Perspective

Towards multimodal foundation models in molecular cell biology

https://doi.org/10.1038/s41586-025-08710-y

Received: 17 October 2023

Accepted: 29 January 2025

Published online: 16 April 2025

Check for updates

Haotian Cui[1,2,3], Alejandro Tejada-Lapuerta[4,5], Maria Brbić[6,7,8], Julio Saez-Rodriguez[9,10], Simona Cristea[11,12], Hani Goodarzi[13,14], Mohammad Lotfollahi[15,16], Fabian J. Theis[4,5,17] ✉ & Bo Wang[1,2,3,18] ✉

The rapid advent of high-throughput omics technologies has created an exponential growth in biological data, often outpacing our ability to derive molecular insights. Large-language models have shown a way out of this data deluge in natural language processing by integrating massive datasets into a joint model with manifold downstream use cases. Here we envision developing multimodal foundation models, pretrained on diverse omics datasets, including genomics, transcriptomics, epigenomics, proteomics, metabolomics and spatial profiling. These models are expected to exhibit unprecedented potential for characterizing the molecular states of cells across a broad continuum, thereby facilitating the creation of holistic maps of cells, genes and tissues. Context-specific transfer learning of the foundation models can empower diverse applications from novel cell-type recognition, biomarker discovery and gene regulation inference, to in silico perturbations. This new paradigm could launch an era of artificial intelligence-empowered analyses, one that promises to unravel the intricate complexities of molecular cell biology, to support experimental design and, more broadly, to profoundly extend our understanding of life sciences.

One central quest of molecular cell biology is to discover and represent the dynamic interactions and regulations among biological molecules: DNAs, RNAs, proteins and metabolites[1,2]. This comprehensive understanding will provide a foundation to capture, simulate and predict the dynamics of cell development and state changes. Efforts towards this quest have spanned decades and centred around the concepts of whole-cell modelling[3–5] or virtual cell[6–8]. Historically, these models were built as amalgams of rule-based submodules or ordinal differential equations (ODEs), in which each submodule was used to simulate one biological process[9]. For instance, the first whole-cell model is a system of 28 ODEs to capture the cellular processes of *Mycoplasma genitalium*[10]. However, these approaches are often limited by the oversimplification of dynamics and mathematical instability of ODEs[9]. As a result, existing virtual cell or whole-cell models are often limited to bacterial organisms and struggle to fully capture the complexity and scale of large non-linear interactions, especially in diverse contexts of tissues and cell states[4,11].

Recently, new opportunities have been enabled by the joint breakthroughs of analytical technologies (for example, next-generation sequencing, single-cell sequencing, cryo-electron microscopy and mass spectrometry-based proteomics; Fig. 1a), and the advancement of data-driven computational methods in large-scale machine learning: (1) for the past decade, advanced high-throughput sequencing technologies have yielded a profound reservoir of knowledge spanning the central dogma of molecular biology, encompassing DNA, RNA and their resulting protein products (Fig. 1b). The pace of biological data generation through genomics, transcriptomics, proteomics and other high-throughput technologies continues to accelerate exponentially[12]. This burgeoning wealth of data holds immense promise for elucidating molecular functions and characteristics in both normal and pathological states. Global consortia efforts, such as the Human Cell Atlas (HCA)[13], the Human Biomolecular Atlas Program (HuBMAP)[11] and the Human Tumor Atlas Network (HTAN)[14], have accumulated vast amounts of data spanning millions of cells across heterogeneous conditions and data modalities at an unprecedented rate. In addition, massively parallel multi-omic measurements have recently enabled measuring two[15–17] or even three different modalities in the same cells[18,19], necessitating the need of modelling across multimodal data[20]. (2) Driven by the recent breakthrough of pretraining large machine learning models, computational approaches are anticipated to ingest, analyse and interpret a

[1]Department of Computer Science, University of Toronto, Toronto, Ontario, Canada. [2]Vector Institute for Artificial Intelligence, Toronto, Ontario, Canada. [3]Peter Munk Cardiac Center, University Health Network, Toronto, Ontario, Canada. [4]Institute of Computational Biology, Helmholtz Center Munich, Munich, Germany. [5]School of Computing, Information and Technology, Technical University of Munich, Munich, Germany. [6]School of Computer and Communication Sciences, Swiss Federal Institute of Technology (EPFL), Lausanne, Switzerland. [7]School of Life Sciences, Swiss Federal Institute of Technology (EPFL), Lausanne, Switzerland. [8]Swiss Institute of Bioinformatics, Lausanne, Switzerland. [9]Institute for Computational Biomedicine, Heidelberg University, Faculty of Medicine, Heidelberg University Hospital, Heidelberg, Germany. [10]European Molecular Biology Laboratory, European Bioinformatics Institute (EMBL-EBI), Hinxton, UK. [11]Department of Data Science, Dana-Farber Cancer Institute, Boston, MA, USA. [12]Harvard T.H. Chan School of Public Health, Boston, MA, USA. [13]Arc Institute, Palo Alto, CA, USA. [14]Department of Biochemistry and Biophysics, University of California San Francisco, San Francisco, CA, USA. [15]Wellcome Sanger Institute, Cambridge, UK. [16]Cambridge Stem Cell Institute, University of Cambridge, Cambridge, UK. [17]TUM School of Life Sciences Weihenstephan, Technical University of Munich, Munich, Germany. [18]Department of Laboratory Medicine and Pathobiology, University of Toronto, Toronto, Ontario, Canada. ✉e-mail: fabian.theis@helmholtz-munich.de; bowang@vectorinstitute.ai

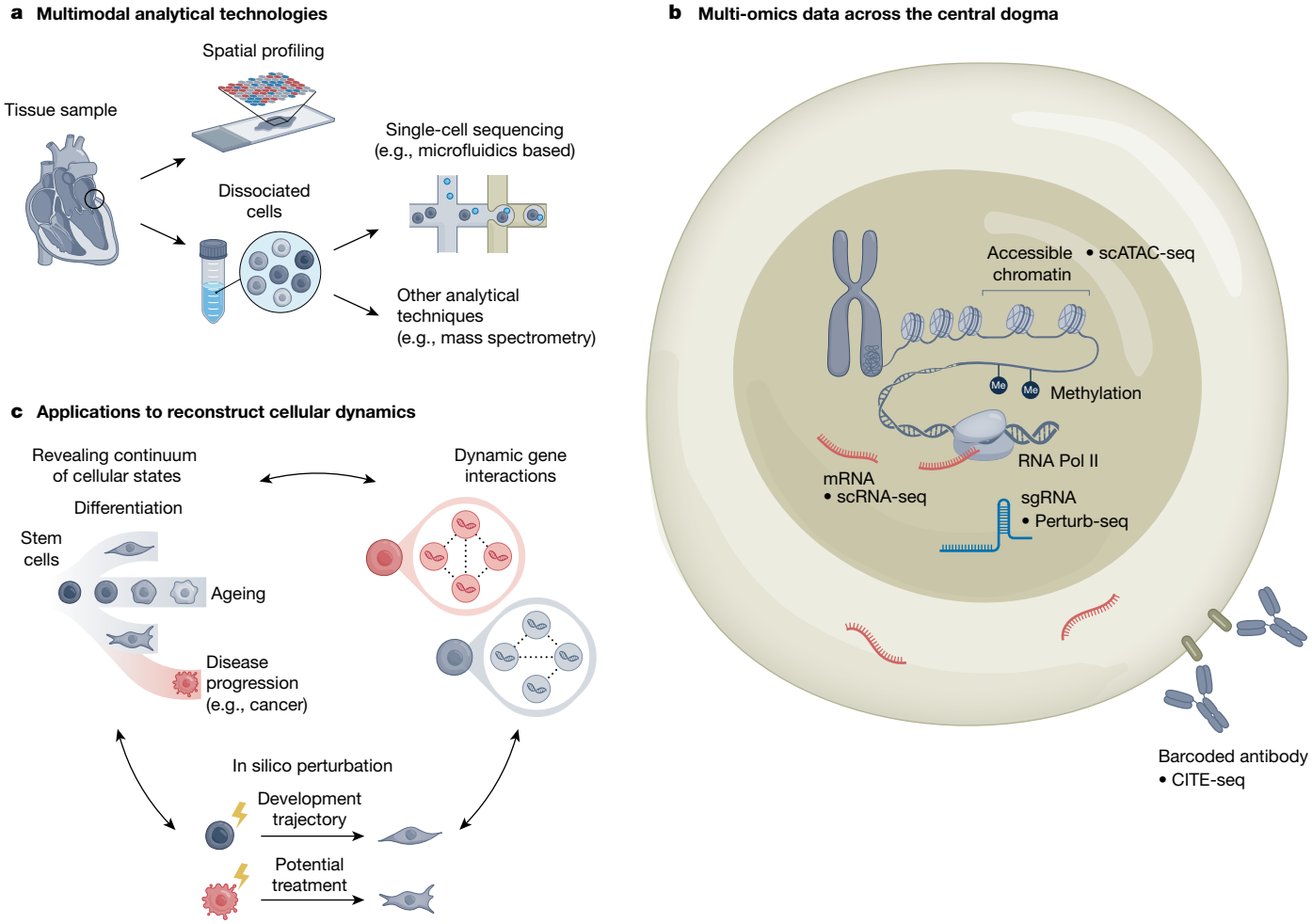


Fig. 1 | Multimodal analytical technologies and their applications. a, Various analytical technologies provide rich and diverse data at single-cell resolution and with spatial profiling. **b**, Data from analytical methods can reveal multiple steps across the central dogma. Inline texts list common sequencing methods for multi-omics profiling. For a complete list of currently available methods, we direct readers to recent reviews[57,125]. Pol II, polymerase II; scRNA-seq, single-cell RNA sequencing; sgRNA, single guide RNA. **c**, Opportunities of important potential applications to reconstruct cellular dynamics. The arrows indicate that the underlying mechanism of these applications is connected, and solving one task using MFMs can contribute to other tasks.

wide array of biological data types or 'modalities' and to evolve with the growing vast amount of data.

Therefore, we envision building multimodal foundation models (MFMs) as a promising new approach with the potential to address this challenge. Specifically, the main strategy is to train models on the extensive data in a self-supervised manner across modalities, thereby acquiring fundamental knowledge and capabilities, an approach epitomized by the concept of a foundation model[21]. The model should thus be able to accept different input data modalities and solve different tasks such as characterizing cell states and gene functions in healthy and disease conditions, and predicting the dynamics of these states (details in the section 'Opportunities of MFMs').

In the coming sections, we delve deeper into the structure and capabilities of MFMs. The section 'Overview of multimodal foundation models' expands on the idea of MFMs and their potential role in accelerating the 'wet lab in the loop', boosting the data generation and model building in feedback cycles. The section 'Opportunities of MFMs' explores the opportunities that these models present in areas such as tissue heterogeneity characterization, gene function prediction and in silico perturbation studies. The section 'Towards building MFMs for molecular cell biology' provides a description of the computational components and data requirements for building effective MFMs, and the section 'Challenges and limitations' sketches the challenges and limitations in their development and applications.

Overview of multimodal foundation models

The idea of foundation models

Foundation models are computational models of deep neural networks trained on expansive datasets with self-supervised learning methods, thus demonstrating strong capabilities for a wide array of downstream tasks via transfer learning[21]. In natural language processing, transformer-based[22] foundation models, such as GPT[23,24] and Llama[25–27] series, were trained on massive text corpora and can rapidly adapt to diverse downstream tasks via fine-tuning or in-context learning. Recently, the success of foundation models has also expanded to natural images[28,29] and videos[30], and gained cross-modal generation abilities between language and images[31]. In the context of molecular cell biology, foundation models offer a compelling approach to unifying our understanding of diverse biological processes. The key advantage of biological foundation models lies in their ability to learn and represent the complex, interconnected nature of cellular systems. By training on diverse omics data, these models can uncover subtle patterns and relationships that may not be apparent in isolated experiments or single-modality analyses, potentially revealing universal

Table 1 | Comparison between traditional machine learning models and MFMs for molecular cell biology

Key characteristics	Traditional specialist machine learning models	MFMs	Detailed in section
Pretraining	No	Yes, on data of varying contexts	–
Applicability	Limited to a subset of cell types and modalities; specialized in one task	Broad applicability across cell types and modalities; capable of handling diverse biological tasks such as cell state and gene function prediction, regulatory network reconstruction and in silico perturbation	Opportunities of MFMs
Learning paradigm	Typically supervised learning with labels provided by human experts	Self-supervised learning, contrastive learning and cross-modal objectives allow for learning from unlabelled datasets	Towards building MFMs for molecular cell biology
Overfitting risk	Prone to overfitting on small-scale collected datasets	Reduces overfitting risks by leveraging large-scale pretraining and generalizing well to smaller datasets via in-context learning	
Generalization to unseen data and contexts	Often struggles with generalization to new data or unseen cell types, requiring extensive retraining	Capable of in-context learning, allowing it to generalize well to unseen cells, benefitting from shared knowledge of extensive pretraining across multiple contexts	

biological principles that might be obscured in more narrowly focused studies (Table 1).

Expected characteristics and architecture

MFMs should readily incorporate diverse data types (such as bulk and single-cell sequencing) with multiple modalities, including transcriptomics, proteomics, metabolomics and epigenomics[12]. Specifically, the model should be pretrained in a unified self-supervised learning manner across modalities and then support varying biological analysis via transfer learning: first, by pretraining on massive aggregated data collections across many encompassing conditions, cell states and time points, the model is geared towards learning informative representations capturing nuanced properties of genes, transcripts, proteins, pathways and other biological processes[32,33]. Next, transfer learning techniques, including fine-tuning and in-context learning (Supplementary Note 1), then specialize these molecular embeddings to apply to diverse prediction tasks, enabling applications such as temporal cell-state mapping[34], novel cell-type characterization[35,36] and perturbation response prediction[37–39] (Fig. 1c).

The core computational architecture for building foundation models has been centred around transformers[22]. The transformer model, with its internal attention mechanism, shows remarkable abilities for modelling the semantics of word and image tokens (see Supplementary Note 1 for definitions of transformer, attention and tokens), and has become the de facto standard in the largest machine learning models. Accordingly, we expect the ability of the attention mechanism to recapitulate interactions of biomolecules, making transformers the backbone for the proposed biology MFMs as well. Several pioneering studies have corroborated the adoption of transformers in life science. Landmark studies, such as AlphaFold2 (ref. 40) and RoseTTA fold[41] for protein structure prediction, and ESM2 (ref. 42) and ESM3 (ref. 43) for novel protein generation, utilized the attention architecture to model protein structures and interactions of amino acids effectively. Enformer[44] used transformer architecture to predict gene expression and chromatin states from DNA sequences. Transformer architecture has been recently adopted in single-cell genomics by scGPT[45], GeneFormer[46] and scBERT[47] to pretrain single-cell RNA sequencing data and learn cell and gene representations. These studies verified the potential capability of modelling various molecular interactions in biological systems using the transformer architecture.

Data-centric workflow with lab-in-the-loop

The potential of MFMs poses a workflow shift unfolding in molecular cell biology. Historically, biology has been guided by hypothesis-driven approaches: recognizing patterns, generating hypotheses, designing experiments to challenge these hypotheses and refining theories based on the outcomes. Despite the long historical success of the hypothesis-driven approach, this approach is rather 'discipline' specific, for example, one studies cancer cells to understand cancer and studies cardiac cells to understand heart health (Fig. 2a, left). This implies a tacit assumption that knowledge in one context is rarely useful for another, which ignores the shared biochemical rules and molecular interactions among varying tissue and cell types.

Now, the advent of MFMs offers an opportunity to instead elevate a data-centric workflow, leveraging the pretraining process to capture and represent complex, non-linear biological rules in the vast model parameter space. Researchers start with large-scale, high-dimensional hypothesis-free data generation, followed by training a foundation model to integrate the data and extract underlying knowledge into biologically meaningful representations. Once the model can faithfully recapitulate the system (which can be verified by in silico reproducing experimental replicates), researchers can query the foundation model to extract valuable insights about the system and infer the underlying biological principles. This workflow is expected to enable accurate and fast modelling of biomolecular systems at unprecedented capacity and scale. This data-centric approach for molecular cell biology marks a departure from the prevalent hypothesis-driven workflow that seeks to derive conclusions solely from the study of specific contexts. Instead, the new approach operates on the premise that a shared foundational knowledge of biology exists, which can be leveraged across diverse contexts. Using the aforementioned example of cancer and cardiology studies, now the data-centric approach is intrinsically transdisciplinary; by training in large and diverse data at scale (for example, including cancerous and cardiac cells), the data-driven MFM workflow enables the acquisition of the foundational principles that rule cell behaviour (Fig. 2a).

In this new workflow, the foundation model facilitates a data-driven comprehension of biology through a process known as lab-in-the-loop[48–50], in which an experimental and a computational laboratory iterate together, integrating experiments and computational simulations to enhance the efficiency and accuracy of the experiments (Fig. 2b). Specifically, once a foundation model has been trained, it can be used to select an informative set of experiments to explore in the next round. For example, the model can predict drug efficacy on unseen cell lines, and then guide the forthcoming experiments to test on cell lines with high uncertainty (this will require the model to generate probabilistic output). The outcomes of these experiments are subsequently integrated into the training dataset of the model. Consequently, after sufficient iterations, the foundation model contains a simulator of molecular cell biology, offering invaluable guidance for the orchestration of targeted experiments[51–53].

Opportunities of MFMs

By assimilating diverse omics measurements, MFMs can develop holistic representations spanning the central dogma from genes to transcripts to proteins, elucidating the roles of specific genes, cell types and novel conditional gene interactions in dynamic contexts.

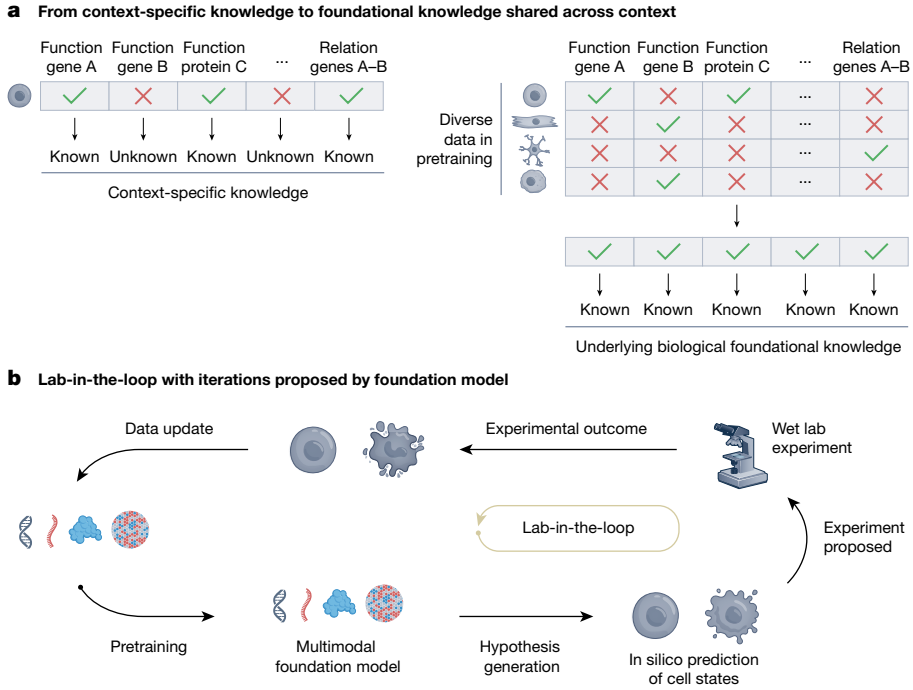


Fig. 2 | Diverse data context in pretraining and iterative improvement by lab-in-the-loop. a, MFMs are trained on biological data from a rich context. Diverse data across context-specific conditions can be recapitulated during pretraining, enriching the biological knowledge representation for both known and unknown conditions. The example scenario in the panel illustrated the idea of generalizing gene functions across diverse cell states, which helps to infer unseen functions in applications. **b**, Model–data–experiment, forming an active learning loop. This lab-in-the-loop yields iterative feedback to constantly update the MFM capability and the quality[57,125] generated biology hypotheses.

This section highlights important applications in which MFMs have particular potential.

Characterizing tissue heterogeneity

Recent advances in single-cell omics have enabled high-resolution dissection of cell subpopulations beyond classical surface markers, and researchers are actively developing techniques to unravel heterogeneity within complex tissues such as tumours. For instance, single-cell RNA sequencing has revealed transcriptional heterogeneity within glioblastomas associated with variable treatment response[54]. The epigenomic analysis further distinguished tumour subclones based on chromatin state indicative of different cells of origin[55]. Proteomic approaches have also parsed functional variability, with cytometry by time of flight identifying unique signalling states in cancer[56]. Integrating diverse measurements from the same cells can enable a more nuanced characterization of transitional states and lineages[57,58].

MFMs offer unique opportunities to define the continuous nature of cell states, in contrast to existing methods that have mainly focused on discrete definitions. The power of this modelling would be to infer the past, future and response of a cell to internal or external stimuli. By learning coordinated embeddings, we expect MFMs to truly excel in their abilities to contextualize, compare and complete cellular states. (1) To contextualize cellular states, MFMs are adept at embedding cells within an expansive continuum by assimilating diverse omics corpora during training. Reference mapping has been pioneered in studies of Seurat (v4)[58] and scArches[59], in which cell types and other meta information can be propagated from a rich context of cell atlases to new cells of interest. Now, in addition to discrete cell types, MFMs enable continuous cell-state description, which may recapitulate the position of a cell in the developmental tree or in disease progression. (2) To compare cellular states, MFMs facilitate swift and robust integration across heterogenous single-cell measurements and across omics modalities assayed separately in the same cells. This could allow joint analysis of heterogeneous datasets and comparison of cellular states across healthy and disease conditions. (3) To complete cellular states, given incomplete observations, MFMs can generate missing modalities to reconstruct full cell profiles in silico. For example, metabolic labelling of RNA and protein can be used to measure dynamics in experimental models[60–63], but it is not applicable in clinical samples. Now, by training on these experimental data and learning the cellular state dynamics, MFMs may help us to fill in missing modalities through their generative properties and predict the cell fates in clinical samples without metabolic labelling. This inherent capability helps to resolve traditional integration difficulties and utilize prior knowledge to tackle the multi-omics challenges of today.

Predicting gene functions and regulations

In biomarker discovery, learning unified patterns in massive heterogeneous disease datasets may reveal predictive multi-omics signatures involving specific gene modules, proteomic markers or metabolic profiles[12]. Recent works have demonstrated success in predicting gene functions from genome sequences alone[64,65] and using models learning from cell atlases of single-cell RNA sequencing data[46]. Furthermore, adding multi-omics contexts such as chromatin accessibility and methylation could improve inferences[58].

In addition to predicting gene functions, MFMs hold promise for reconstructing context-specific gene regulatory networks (GRNs). This promise is mainly driven by two observations: (1) the gene regulatory mechanism is inherently a process across multi-omics. Historically, GRNs have been predominantly compiled from experimentally validated regulatory events, as catalogued in various databases[66,67], or inferred from gene co-expression analyses in bulk transcriptomics data[68,69]. However, capturing complete regulatory mechanisms requires not only transcriptomics data but also other events along the central dogma, such as the DNA-binding events, alternative splicing of RNAs and post-translational protein modifications. Therefore, by

assimilating diverse omics data, multimodal models can potentially offer an integral and more accurate view[70]. For example, combining expression with chromatin accessibility can implicate influential regulators by incorporating *cis*-regulatory elements[71,72]. (2) The gene regulatory mechanism is inherently context specific. Transcription factor binding is known to be a highly dynamic process specific to tissues and conditions[70,73]. MFMs can address this challenge by uncovering conditional gene networks specific to cell types, developmental stages and disease states. MFMs can be expected to learn a default regulatory network during the large-scale pretraining with diverse contexts of multi-omics data, and the model can be flexibly adapted via transfer learning to elucidate specific GRNs by interpreting learned embeddings under different contexts. Thus, MFMs may fill key gaps in deciphering conditional GRNs for understanding the dynamical biological systems.

We also highlight the promising directions of incorporating prior knowledge from existing GRNs into MFMs, and the application of learned molecular regulations to recapitulate developmental and temporal cell states better than existing approaches (Supplementary Note 2). The potential computation mechanism for incorporating prior knowledge is further discussed in the section 'Desired computational components'.

In silico perturbation

MFMs trained on diverse omics data may predict the effects of hypothetical genetic or chemical perturbations on cell states. Recently, models such as scGPT[45], CellOracle[74], Geneformer[46], CellOT[75], CPA[38], chemcCPA[76] and GEARS[39] have shown initial success in perturbing learned cell embeddings to predict resulting expression profiles. Future developments can expand the applications beyond transcriptomics. By assimilating multi-omics measurements, MFMs can be more effective in perturbational modelling.

The in silico perturbation can be built on the abilities described in the aforementioned sections: MFMs can first construct complete cell representations by integrating gene expression, epigenetics and proteomics. Conditioning these embeddings on different cell types and perturbing states would allow nuanced perturbation analysis. Incorporating spatial and temporal datasets provides additional opportunities to trace impacts across tissues and time points. Models can then leverage learned pathway knowledge and gene regulatory networks to predict coordinated downstream effects of perturbations beyond just transcription. Particularly with the growing data[77,78] that combine single-cell sequencing and large-scale CRISPR perturbations, such as Perturb-seq[79], MFMs can be trained to predict post-perturbational response conditioned on the original cell profile and individual possible perturbation conditions. Of note, the combinatorial space of possible genetic perturbations is exponential: there exist $2^k - 1$ distinct combinations for knockout experiments of k genes. Therefore, accurate in silico predictions of perturbation responses could greatly accelerate the understanding of gene regulations and the discovery of new treatments.

Towards building MFMs for molecular cell biology

To fulfil the potential applications previously described, multimodal foundation models for molecular cell biology should possess certain key technical properties. We outline design and technical considerations to develop these capable foundation models.

Data for training MFMs

Pretraining versatile multimodal foundation models requires large and diverse multi-omics datasets spanning bulk sequencing, single-cell assays, spatial transcriptomics, chromatin accessibility and proteomics. Several valuable multi-omics data repositories exist, such as HuBMAP, ENCODE, the International Human Epigenome Consortium (IHEC) and the HCA[11,13,80–82]. However, current resources have limited paired measurements profiling the same cells or samples across modalities[83]. Paired data are generated by more recent sequencing protocols (such as 10X Multiome, single-cell cellular indexing of transcriptomes and epitopes by sequencing (CITE-seq), and single-cell assay for transposase-accessible chromatin (ATAC) with select antigen profiling by sequencing (ASAP-seq)) to capture different modalities simultaneously. For illuminating processes spanning the central dogma, such paired data would perform an essential role as anchor points when integrating other samples. Cross-species datasets may also provide helpful evolutionary context[84,85].

Single-cell sequencing data can particularly play a central part in training MFMs, due to the revealing of individual-level heterogeneity that would otherwise be missing in bulk experiments. Here we highlight the need for data generation and curation for future MFM training, and use the observation on single-cell data as an example. First, a growing trend of creating and sharing data has been observed in recent years. For instance, the number of cells in the CellxGENE[86] service (an online collection containing data from HuBMAP and HCA, among others) has tripled in the past year, from around 30 million to 93 million. Foundation models utilizing single-cell RNA sequencing data at the scale of tens of millions have already been developed[45,46], and we anticipate that the volume of publicly accessible data will continue to expand. However, the core challenge arises with data modalities beyond RNA sequencing. For instance, CELLxGENE currently hosts only about 200,000 cells of human single-cell ATAC-seq data (and 880,000 cells of mouse single-cell ATAC-seq data). Moving forwards, although single-cell RNA sequencing data may form the majority of the training data and provide backbone foundational knowledge, acquiring sufficient data that can encompass the tissue heterogeneity in each other modalities is equally critical. The generation of more comprehensive multimodal data will be instrumental in enriching the training datasets for MFMs.

Aggregating and curating data from a large number of studies is an equally important step. This includes straightforward efforts such as harmonizing meta labels across studies, and also non-trivial challenges such as unified quality control and normalization. Specifically, there is also the opportunity that MFMs also contribute to solving these issues themselves (Supplementary Note 3).

Desired computational components

Unified tokenization for multimodal data representation. Omics data bring about additional challenges in their diverse data types and different molecular resolutions from single nucleotides to whole proteins. To address this challenge, a potential solution can be inspired by the general machine learning research field, constructing unified tokens (Supplementary Note 1) across diverse data types. Representing the basic semantic unit of various data (such as words in natural language, pixel patches in images and nucleotides in DNA sequences) into token embeddings in a shared vector space has emerged as an inspiring way in recent unified large language models (LLMs) for computer visions and human languages[31,87]. Although tokenizing single modalities can be straightforward (for example, the same byte-pair-encoding[88] tokenization workflow has been used in OpenAI GPT series[89], as well as in biological sequence modelling such as DNABERT[64]), the greater potential comes from unifying the token representations across modalities. Specifically, this idea connects to the concept of 'early fusion'[90] (Supplementary Note 1), which emphasizes integrating the multimodal representations at the earliest stage of modelling (that is, the tokenization stage for transformer models). Molecular data are provided at vastly different resolutions, from single nucleotides (for example, the raw reads in next-generation sequencing) to whole proteins, and we envision tokenization techniques that can happen at multiple levels. For example, there should be low-level tokens as summarization of nucleotide k-mers, medium-level tokens that cover longer motifs and high-level tokens at the resolution of whole genes (Fig. 3a). Techniques such as subword tokenization can encode raw nucleotides

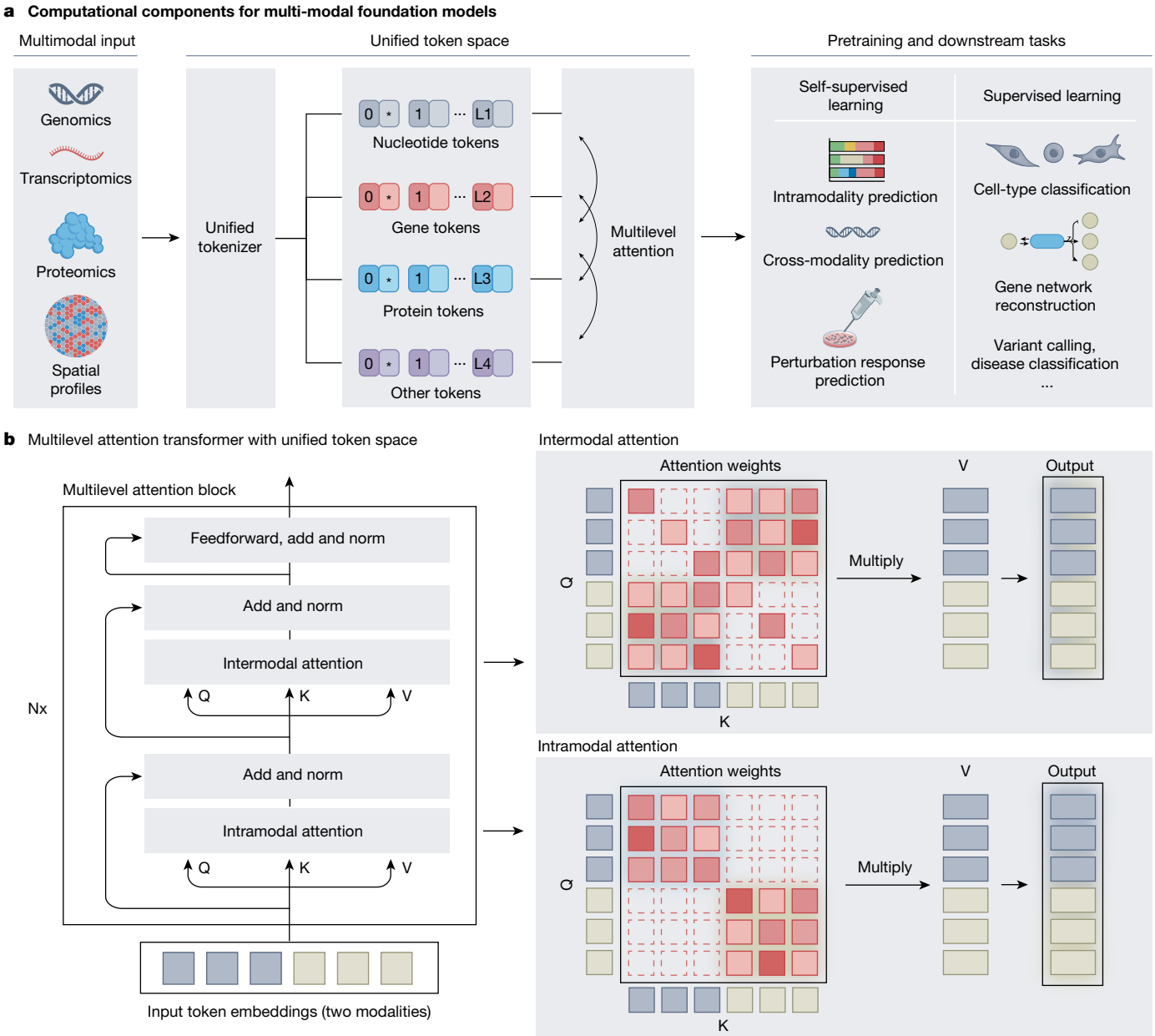
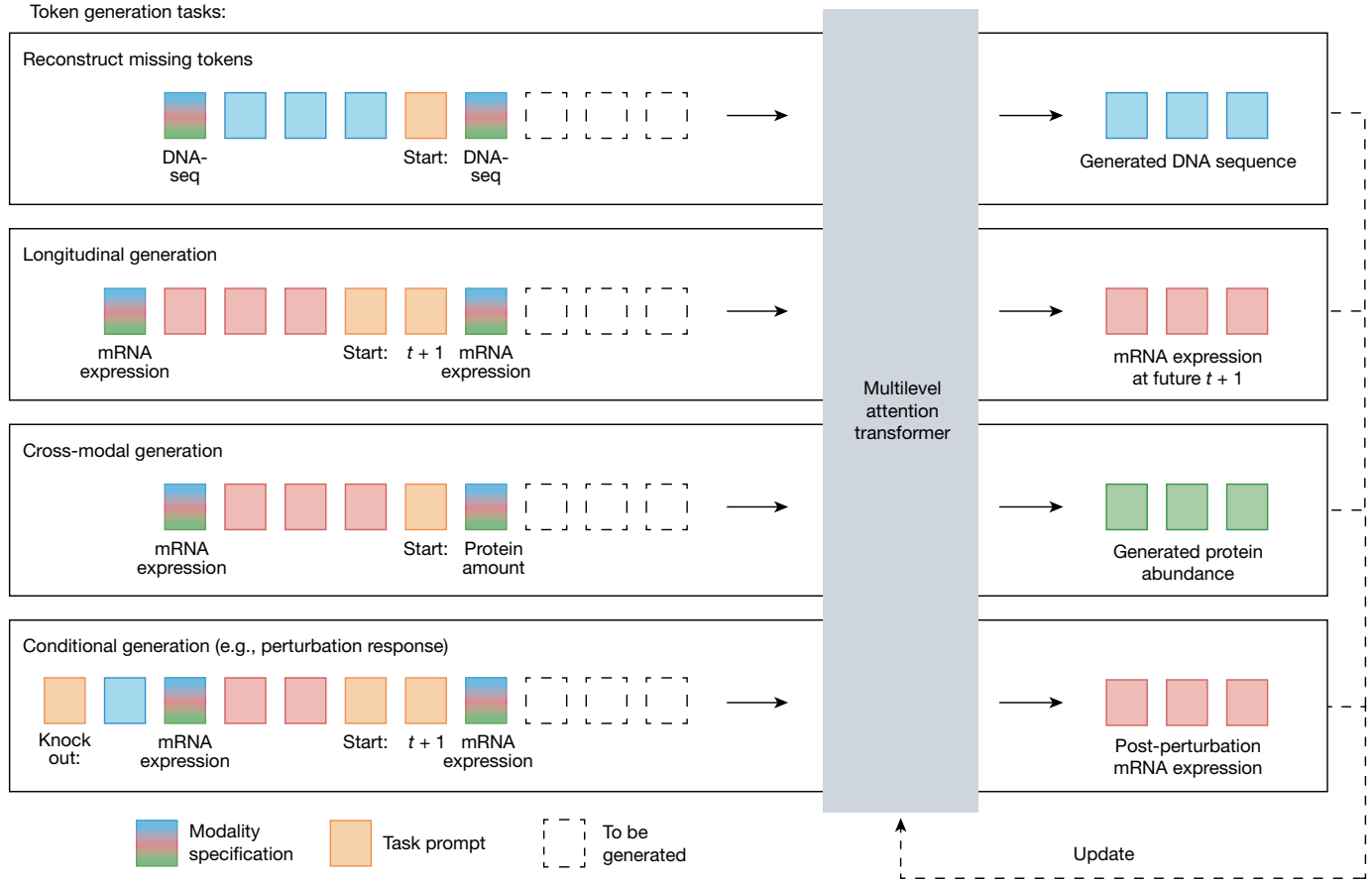


Fig. 3 | Computational components of multimodal foundation models. **a**, Desired components of MFMs. The model consists of multimodal input data, which is processed by hybrid unified tokens and multilevel attention operations. Various self-supervised and supervised learning objectives can be used to train the model in pretraining and transfer learning. **b**, Zoomed-in model of intramodal and intermodal attention mechanisms, showing variants of multihead attentions used in the model. The zoomed-in panels visualize intermodal and intramodal attention operations on a single head. The dense squares indicate the attention between the corresponding query (Q) and key (K) pairs, whereas the dashed squares indicate that no attention is computed for the specific queries and keys. Query, key and value (V) are real-number vectors computed in transformer models. Nx, *N* number of the attention blocks are stacked consecutively.

or amino acids into discrete vocabularies for modelling (this strategy has been used in the recent version of DNABERT-2 (ref. 65)), whereas higher-level tokens can represent genes or proteins[45,46].

Hybrid multilevel attention. As mentioned above, molecular data naturally exhibit structure at multiple scales, from individual base pairs to genes and pathways. To address this, hybrid transformer architectures with separate local (intramodal) and global (intermodal) self-attentions may effectively model interactions at each biologically relevant scale (Fig. 3b). Here intramodal attention stands for self-attention operations between tokens of the same level, such as the interactions that connect genes to genes, or nucleotides to nucleotides. Global attention refers to the inter-level operations that connect multilevel tokens, ideally generating an integral view of input data. Although multilevel attention has been applied in milestone studies in computer vision, such as SwinTransformer[91] and MultiScale ViT[92], similar ideas in biological foundation models are yet to be explored. Local attention mechanisms would understand relations within a specific modality, whereas the global attention mechanism would operate on a bigger scale, drawing connections between data modalities (gene–protein interactions, among others).

Intramodal and cross-modal training tasks with prompts. Models can be pretrained on unlabelled multi-omics data using objectives such as masked language modelling[93] and next token generation but applied to biology data[45,46]. Self-supervised learning tasks can again be



b Challenges

Data and computing resources

- Requirement of diverse and large volume of multi-omic data
- Parallel and accelerated computing resources
- Efforts to expand the accessibility of training and deploying foundation models

Open science and ethical considerations

- Biological foundation models should be publicly accessible
- Clearly conveying capabilities, limitations and use cases
- Safeguards for data privacy

Rigorous evaluations

- Diverse benchmarks on standardized datasets
- Evaluating abilities including predicting, generation, perturbation and other biological insight
- Open leaderboards and competitions

Interpretability and hallucination risks

- Interpreting large deep learning networks is challenging
- Predictions needed to be grounded by the training data and provided biological context
- Models should be able to admit uncertain outputs

Fig. 4 | Potential training tasks and challenges. a, Examples of training tasks for the pretraining of MFMs, including reconstructing missing tokens, longitudinal (temporal) generation, cross-modal and conditional generation. These tasks can all be framed in a unified way of token generation, with different meta tokens of modality specification and task prompting. **b**, Potential challenges in building MFMs for molecular cell biology.

categorized into intramodal and cross-modal types. The intramodal tasks optimize the modal to reconstruct unseen data, such as predicting randomly masked gene expression, imputing missing proteomic values or predicting post-perturbation responses from the naive cell state. In addition to intramodal self-supervised learning, we highlight two other promising cross-modal directions: (1) contrastive self-supervision is a promising pretraining approach that has been used in recent vision and language models[94–96], in which models are trained by maximizing the similarity difference between positive and negative pairs of input data. Analogously, MFMs can be trained by positive input pairs of different modalities' data of the same cell. (2) Multiple cross-modality prediction tasks can be included in the training, and the model can be guided by special task tokens when performing corresponding tasks. For example, to perform mRNA to protein prediction task, one can append task tokens '<mRNA>' to '<protein>' to the input data of mRNA sequencing profile and then train the model to output predictions of protein abundance. Moreover, this approach can extend to other tasks such as temporal predictions and perturbation response predictions.

In addition, all the training tasks mentioned above can be unified in the same token generation framework controlled by a few prompt tokens (Fig. 4a). By learning a few prompt tokens, such as modality

specification, condition specification (for example, <t + 1>, <knock out>) and meta control (for example, <start generation>), we can greatly expand the model capabilities and ensure the maximum reusing of model parameters between tasks (Fig. 4a). In addition, training objectives may not be limited to pure self-supervised learning. Informative meta information, such as age, gender and disease conditions, among others, is often paired with tissue samples. These can be readily used as supervised training signals, marking a unique characteristic of MFM training compared with LLMs in generic domains.

Integration of human knowledge. Integrating external knowledge such as pathways, gene ontologies, protein interaction networks and literature into pretraining can provide useful inductive biases for the otherwise purely data-driven models[97].

We highlight two possible technique directions, in particular, for structured and unstructured knowledge, respectively. (1) For structured knowledge integration[98], current biomedical databases represent structured interactions of biomolecules (for example, genes) in the form of knowledge graphs, and injecting these interaction priors into the attention mechanism in transformers can be a natural bridge connecting data-driven and human knowledge. For instance, given two genes that are annotated with related functions in databases (for example, Gene Ontology[99] and Reactome[100]), gene embeddings can be learned via graph-embedding methods[101,102]. Next, these knowledge graph-inspired embeddings can be used as the initialization of gene token embeddings in MFMs, potentially boosting the pretraining process. Of note, although this workflow exemplifies the combination of knowledge graphs and MFM training via gene token initialization, similar approaches can naturally extend to other tokens. (2) For unstructured knowledge integration, the raw text in biomedical literature contains vast unstructured knowledge. Recent retrieval-based chatbots have succeeded in industrial and clinical applications[103,104], in which existing unstructured text is represented as databases of vector embeddings with the help of current large-language models, such as recent BioGPT[105] and Med-PaLM[106]. These knowledge embeddings can be appended to the input of MFMs, enabling the joint training of MFMs with both experimental data and literature knowledge representations. Recent work using ProLlama[107] is an example of piloting such ideas, in which the authors introduced multitask training and instruction tuning for protein sequence data.

Challenges and limitations

Technical and regulatory challenges as well as limitations remain in the way towards broad use of molecular foundation models (Fig. 4b). Although these challenges for building MFMs for molecular cell biology share several similar topics with foundation models in generic domains, we found that the specific requests and potential solutions are often unique for this domain. We emphasize several considerations as follows.

Data and computing resources

Pretraining MFMs demands paired and aligned multi-omics datasets, ideally including spatial profiling and longitudinal samples. Although data of these kinds exist in global cell atlases, they are often lacking in sample numbers and scattered across studies. Therefore, global coordination across consortia would be essential for the development of data collection and versatile algorithms (Supplementary Note 4).

When building large-scale foundation models, great volumes of computing resources (for example, high-end GPUs) are usually utilized for training and deployment. This limits the accessibility of MFMs and also increases electricity consumption. To address the challenge, low-resource techniques would be important for building environmentally friendly artificial intelligence[108], and also greatly expand the accessibility of MFMs for users. Currently, it is inspiring to see that open-source low-resource techniques have already drawn attention in the broad machine learning domain, with the development of esteemed tools from low-rank adaptation (LORA)[109] to adapter-transformer[110]. These efforts may well be inherited to relieve the challenge of building biological MFMs.

Synthetic data hold potential as a complementary tool in training MFMs, especially in scenarios in which real data are scarce or incomplete (Supplementary Note 5). For instance, in molecular cell biology, data on paired modalities will specifically be needed to optimize the proposed cross-modal objectives. Such datasets are admittedly of limited volumes as mentioned in the section 'Characterizing tissue heterogeneity'. Synthetic data can help to fill these gaps, allowing for more comprehensive and effective model training.

Rigorous evaluation methodology

Extensive evaluations of model utility and suitability are crucial for genuine progress. Diverse benchmarks on standardized datasets assessing distinct capabilities will be needed. Examples of such may include predicting cell types and specific development dynamics, generating pseudo-samples for specific diseases, in silico perturbations and other abilities to provide biological insights. Of note, the tasks depicted in the section 'Opportunities of MFMs' can also be used to evaluate the essential abilities of MFMs. However, the evaluation metrics can be limited by the provided human annotation as ground truth. For example, cell-type annotations by human experts are currently widely based on marker genes or linear methods, potentially limiting the classifications of subtypes and rare cell types. When using these annotations as evaluations, models are favoured by high alignment (for example, by mutual information metric) between predicted cell clusters and human-labelled cell types[111]. This can penalize models when novel cell types or subclusters are recognized, and thus the evaluation metric can be exactly against the capability of MFMs for discovering new biological insights. Such a paradox can happen similarly when a model predicts new gene interactions or drug targets that were not originally in existing databases. This poses a unique challenge for biomolecular data analysis that human judgement and annotations can be unfaithful. Therefore, we anticipate the development of more 'objective', human-agnostic metrics to improve the evaluation process.

The assessment of MFMs needs to be conducted in a continuous and transparent manner. Open leaderboards and competitions on shared computing resources can enable rapid experimentation and innovation. Efforts of such have been pioneered in the OpenProblems (https://openproblems.bio) and DREAM[112] challenges, which have hosted numerous competitions and open datasets to accelerate methods development from community efforts[113,114]. We anticipate that these endeavours will continue to grow in various aspects, including generating standardized training and benchmarking datasets, developing trustworthy evaluation metrics, and particularly expanding the scope and scale of multi-omics data.

Interpretability and hallucination risks

Despite the promising opportunities, there are unsolved limitations of MFMs that may even fall short of traditional machine learning models or rule-based systems. In particular, we highlight interpretability and hallucination risks.

Interpreting large deep learning networks is challenging in general. For molecular cell biology, MFMs can generate gene expression profiles, predict DNA mutations, identify epigenetic signatures of new cells and predict new cell types. Explaining why the expression of a particular gene is upregulated or justifying the accuracy of a predicted gene–gene network can be complex. Recent advancements such as Kolmogorov–Arnold networks[115] show promise in extracting symbolic functions within gradient descent optimizations. These networks can be integrated with transformers to enhance the interpretability of

MFMs, providing clearer explanations for the predictions and decisions of the model.

Hallucination, a substantial challenge and potential limitation for MFMs, originally refers to the generation of plausible but factually incorrect or nonsensical outputs[116,117]. Although the hallucination of biological foundation models has not yet been formally defined, we propose the following factuality requirements for MFMs: (1) the output of the model should be grounded by the training data. (2) The output should be consistent with the context. For instance, if the model is prompted to generate a $CD4^+$ T cell, the generation should have such a gene profile signature. (3) When the model is not able to give accurate predictions or generations, it can admit so. Admittedly, satisfying these requirements can be challenging for MFM development, particularly the third requirement of self-identification. One potential direction to address hallucination involves implementing measures of uncertainty in model predictions[118–120]. By quantifying the uncertainty in model predictions, the model can be used to identify possible hallucinations and warn the uncertain cases.

Open science and ethical considerations

Pretrained models should be open and accessible with clear statements conveying capabilities, limitations and intended use cases. Transparency of foundation models is gaining surging importance. Recent efforts in natural language try to evaluate the accessibility and transparency of LLMs in multiple important dimensions[121], including data access, methods, usage policy, ethical risks and distribution fairness, among others. Similar assessment dimensions can also be valuable for biological foundation models. Overcoming these pressing challenges through collective efforts and research will be key to realizing the potential of MFMs.

Deploying multimodal models for biomedicine scenarios raises critical challenges. Models requiring large patient datasets warrant stringent safeguards for privacy, security and preventing unauthorized access or harm from leaked data. Great effort is essential to ensure that datasets are inclusive and representative across populations to avoid marginalizing groups and prevent skewed model performance. Predictions must be carefully validated on clinical cohorts before use in patient care. It has been reported in the existing large natural language models that great flexibility comes with a high probability of hallucination[117,122]. Similar concerns can happen for biological MFMs. For instance, when a doctor receives suggestions that recommend certain target therapies based on patient biopsy data, it is extremely important to guarantee the accuracy and interpret the rationale of the recommendation.

In addition, ensuring equitable access to models and data is crucial to fostering inclusivity in the field. We see the need for open-source and open-access infrastructures, and these will help to maintain a transparent and forward-looking perspective in the field (Supplementary Note 6).

A future of collective innovation

The development of MFMs to integrate diverse omics data promises to revolutionize molecular biology by uncovering insights at unprecedented scale and resolution. Achieving this potential requires a collaborative effort among biologists, data scientists, artificial intelligence researchers and ethicists to generate high-quality data, refine models and ensure accessibility. Looking forwards, the integration of MFMs into medicine could drive innovations in areas such as personalized treatment, disease modelling and drug discovery. This mirrors the transformative role that cell atlases, such as the HCA, already have in medical research[123,124]. In essence, the future of molecular discovery will be nurtured by a vibrant, collaborative ecosystem with a shared vision, empowering the scientific community to solve some of the most pressing challenges in biology and medicine.

1. Alberts, B. et al. *Molecular Biology of the Cell* 6th edn (W. W. Norton, 2020).
2. Keller, E. F. *Making Sense of Life: Explaining Biological Development with Models, Metaphors, and Machines* (Harvard Univ. Press, 2002).
3. Barabási, A.-L. & Oltvai, Z. N. Network biology: understanding the cell's functional organization. *Nat. Rev. Genet.* **5**, 101–113 (2004).
A seminal review on network biology, elucidating how molecular interactions shape cellular and organismal function.
4. Karlebach, G. & Shamir, R. Modelling and analysis of gene regulatory networks. *Nat. Rev. Mol. Cell Biol.* **9**, 770–780 (2008).
5. Goldberg, A. P. et al. Emerging whole-cell modeling principles and methods. *Curr. Opin. Biotechnol.* **51**, 97–102 (2018).
6. Johnson, G. T. et al. Building the next generation of virtual cells to understand cellular biology. *Biophys. J.* **122**, 3560–3569 (2023).
7. Karr, J. R., Takahashi, K. & Funahashi, A. The principles of whole-cell modeling. *Curr. Opin. Microbiol.* **27**, 18–24 (2015).
8. Freddolino, P. L. & Tavazoie, S. The dawn of virtual cell biology. *Cell* **150**, 248–250 (2012).
9. Georgouli, K., Yeom, J.-S., Blake, R. C. & Navid, A. Multi-scale models of whole cells: progress and challenges. *Front. Cell Dev. Biol.* **11**, 1260507 (2023).
10. Karr, J. R. et al. A whole-cell computational model predicts phenotype from genotype. *Cell* **150**, 389–401 (2012).
11. HuBMAP Consortium. The human body at cellular resolution: the NIH Human Biomolecular Atlas Program. *Nature* **574**, 187–192 (2019).
12. Hasin, Y., Seldin, M. & Lusis, A. Multi-omics approaches to disease. *Genome Biol.* **18**, 83 (2017).
The potential of multi-omics in uncovering molecular underpinnings of diseases and informing precision medicine.
13. Regev, A. et al. Science Forum: the Human Cell Atlas. *eLife* https://doi.org/10.7554/eLife.27041 (2017).
An introduction of the HCA initiative, a pivotal project for mapping cellular diversity across human tissues.
14. Rozenblatt-Rosen, O. et al. The Human Tumor Atlas Network: charting tumor transitions across space and time at single-cell resolution. *Cell* **181**, 236–249 (2020).
15. Stoeckius, M. et al. Simultaneous epitope and transcriptome measurement in single cells. *Nat. Methods* **14**, 865–868 (2017).
16. Ma, S. et al. Chromatin potential identified by shared single-cell profiling of RNA and chromatin. *Cell* **183**, 1103–1116.e20 (2020).
17. Deng, Y. et al. Spatial-CUT&Tag: spatially resolved chromatin modification profiling at the cellular level. *Science* **375**, 681–686 (2022).
18. Swanson, E. et al. Simultaneous trimodal single-cell measurement of transcripts, epitopes, and chromatin accessibility using TEA-seq. *eLife* **10**, e63632 (2021).
19. Mimitou, E. P. et al. Scalable, multimodal profiling of chromatin accessibility, gene expression and protein levels in single cells. *Nat. Biotechnol.* **39**, 1246–1258 (2021).
20. Moor, M. et al. Foundation models for generalist medical artificial intelligence. *Nature* **616**, 259–265 (2023).
21. Bommasani, R. et al. On the opportunities and risks of foundation models. Preprint at https://arxiv.org/abs/2108.07258 (2021).
An overview of the concept, opportunities and challenges of foundation models for diverse artificial intelligence applications.
22. Vaswani, A. et al. Attention is all you need. Preprint at https://arxiv.org/abs/1706.03762 (2017).
An introduction of the transformer architecture, the cornerstone of modern foundation models.
23. Brown, T. et al. Language models are few-shot learners. In *Proc. 34th International Conference on Neural Information Processing Systems* 1877–1901 (Curran Associates Inc., 2020).
An introduction of GPT-3, a 175-billion parameter language model demonstrating strong few-shot learning capabilities across diverse natural language processing tasks.
24. Ouyang, L. et al. Training language models to follow instructions with human feedback. In *Proc. 36th International Conference on Neural Information Processing Systems* 27730–27744 (Curran Associates Inc., 2022).
25. Touvron, H. et al. LLaMA: open and efficient foundation language models. Preprint at https://arxiv.org/abs/2302.13971 (2023).
An introduction to LLaMA, a suite of open-source language models (7B to 65B parameters) trained on publicly available data.
26. Touvron, H. et al. Llama 2: open foundation and fine-tuned chat models. Preprint at https://arxiv.org/abs/2307.09288 (2023).
27. llama3: The official meta Llama 3 GitHub site. *GitHub* https://github.com/meta-llama/llama3 (2024).
28. Rombach, R., Blattmann, A., Lorenz, D., Esser, P. & Ommer, B. High-resolution image synthesis with latent diffusion models. In *Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition* 10674–10685 (IEEE/CVF, 2022).
29. Podell, D. et al. SDXL: improving latent diffusion models for high-resolution image synthesis. Preprint at https://arxiv.org/abs/2307.01952 (2023).
30. Blattmann, A. et al. Stable video diffusion: scaling latent video diffusion models to large datasets. Preprint at https://arxiv.org/abs/2311.15127 (2023).
31. Liu, H., Li, C., Wu, Q. & Lee, Y. J. Visual instruction tuning. In *Proc. 37th International Conference on Neural Information Processing Systems* 34892–34916 (Curran Associates Inc., 2023).
32. Eraslan, G., Simon, L. M., Mircea, M., Mueller, N. S. & Theis, F. J. Single-cell RNA-seq denoising using a deep count autoencoder. *Nat. Commun.* **10**, 390 (2019).
33. Li, R., Li, L., Xu, Y. & Yang, J. Erratum to: Machine learning meets omics applications and perspectives. *Brief. Bioinform.* **23**, bbab560 (2022).
34. Klein, D. et al. Mapping cells through time and space with moscot. *Nature* **638**, 1065–1075 (2025).
35. Brbić, M. et al. MARS: discovering novel cell types across heterogeneous single-cell experiments. *Nat. Methods* **17**, 1200–1206 (2020).

36. Brbić, M. et al. Annotation of spatially resolved single-cell data with STELLAR. *Nat. Methods* **19**, 1411–1418 (2022).
37. Lotfollahi, M., Wolf, F. A. & Theis, F. J. scGen predicts single-cell perturbation responses. *Nat. Methods* **16**, 715–721 (2019).
38. Lotfollahi, M. et al. Predicting cellular responses to complex perturbations in high-throughput screens. *Mol. Syst. Biol.* **19**, e11517 (2023).
39. Roohani, Y., Huang, K. & Leskovec, J. Predicting transcriptional outcomes of novel multigene perturbations with GEARS. *Nat. Biotechnol.* **42**, 927–935 (2024).
A deep learning model integrating gene–gene relationship knowledge graphs to predict transcriptional responses.
40. Jumper, J. et al. Highly accurate protein structure prediction with AlphaFold. *Nature* **596**, 583–589 (2021).
An introduction to AlphaFold, a deep learning model achieving near-experimental accuracy in predicting protein structures.
41. Baek, M. et al. Accurate prediction of protein structures and interactions using a three-track neural network. *Science* **373**, 871–876 (2021).
42. Lin, Z. et al. Language models of protein sequences at the scale of evolution enable accurate structure prediction. Preprint at *bioRxiv* https://doi.org/10.1101/2022.07.20.500902 (2022).
43. ESM3: simulating 500 million years of evolution with a language model. *EvolutionaryScale* https://www.evolutionaryscale.ai/blog/esm3-release (2024).
A frontier language model for biology that simultaneously reasons over the sequence, structure and function of proteins.
44. Avsec, Ž. et al. Effective gene expression prediction from sequence by integrating long-range interactions. *Nat. Methods* **18**, 1196–1203 (2021).
45. Cui, H. et al. scGPT: towards building a foundation model for single-cell multi-omics using generative AI. *Nat. Methods* **21**, 1470–1480 (2024).
The development of scGPT, a generative pre-trained transformer model, leveraging over 33 million single-cell datasets to advance single-cell biology.
46. Theodoris, C. V. et al. Transfer learning enables predictions in network biology. *Nature* **618**, 616–624 (2023).
A large model pretrained on 30 million single-cell transcriptomes, facilitating accurate predictions in gene network biology.
47. Yang, F. et al. scBERT as a large-scale pretrained deep language model for cell type annotation of single-cell RNA-seq data. *Nat. Mach. Intell.* **4**, 852–866 (2022).
48. Wang, H. et al. Scientific discovery in the age of artificial intelligence. *Nature* **620**, 47–60 (2023).
49. Sverchkov, Y. & Craven, M. A review of active learning approaches to experimental design for uncovering biological networks. *PLoS Comput. Biol.* **13**, e1005466 (2017).
50. Szymanski, N. J. et al. An autonomous laboratory for the accelerated synthesis of novel materials. *Nature* **624**, 86–91 (2023).
51. Foster, A., Ivanova, D. R., Malik, I. & Rainforth, T. Deep adaptive design: amortizing sequential Bayesian experimental design. In *Proc. 38th International Conference on Machine Learning* Vol. 139 3384–3395 (PMLR, 2021).
52. Rainforth, T., Foster, A., Ivanova, D. R. & Smith, F. B. Modern Bayesian experimental design. *Statist. Sci.* **39**, 100–114 (2024).
53. Vanlier, J., Tiemann, C. A., Hilbers, P. A. J. & van Riel, N. A. W. A Bayesian approach to targeted experiment design. *Bioinformatics* **28**, 1136–1142 (2012).
54. Patel, A. P. et al. Single-cell RNA-seq highlights intratumoral heterogeneity in primary glioblastoma. *Science* **344**, 1396–1401 (2014).
55. Eyler, C. E. et al. Single-cell lineage analysis reveals genetic and epigenetic interplay in glioblastoma drug resistance. *Genome Biol.* **21**, 174 (2020).
56. Chevrier, S. et al. An immune atlas of clear cell renal cell carcinoma. *Cell* **169**, 736–749.e18 (2017).
57. Zhu, C., Preissl, S. & Ren, B. Single-cell multimodal omics: the power of many. *Nat. Methods* **17**, 11–14 (2020).
58. Hao, Y. et al. Integrated analysis of multimodal single-cell data. *Cell* **184**, 3573–3587.e29 (2021).
59. Lotfollahi, M. et al. Mapping single-cell data to reference atlases by transfer learning. *Nat. Biotechnol.* **40**, 121–130 (2022).
60. Battich, N. et al. Sequencing metabolically labeled transcripts in single cells reveals mRNA turnover strategies. *Science* **367**, 1151–1156 (2020).
61. Cao, J., Zhou, W., Steemers, F., Trapnell, C. & Shendure, J. Sci-fate characterizes the dynamics of gene expression in single cells. *Nat. Biotechnol.* **38**, 980–988 (2020).
62. Qiu, Q. et al. Massively parallel and time-resolved RNA sequencing in single cells with scNT-seq. *Nat. Methods* **17**, 991–1001 (2020).
63. Qiu, X. et al. Mapping transcriptomic vector fields of single cells. *Cell* **185**, 690–711.e45 (2022).
64. Ji, Y., Zhou, Z., Liu, H. & Davuluri, R. V. DNABERT: pre-trained bidirectional encoder representations from transformers model for DNA-language in genome. *Bioinformatics* **37**, 2112–2120 (2021).
65. Zhou, Z. et al. DNABERT-2: efficient foundation model and benchmark for multi-species genome. Preprint at https://arxiv.org/abs/2306.15006 (2023).
66. Han, H. et al. TRRUST v2: an expanded reference database of human and mouse transcriptional regulatory interactions. *Nucleic Acids Res.* **46**, D380–D386 (2018).
67. Liu, Z.-P., Wu, C., Miao, H. & Wu, H. RegNetwork: an integrated database of transcriptional and post-transcriptional regulatory networks in human and mouse. *Database* **2015**, bav095 (2015).
68. Margolin, A. A. et al. ARACNE: an algorithm for the reconstruction of gene regulatory networks in a mammalian cellular context. *BMC Bioinformatics* **7**, S7 (2006).
69. Langfelder, P. & Horvath, S. WGCNA: an R package for weighted correlation network analysis. *BMC Bioinformatics* **9**, 559 (2008).
70. 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).
71. Buenrostro, J. D., Giresi, P. G., Zaba, L. C., Chang, H. Y. & Greenleaf, W. J. Transposition of native chromatin for fast and sensitive epigenomic profiling of open chromatin, DNA-binding proteins and nucleosome position. *Nat. Methods* **10**, 1213–1218 (2013).
72. Qin, Q. et al. Lisa: inferring transcriptional regulators through integrative modeling of public chromatin accessibility and ChIP-seq data. *Genome Biol.* **21**, 32 (2020).
73. Kim, S. & Wysocka, J. Deciphering the multi-scale, quantitative *cis*-regulatory code. *Mol. Cell* **83**, 373–392 (2023).
74. Kamimoto, K. et al. Dissecting cell identity via network inference and in silico gene perturbation. *Nature* **614**, 742–751 (2023).
75. Bunne, C. et al. Learning single-cell perturbation responses using neural optimal transport. *Nat. Methods* **20**, 1759–1768 (2023).
76. Hetzel, L. et al. Predicting cellular responses to novel drug perturbations at a single-cell resolution. In *Proc. 36th International Conference on Neural Information Processing Systems* 26711–26722 (Curran Associates Inc., 2022).
77. Joung, J. et al. A transcription factor atlas of directed differentiation. *Cell* **186**, 209–229.e26 (2023).
78. Replogle, J. M. et al. Mapping information-rich genotype–phenotype landscapes with genome-scale Perturb-seq. *Cell* **185**, 2559–2575.e28 (2022).
79. Dixit, A. et al. Perturb-seq: dissecting molecular circuits with scalable single-cell RNA profiling of pooled genetic screens. *Cell* **167**, 1853–1866.e17 (2016).
80. Rozenblatt-Rosen, O., Stubbington, M. J. T., Regev, A. & Teichmann, S. A. The Human Cell Atlas: from vision to reality. *Nature* **550**, 451–453 (2017).
81. Luo, Y. et al. New developments on the Encyclopedia of DNA Elements (ENCODE) data portal. *Nucleic Acids Res.* **48**, D882–D889 (2020).
82. Stunnenberg, H. G., International Human Epigenome Consortium & Hirst, M. The International Human Epigenome Consortium: a blueprint for scientific collaboration and discovery. *Cell* **167**, 1897 (2016).
83. Butler, A., Hoffman, P., Smibert, P., Papalexi, E. & Satija, R. Integrating single-cell transcriptomic data across different conditions, technologies, and species. *Nat. Biotechnol.* **36**, 411–420 (2018).
84. Tabula Muris Consortium. et al. Single-cell transcriptomics of 20 mouse organs creates a Tabula Muris. *Nature* **562**, 367–372 (2018).
85. Yao, Z. et al. A transcriptomic and epigenomic cell atlas of the mouse primary motor cortex. *Nature* **598**, 103–110 (2021).
86. CZI Single-Cell Biology Program et al. CZ CELL×GENE Discover: a single-cell data platform for scalable exploration, analysis and modeling of aggregated data. *Nucleic Acids Res.* **53**, D886–D900 (2025).
87. Chameleon Team. Chameleon: mixed-modal early-fusion foundation models. Preprint at https://arxiv.org/abs/2405.09818 (2024).
88. Gage, P. A new algorithm for data compression. *C Users J. Arch.* **12**, 23–38 (1994).
89. OpenAI et al. GPT-4 technical report. Preprint at https://arxiv.org/abs/2303.08774 (2023).
90. Barnum, G., Talukder, S. & Yue, Y. On the benefits of early fusion in multimodal representation learning. Preprint at https://arxiv.org/abs/2011.07191 (2020).
An investigation into early-fusion strategies in multimodal learning, demonstrating that immediate integration of inputs enhances model performance and robustness.
91. Liu, Z. et al. Swin Transformer: hierarchical vision transformer using Shifted Windows. In *Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition* 9992–10002 (IEEE/CVF, 2021).
92. Fan, H. et al. Multiscale vision transformers. In *Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition* 6804–6815 (IEEE/CVF, 2021).
93. Devlin, J., Chang, M.-W., Lee, K. & Toutanova, K. BERT: pre-training of deep bidirectional transformers for language understanding. In *Proc. 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies* 4171–4186 (Association for Computational Linguistics, 2019).
94. Grill, J.-B. et al. Bootstrap your own latent—a new approach to self-supervised learning. In *Proc. 34th International Conference on Neural Information Processing Systems* 21271–21284 (Curran Associates Inc., 2020).
95. Chen, T., Kornblith, S., Norouzi, M. & Hinton, G. A simple framework for contrastive learning of visual representations. In *Proc. 37th International Conference on Machine Learning* Vol. 119 (eds. Iii, H. D. & Singh, A.) 1597–1607 (PMLR, 2020).
96. Radford, A. et al. Learning transferable visual models from natural language supervision. In *Proc. 38th International Conference on Machine Learning* Vol. 139 8748–8763 (PMLR, 2021).
97. AlQuraishi, M. & Sorger, P. K. Differentiable biology: using deep learning for biophysics-based and data-driven modeling of molecular mechanisms. *Nat. Methods* **18**, 1169–1180 (2021).
98. Pan, S. et al. Unifying large language models and knowledge graphs: a roadmap. *IEEE Trans. Knowl. Data Eng.* **36**, 3580–3599 (2024).
99. Harris, M. A. et al. The Gene Ontology (GO) database and informatics resource. *Nucleic Acids Res.* **32**, D258–D261 (2004).
100. Fabregat, A. et al. The Reactome Pathway Knowledgebase. *Nucleic Acids Res.* **46**, D649–D655 (2018).
101. Grover, A. & Leskovec, J. node2vec: Scalable feature learning for networks. In *Proc. 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* 855–864 (Association for Computing Machinery, 2016).
102. Hamilton, W. L., Ying, R. & Leskovec, J. Inductive representation learning on large graphs. In *Proc. 31st International Conference on Neural Information Processing Systems* 1–19 (Curran Associates Inc., 2017).
103. Zhao, W. X., Liu, J., Ren, R. & Wen, J.-R. Dense text retrieval based on pretrained language models: a survey. *ACM Trans. Inf. Syst. Secur.* **42**, 1–60 (2024).
104. Jeong, J. et al. Multimodal image-text matching improves retrieval-based chest X-ray report generation. In *Proc. Machine Learning Research. Medical Imaging with Deep Learning* Vol. 227 (eds Oguz, I. et al.) 978–990 (PMLR, 2024).
105. Luo, R. et al. BioGPT: generative pre-trained transformer for biomedical text generation and mining. *Brief. Bioinform.* **23**, bbac409 (2022).
106. Singhal, K. et al. Large language models encode clinical knowledge. *Nature* **620**, 172–180 (2023).
107. Lv, L. et al. ProLLaMA: a protein large language model for multi-task protein language processing. Preprint at https://arxiv.org/abs/2402.16445 (2024).

108. Debus, C., Piraud, M., Streit, A., Theis, F. & Götz, M. Reporting electricity consumption is essential for sustainable AI. *Nat. Mach. Intell.* **5**, 1176–1178 (2023).
109. Hu, E. J. et al. LoRA: low-rank adaptation of large language models. Preprint at https://arxiv.org/abs/2106.09685 (2021).
110. Pfeiffer, J. et al. AdapterHub: a framework for adapting transformers. In *Proc. 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations* 46–54 (Association for Computational Linguistics, 2020).
111. Luecken, M. D. et al. Benchmarking atlas-level data integration in single-cell genomics. *Nat. Methods* **19**, 41–50 (2022).
112. Meyer, P. & Saez-Rodriguez, J. Advances in systems biology modeling: 10 years of crowdsourcing DREAM challenges. *Cell Syst.* **12**, 636–653 (2021).
113. Saez-Rodriguez, J. et al. Crowdsourcing biomedical research: leveraging communities as innovation engines. *Nat. Rev. Genet.* **17**, 470–486 (2016).
114. Lance, C. et al. Multimodal single cell data integration challenge: results and lessons learned. In *Proc. NeurIPS 2021 Competitions and Demonstrations Track* Vol. 176 (eds Kiela, D., Ciccone, M. & Caputo, B.) 162–176 (PMLR, 2022).
115. Liu, Z. et al. KAN: Kolmogorov–Arnold networks. Preprint at https://arxiv.org/abs/2404.19756 (2024).
116. Maynez, J., Narayan, S., Bohnet, B. & McDonald, R. On faithfulness and factuality in abstractive summarization. In *Proc. 58th Annual Meeting of the Association for Computational Linguistics* 1906–1919 (Association for Computational Linguistics, 2020).
117. Ji, Z. et al. Survey of hallucination in natural language generation. *ACM Comput. Surv.* **55**, 248 (2022).
118. Manakul, P., Liusie, A. & Gales, M. J. F. SelfCheckGPT: zero-resource black-box hallucination detection for generative large language models. In *Proc. 2023 Conference on Empirical Methods in Natural Language Processing* 9004–9017 (Association for Computational Linguistics, 2023).
119. Yin, Z. et al. Do large language models know what they don't know? In *Proc. Findings of the Association for Computational Linguistics: ACL 2023* 8653–8665 (Association for Computational Linguistics, 2023).
120. Tian, K., Mitchell, E., Yao, H., Manning, C. D. & Finn, C. Fine-tuning language models for factuality. Preprint at https://arxiv.org/abs/2311.08401 (2023).
121. Bommasani, R. et al. The foundation model transparency index. Preprint at https://arxiv.org/abs/2310.12941 (2023).
122. Bubeck, S. et al. Sparks of artificial general intelligence: early experiments with GPT–4. Preprint at https://arxiv.org/abs/2303.12712 (2023).
123. Rood, J. E., Maartens, A., Hupalowska, A., Teichmann, S. A. & Regev, A. Impact of the Human Cell Atlas on medicine. *Nat. Med.* **28**, 2486–2496 (2022).
124. Han, X. et al. Construction of a human cell landscape at single-cell level. *Nature* **581**, 303–309 (2020).
125. Baysoy, A., Bai, Z., Satija, R. & Fan, R. The technological landscape and applications of single-cell multi-omics. *Nat. Rev. Mol. Cell Biol.* **24**, 695–713 (2023).

Acknowledgements We thank our collaborators across institutions for their support and insightful contributions throughout the development of the manuscript.

Author contributions H.C., B.W. and F.J.T. conceptualized the study. H.C. led the development of the manuscript and drafted the initial version. H.C., A.T.-L., M.B., J.S.-R., S.C., H.G., M.L. and F.J.T. contributed to refining the concepts and methodology. A.T.-L., M.B. and H.G. provided substantial input on the sections related to single-cell data and opportunities. J.S.-R. and S.C. contributed to discussions on gene function prediction and regulatory network reconstruction. F.J.T. and B.W. supervised the overall research direction and advised on critical revisions. All authors reviewed, edited and approved the final manuscript.

Competing interests F.J.T. consults for Immunai, CytoReason, Cellarity, BioTuring and Genbio. AI, and has an ownership interest in Dermagnostix GmbH and Cellarity. B.W. serves as a scientific advisor to Shift Bioscience, Deep Genomics and Vevo Therapeutics, and acts as a consultant for Arsenal Bioscience. H.G. has an ownership interest in Vevo Therapeutics, and is an advisor to Verge Genomics and Deep Forest Biosciences. J.S.-R. reports funding from GSK, Pfizer and Sanofi, and fees and/or honoraria from Travere Therapeutics, Stadapharm, Astex, Owkin, Pfizer, Grunenthal, Moderna and Tempus. M.L. owns interests in Relation Therapeutics and AIVIVO, and is a scientific cofounder and part-time employee at AIVIVO. The other authors declare no competing interests.

Additional information

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41586-025-08710-y.

Correspondence and requests for materials should be addressed to Fabian J. Theis or Bo Wang.

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