is expressed but not the other. Instead of modeling the cycle structure directly, a latent variable can be included in the model, which accounts for the current state of the bistable mechanism.

Single-cell models often include latent variables to represent important but unobserved biological quantities (Box 2). They could, in principle, correspond to some attributes of known mechanisms, such as the activation of gene programs or pathways, or interactions between cells, along with features of their local microenvironment. Fig. 3a illustrates the specific case of incorporating spatial information into a causal model by adding latent variables that encode the characteristics of the extracellular environment of the cell and directly affect the expression levels of some genes.

These latent variables are typically inferred from a collection of low-level observations (that is, direct measures of biological quantities, typically gene expression levels), which can make them less subject to technical noise, as their value is not based on a single experimental measurement but usually derived from several measurements (the expression of several genes can serve for predicting the activation of a gene program). Latent variables can also provide a lower-dimensional representation of the cellular state, helping to avoid scalability issues. Furthermore, latent variables can solve other algorithmic limitations, such as the existence of cycles in the causal graph. For instance, in a bistable switch mechanism, models can infer a latent variable that is parent to both genes involved and represents the state of the switch. This approach removes the necessity to model direct interactions between the genes involved and therefore the cyclic regulation among them (shown in Fig. 3b).

Learning representations of single-cell data have been widely explored in single-cell genomes, mainly through matrix factorization[59] or disentanglement techniques[60,61]. These representations capture the major axes of variation in the data, helping to reveal key biological patterns. However, these representations are not necessarily causal, and interactions among latent variables may either be unmodeled or vary across environments. Causal representation learning[62] aims to discover latent causal variables from low-level observations. Current training approaches rely on interventional data to learn causal latent variables that change sparsely across conditions[62–65].

Latent variables can hinder model interpretability, as it often remains unclear which biological entities or quantities they represent, complicating the mapping of gene programs and known structures to these inferred variables. Without constraint, latent variables lack clear links with known processes or entities, making them hard to interpret. Interpretability can be enhanced through strategies that sparsify the dependencies between latent variables and known biological variables such as genes, so that each latent variable depends on only a small subset of genes. For instance, latent variables can be linked to known biological pathways on the basis of the set of genes they are most dependent on[66]. These strategies can either rely on hard constraints based on prior knowledge or soft constraints based on regularization during training. For example, one approach involves constraining perturbations to target a reduced set of latent variables[67,68]. As CRISPR knockouts or small molecules target specific biological processes, latent variables learned this way are more likely to correspond to one specific biological process, as opposed to a mixture of them (which would involve more genes) and therefore are more likely to be interpretable. This opens the way to the analysis of when specific pathways become activated in particular cell subtypes.

Learning causal kinetic models

The last challenge we focus on is modeling the temporal aspect of biological processes that occur over time. In many cases, such as in cell differentiation, development or disease progression, the temporal aspect cannot be ignored.

https://doi.org/10.1038/s41588-025-02124-2

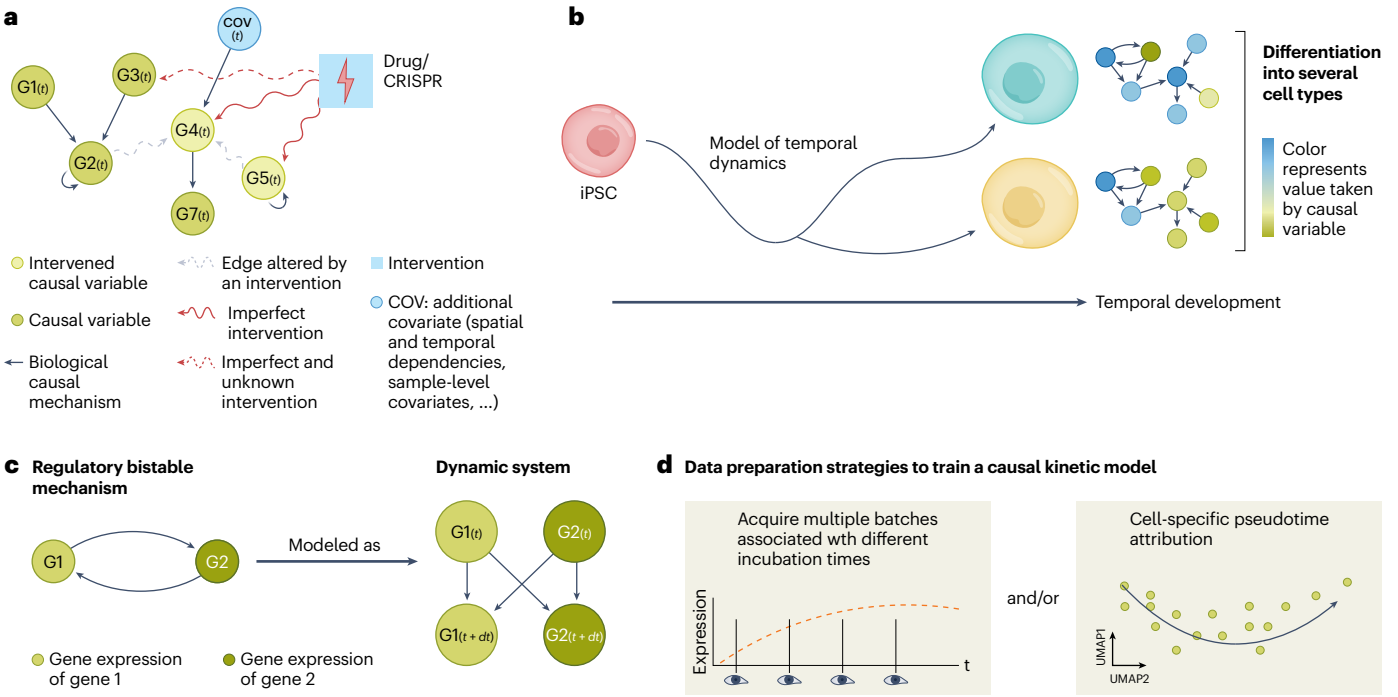

Fig. 4 | Modeling of a cell through causal kinetic models. a, Overview of a causal kinetic model for a cell. Here, variables represent expression levels, but they now explicitly depend on time. Their development through time is a function of their causal parents. Interventions may be applied in the same way as in simpler causal models and latent variables learned. However, the graph now contains cyclic regulatory motifs and autoregulation. **b**, Cellular development is a dynamic process that can be modeled by either ordinary or stochastic differential equations. Such models can account for branching phenomena, in which cells driven by similar mechanisms evolve toward distinct cell types that are associated with different expression profiles. iPSC, induced pluripotent stem cell. **c**, Causal kinetic models can account for cyclic regulatory motifs. In the regulatory mechanism (left), G1 and G2 downregulate each other, creating a feedback loop that forms a cycle. This cyclic structure poses algorithmic challenges for SCMs, which typically require acyclic graphs to define clear causal relationships. On the right, the system is represented as a time-dependent dynamical model, where the expression levels of G1 and G2 at time t influence their states at $t + dt$. This approach allows tracking of the temporal evolution of gene expression, bypassing the limitations imposed by cycles in static causal models. **d**, Data preparation strategies to train a causal kinetic model. Here, multiple acquisition times can be used or cell-specific pseudotimes can be attributed to obtain multiple series of expression profiles that are then used to construct training data of temporal dynamics. These temporal profiles enable the model to approximate causal relationships between gene expression over time, potentially uncovering the underlying kinetics of cellular processes.

In the causal models discussed so far, time is not taken into account, and relationships are drawn between static values of causal variables. By contrast, causal kinetic models, introduced by Peters et al.[69] (depicted in Fig. 4a), incorporate temporal information and account for dynamical properties of a system (Box 1). More precisely, these models assume that the rate of change of causal variables is governed by either ordinary differential equations (ODEs) or stochastic differential equations and depends on the expression of a small set of parent variables[69] (Box 1). For instance, such models can account for branching phenomena that occur during differentiation (illustrated in Fig. 4b). Interestingly, cyclic structures such as autoregulation do not pose any difficulty for such causal models, in which cells are framed as a dynamical system (shown in Fig. 4c).

The main difficulty in the application of causal kinetic models to single-cell genomics resides in the single-cell data reflecting only a snapshot as cells are destroyed before being measured and therefore observed at a single time point. One strategy to overcome this limitation is to use pseudotime inference methods[70–72] (Fig. 4d), which associate each cell with a different pseudotime, recapitulating its differentiation stage. Dynamical models have been applied to single-cell data by relying on pseudotime[73,74], in which the pseudotemporal information associated with each cell is used to construct multiple series of cells on which the dynamical model is trained. However, the performance of pseudotime analysis methods depends on the type of trajectory in the data[75], which is often unknown. While pseudotime analysis can be valuable, its limitations in complex datasets with unknown trajectories should be considered when interpreting the results.

Trends at the cell population level can be analyzed across experiments, but because single-cell sequencing is a destructive process, explicit matching between cells at different time points is impossible. Optimal transport methods[75–78] are being explored to match individual cells within populations across time points, helping to track cell state changes over time. Defining the necessary number of time points for reliable inference of kinetic causal dynamics is complex and should be determined empirically in simple contexts in which the ground truth mechanisms that drive the dynamics are known. However, the results likely depend on the context of the data and in particular on the complexity of the dynamics observed in the data. The development of single-cell simulators can help to explore the problem, as they can generate idealized datasets that mimic biological processes, helping researchers to understand and refine their models.

Causal kinetic models offer the promise to generate trajectories under different interventions, thereby opening the door to investigating the different developmental pathways that cells follow from their progenitor states to fully differentiated and functional cells. To make further progress in this direction, we believe that it is of crucial importance to have access to large temporal interventional datasets.

Conclusions

Causal questions lie at the core of biological research; however, causal machine learning is still in its infancy regarding its applications to

single-cell biology. Here, we have discussed the framework of causal machine learning that is classically applied to single-cell genomics and outlined three challenges in making accurate predictions under novel conditions, the interpretability of causal models and the inference of transcriptional dynamics.

A large data generation effort is needed to improve model training and to experimentally validate model predictions. In particular, we highlight the need for an increased availability of reliable interventional data (with large numbers of interventions and numbers of cells per intervention), temporal observations under intervention and an increased standardization of experimental protocols across batches and studies. Additionally, experimental replicates are crucial for addressing batch effects, enabling models to identify and learn the biological signal. Together, these improvements will allow the constitution of large and reliable interventional single-cell datasets that can serve as benchmark datasets to evaluate the generalization power of causal approaches across novel conditions. Notably, there are already ongoing efforts within the community to establish such benchmarks across a range of topics, from perturbation prediction[79,80] to experimental design for cell biology[81], which aim to provide standards for assessing model performance and fostering consistency and rigor in the field. Ultimately, the existence of consequential unknowns and unmeasured factors complicates efforts to faithfully capture biological mechanisms. Fortunately, continuous advances in experimental technologies for single-cell sequencing are anticipated to enhance data availability and quality, such as measuring multiple modalities simultaneously[82].

While data quality and availability are critical factors in advancing causal machine learning for single-cell genomics, they are not the only hurdles to overcome. One major issue is the lack of effective computational methods capable of scaling to the complexities of biological interactions. Many existing approaches operate with a limited number of variables and do not provide essential convergence guarantees[83,84]. Moreover, methodologies that combine causality with cross-modal data integration are lacking. Addressing varying spatial and temporal scales is also essential, as biological processes occur at different resolutions.

Causal machine learning for single-cell genomics offers the promise of providing a mechanistic view of cellular decision making. When causal variables are interpretable, such as genes and their amount of messenger RNA (mRNA) transcripts, models can yield biological insights that are then validated through targeted experiments, leading to new scientific knowledge. However, validating causal relationships in complex biological systems, such as those studied in single-cell genomics, can be particularly challenging due to the interplay of numerous factors (Box 2). In addition, validation experiments can be used to update and improve the causal model within an experimental design pipeline, which can then guide the design of the most informative experiments to perform. This strategy reduces the need for extensive experimentation and associated costs.

With the advent of single-cell atlases and increasing perturbation data, we expect causal models to become a crucial tool for informed experimental design and for deciphering the biological mechanisms that rule cellular decision making. Causal models hold the potential to help scientists navigate the vast complexity of biological systems by uncovering novel insights and accelerating the discovery of new therapeutic interventions with greater precision and efficacy.

References

1. McInnes, L., Healy, J., Saul, N. & Großberger, L. UMAP: Uniform Manifold Approximation and Projection. *J. Open Source Softw.* **3**, 861 (2018).
2. van der Maaten, L. & Hinton, G. Visualizing data using *t*-SNE. *J. Mach. Learn. Res.* **9**, 2579–2605 (2008).
3. Luecken, M. D. et al. Benchmarking atlas-level data integration in single-cell genomics. *Nat. Methods* **19**, 41–50 (2022).
4. Lange, M. et al. CellRank for directed single-cell fate mapping. *Nat. Methods* **19**, 159–170 (2022).
5. Hetzel, L. et al. Predicting cellular responses to novel drug perturbations at a single-cell resolution. In *Proceedings of the 36th International Conference on Neural Information Processing Systems* 26711–26722 (Curran Associates, 2022).
6. Liu, J. et al. Towards out-of-distribution generalization: a survey. Preprint at https://arxiv.org/abs/2108.13624 (2021).
7. Sekhon, J. The Neyman–Rubin model of causal inference and estimation via matching methods. In *The Oxford Handbook of Political Methodology* (eds Box-Steffensmeier, J. M. et al.) Ch. 11 (Oxford Academic, 2008).
8. Imbens, G. W. & Rubin, D. B. *Causal Inference in Statistics, Social, and Biomedical Sciences* (Cambridge University Press, 2015).
9. Segal, E. et al. Module networks: identifying regulatory modules and their condition-specific regulators from gene expression data. *Nat. Genet.* **34**, 166–176 (2003).
10. Qiao, L., Khalilimeybodi, A., Linden-Santangeli, N. J. & Rangamani, P. The evolution of systems biology and systems medicine: from mechanistic models to uncertainty quantification. *Annu. Rev. Biomed. Eng.* https://doi.org/10.1146/annurev-bioeng-102723-065309 (2025).
11. Wen, Y. et al. Applying causal discovery to single-cell analyses using CausalCell. *eLife* **12**, e81464 (2023).
12. Belyaeva, A., Squires, C. & Uhler, C. DCI: learning causal differences between gene regulatory networks. *Bioinformatics* **37**, 3067–3069 (2021).
13. Tam, G. H. F., Chang, C. & Hung, Y. S. Gene regulatory network discovery using pairwise Granger causality. *IET Syst. Biol.* **7**, 195–204 (2013).
14. Ke, N. R. et al. DiscoGen: learning to discover gene regulatory networks. Preprint at *bioRxiv* https://doi.org/10.1101/2023.04.11.536361 (2023).
15. Badia-I-Mompel, P. et al. Gene regulatory network inference in the era of single-cell multi-omics. *Nat. Rev. Genet.* **24**, 739–754 (2023).
16. Fleck, J. S. et al. Inferring and perturbing cell fate regulomes in human brain organoids. *Nature* **621**, 365–372 (2023).
17. Bravo González-Blas, C. et al. SCENIC+: single-cell multiomic inference of enhancers and gene regulatory networks. *Nat. Methods* **20**, 1355–1367 (2023).
18. Santos-Zavaleta, A. et al. RegulonDB v 10.5: tackling challenges to unify classic and high throughput knowledge of gene regulation in *E. coli* K-12. *Nucleic Acids Res.* **47**, D212–D220 (2019).
19. Peters, J., Janzing, D. & Scholkopf, B. *Elements of Causal Inference: Foundations and Learning Algorithms* (MIT Press, 2017).
20. Lopez, R., Hutter, J.-C., Pritchard, J. & Regev, A. Large-scale differentiable causal discovery of factor graphs. *Neural Inf. Process. Syst.* **abs/2206.07824**, 19290–19303 (2022).
21. Chevalley, M., Roohani, Y., Mehrjou, A., Leskovec, J. & Schwab, P. CausalBench: a large-scale benchmark for network inference from single-cell perturbation data. Preprint at https://arxiv.org/abs/2210.17283 (2022).
22. Wang, Y., Solus, L., Yang, K. D. & Uhler, C. Permutation-based causal inference algorithms with interventions. *Neural Inf. Process. Syst.* **30**, 5822–5831 (2017).
23. Aliee, H., Kapl, F., Hediyeh-Zadeh, S. & Theis, F. J. Conditionally invariant representation learning for disentangling cellular heterogeneity. Preprint at https://arxiv.org/abs/2307.00558 (2023).
24. Levine, M. & Davidson, E. H. Gene regulatory networks for development. *Proc. Natl Acad. Sci. USA* **102**, 4936–4942 (2005).
25. Lazar, N. H. et al. High-resolution genome-wide mapping of chromosome-arm-scale truncations induced by CRISPR–Cas9 editing. *Nat. Genet.* **56**, 1482–1493 (2024).

26. Hsu, P. D. et al. DNA targeting specificity of RNA-guided Cas9 nucleases. *Nat. Biotechnol.* **31**, 827–832 (2013).
27. Adikusuma, F. et al. Large deletions induced by Cas9 cleavage. *Nature* **560**, E8–E9 (2018).
28. Tsuchida, C. A. et al. Mitigation of chromosome loss in clinical CRISPR–Cas9-engineered T cells. *Cell* **186**, 4567–4582 (2023).
29. Papalexi, E. et al. Characterizing the molecular regulation of inhibitory immune checkpoints with multimodal single-cell screens. *Nat. Genet.* **53**, 322–331 (2021).
30. Bunne, C. et al. Learning single-cell perturbation responses using neural optimal transport. *Nat. Methods* **20**, 1759–1768 (2023).
31. Heumos, L. et al. Pertpy: an end-to-end framework for perturbation analysis. Preprint at *bioRxiv* https://doi.org/10.1101/2024.08.04.606516 (2024).
32. Dixit, A. et al. Perturb-seq: dissecting molecular circuits with scalable single-cell RNA profiling of pooled genetic screens. *Cell* **167**, 1853–1866 (2016).
33. Norman, T. M. et al. Exploring genetic interaction manifolds constructed from rich single-cell phenotypes. *Science* **365**, 786–793 (2019).
34. Replogle, J. M. et al. Mapping information-rich genotype–phenotype landscapes with genome-scale Perturb-seq. *Cell* **185**, 2559–2575 (2022).
35. Lotfollahi, M., Wolf, F. A. & Theis, F. J. scGen predicts single-cell perturbation responses. *Nat. Methods* **16**, 715–721 (2019).
36. Tran, H. T. N. et al. A benchmark of batch-effect correction methods for single-cell RNA sequencing data. *Genome Biol.* **21**, 12 (2020).
37. Tung, P.-Y. et al. Batch effects and the effective design of single-cell gene expression studies. *Sci. Rep.* **7**, 39921 (2017).
38. Rainforth, T., Foster, A., Ivanova, D. R. & Bickford Smith, F. Modern Bayesian experimental design. *Stat. Sci.* **39**, 100–114 (2024).
39. Jain, M. et al. GFlowNets for AI-driven scientific discovery. *Digit. Discov.* **2**, 557–577 (2023).
40. Williams, C. & Rasmussen, C. Gaussian processes for regression. In *Advances in Neural Information Processing Systems* (eds Touretzky, D. et al.) 514–520 (MIT Press, 1995).
41. Gal, Y. & Ghahramani, Z. Dropout as a Bayesian approximation: representing model uncertainty in deep learning. *ICML* **48**, 1050–1059 (2015).
42. Lakshminarayanan, B., Pritzel, A. & Blundell, C. Simple and scalable predictive uncertainty estimation using deep ensembles. In *Advances in Neural Information Processing Systems* **405**, 6402–6413 (2017).
43. Lahlou, S. et al. DEUP: direct epistemic uncertainty prediction. *Trans. Mach. Learn. Res.* (in the press).
44. Ke, N. R. et al. Learning neural causal models from unknown interventions. Preprint at https://arxiv.org/abs/1910.01075 (2019).
45. Deleu, T. et al. Bayesian structure learning with generative flow networks. In *Proceedings of the Thirty-Eighth Conference on Uncertainty in Artificial Intelligence* 518–528 (2022).
46. Močkus, J. On Bayesian methods for seeking the extremum. In *Optimization Techniques IFIP Technical Conference Novosibirsk* (ed. Marchuk, G. I.) 400–404 (Springer, 1975).
47. Toth, C. et al. Active Bayesian causal inference. *Adv. Neural Inf. Proc. Syst.* **35**, 16261–16275 (2022).
48. Scherrer, N. et al. Learning neural causal models with active interventions. Preprint at https://arxiv.org/abs/2109.02429 (2021).
49. Smith, J. S., Nebgen, B., Lubbers, N., Isayev, O. & Roitberg, A. E. Less is more: sampling chemical space with active learning. *J. Chem. Phys.* **148**, 241733 (2018).
50. Tran, K. et al. Computational catalyst discovery: active classification through myopic multiscale sampling. *J. Chem. Phys.* **154**, 124118 (2021).
51. Kim, S. et al. Deep learning for Bayesian optimization of scientific problems with high-dimensional structure. Preprint at https://arxiv.org/abs/2104.11667 (2021).
52. Bertin, P. et al. RECOVER identifies synergistic drug combinations in vitro through sequential model optimization. *Cell Rep. Methods* **3**, 100599 (2023).
53. Tosh, C. et al. A Bayesian active learning platform for scalable combination drug screens. *Nat. Commun.* **16**, 156 (2025).
54. Szklarczyk, D. et al. The STRING database in 2023: protein–protein association networks and functional enrichment analyses for any sequenced genome of interest. *Nucleic Acids Res.* **51**, D638–D646 (2022).
55. Türei, D., Korcsmáros, T. & Saez-Rodriguez, J. OmniPath: guidelines and gateway for literature-curated signaling pathway resources. *Nat. Methods* **13**, 966–967 (2016).
56. Lobentanzer, S. et al. Democratizing knowledge representation with BioCypher. *Nat. Biotechnol.* **41**, 1056–1059 (2023).
57. Bertin, P. et al. Analysis of gene interaction graphs as prior knowledge for machine learning models. Preprint at https://arxiv.org/abs/1905.02295 (2019).
58. Lopez, R., Regier, J., Cole, M. B., Jordan, M. I. & Yosef, N. Deep generative modeling for single-cell transcriptomics. *Nat. Methods* **15**, 1053–1058 (2018).
59. Stein-O'Brien, G. L. et al. Enter the matrix: factorization uncovers knowledge from omics. *Trends Genet.* **34**, 790–805 (2018).
60. Piran, Z., Cohen, N., Hoshen, Y. & Nitzan, M. Disentanglement of single-cell data with biolord. *Nat. Biotechnol.* **42**, 1678–1683 (2024).
61. Lotfollahi, M. et al. Predicting cellular responses to complex perturbations in high-throughput screens. *Mol. Syst. Biol.* **19**, e11517 (2023).
62. Schölkopf, B. et al. Toward causal representation learning. *Proc. IEEE* **109**, 612–634 (2021).
63. Ahuja, K., Mahajan, D., Wang, Y. & Bengio, Y. Interventional causal representation learning. *Proc. 40th Intl Conf. Mach. Learn.* **202**, 372–407 (2023).
64. Varici, B., Acarturk, E., Shanmugam, K., Kumar, A. & Tajer, A. Score-based causal representation learning with interventions. Preprint at https://arxiv.org/abs/2301.08230 (2023).
65. Michael, B. & Karaletsos, T. Modelling cellular perturbations with the sparse additive mechanism shift variational autoencoder. *Adv. Neural Inf. Proc. Syst.* **36**, 1–12 (2023).
66. Lotfollahi, M. et al. Biologically informed deep learning to query gene programs in single-cell atlases. *Nat. Cell Biol.* **25**, 337–350 (2023).
67. Lopez, R. et al. Learning causal representations of single cells via sparse mechanism shift modeling. *Proc. Mach. Learn. Res.* **213**, 1–30 (2023).
68. Kartik, A., Hartford, J. S. & Bengio, Y. Weakly supervised representation learning with sparse perturbations. *Adv. Neural Inf. Process. Syst.* **35**, 15516–15528 (2022).
69. Peters, J., Bauer, S. & Pfister, N. in *Causal Models for Dynamical Systems. Probabilistic and Causal Inference: The Works of Judea Pearl* 1st edn. 671–690 (Association for Computing Machinery, 2022).
70. Haghverdi, L., Büttner, M., Wolf, F. A., Buettner, F. & Theis, F. J. Diffusion pseudotime robustly reconstructs lineage branching. *Nat. Methods* **13**, 845–848 (2016).
71. Saelens, W., Cannoodt, R., Todorov, H. & Saeys, Y. A comparison of single-cell trajectory inference methods. *Nat. Biotechnol.* **37**, 547–554 (2019).
72. Moon, K. R. et al. Manifold learning-based methods for analyzing single-cell RNA-sequencing data. *Curr. Opin. Syst. Biol.* **7**, 36–46 (2018).

73. Aliee, H., Theis, F. J. & Kilbertus, N. Beyond predictions in neural ODEs: identification and interventions. Preprint at https://arxiv.org/abs/2106.12430 (2021).
74. Hananeh, A. et al. Sparsity in continuous-depth neural networks. *Adv. Neural Inf. Process. Syst.* **35**, 901–914 (2022).
75. Heumos, L. et al. Best practices for single-cell analysis across modalities. *Nat. Rev. Genet.* **24**, 550–572 (2023).
76. Tong, A. et al. Trajectorynet: A dynamic optimal transport network for modeling cellular dynamics. *Internatl Conf. Mach. Learn.* http://proceedings.mlr.press/v119/tong20a/tong20a-supp.pdf (PMLR, 2020).
77. Schiebinger, G. et al. Optimal-transport analysis of single-cell gene expression identifies developmental trajectories in reprogramming. *Cell* **176**, 928–943 (2019).
78. Eyring, L. V. et al. Modeling single-cell dynamics using unbalanced parameterized Monge maps. Preprint at *bioRxiv* https://doi.org/10.1101/2022.10.04.510766 (2022).
79. Wu, Y. et al. PerturBench: benchmarking machine learning models for cellular perturbation analysis. Preprint at https://arxiv.org/abs/2408.10609 (2024).
80. Csendes, G., Szalay, K. Z. & Szalai, B. Benchmarking a foundational cell model for post-perturbation RNAseq prediction. Preprint at *bioRxiv* https://doi.org/10.1101/2024.09.30.615843 (2024).
81. Mehrjou, A. et al. GeneDisco: a benchmark for experimental design in drug discovery. Preprint at https://arxiv.org/abs/2110.11875 (2021).
82. Metzner, E., Southard, K. M. & Norman, T. M. Multiome Perturb-seq unlocks scalable discovery of integrated perturbation effects on the transcriptome and epigenome. *Cell Syst.* **16**, 101161 (2025).
83. Sethuraman, M. G. et al. NODAGS-Flow: nonlinear cyclic causal structure learning. In *International Conference on Artificial Intelligence and Statistics* (eds Ruiz, F. et al.) 6371–6387 (PMLR, 2023).
84. Nguyen, T., Tong, A., Madan, K., Bengio, Y. & Liu, D. Causal inference in gene regulatory networks with GFlowNet: towards scalability in large systems. Preprint at https://arxiv.org/abs/2310.03579 (2023).
85. Tung, K.-F., Pan, C.-Y., Chen, C.-H. & Lin, W.-C. Top-ranked expressed gene transcripts of human protein-coding genes investigated with GTEx dataset. *Sci. Rep.* **10**, 16245 (2020).
86. Dhamija, S. & Menon, M. B. Non-coding transcript variants of protein-coding genes — what are they good for? *RNA Biol.* **15**, 1025–1031 (2018).
87. Aebersold, R. et al. How many human proteoforms are there? *Nat. Chem. Biol.* **14**, 206–214 (2018).
88. Dubey, A. et al. The Llama 3 herd of models. Preprint at https://arxiv.org/abs/2407.21783 (2024).
89. Gavrilov, A. A. et al. Studying RNA–DNA interactome by Red-C identifies noncoding RNAs associated with various chromatin types and reveals transcription dynamics. *Nucleic Acids Res.* **48**, 6699–6714 (2020).
90. Noh, J. Y. et al. CCIDB: a manually curated cell–cell interaction database with cell context information. *Database* **2023**, baad057 (2023).
91. Pearce, A. C. et al. Vav1 and Vav3 have critical but redundant roles in mediating platelet activation by collagen. *J. Biol. Chem.* **279**, 53955–53962 (2004).

Author contributions

A.T.-L. and P.B. conceptualized the work. P.B., A.T.-L. and H.A. wrote the original draft and created the figures. S.B. provided useful feedback. All authors critically reviewed and edited the manuscript. F.J.T., Y.B. and H.A. supervised the project.

Competing interests

Y.B. is an advisor to Recursion Pharmaceuticals. F.J.T. consults for Immunai, CytoReason, Cellarity, BioTuring and Genbio.AI and has an ownership interest in Dermagnostix and Cellarity. All other authors declare no competing interests.

Additional information

Correspondence should be addressed to Hananeh Aliee, Yoshua Bengio or Fabian J. Theis.

Peer review information *Nature Genetics* thanks Patrick Schwab and the other, anonymous, reviewer(s) for their contribution to the peer review of this work.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Springer Nature or its licensor (e.g. a society or other partner) holds exclusive rights to this article under a publishing agreement with the author(s) or other rightsholder(s); author self-archiving of the accepted manuscript version of this article is solely governed by the terms of such publishing agreement and applicable law.

© Springer Nature America, Inc. 2025