Acta Pharmaceutica Sinica B 2026;16(1):122−136

Chinese Pharmaceutical Association
Institute of Materia Medica, Chinese Academy of Medical Sciences

Acta Pharmaceutica Sinica B

www.elsevier.com/locate/apsb
www.sciencedirect.com

REVIEW

Omics-based large language models: A new engine for drug discovery innovation

Xia Sheng[a,b,†], Xiaoya Zhang[a,b,†], Yuxin Xing[a,b,c,†], Yuqi Shi[a,b], Chuanlong Zeng[a,b], Xiaochu Tong[a,b,*], Mingyue Zheng[a,b,c,*], Xutong Li[a,b,*]

[a]*Drug Discovery and Design Center, State Key Laboratory of Drug Research, Shanghai Institute of Materia Medica Chinese Academy of Sciences, Shanghai 201103, China*
[b]*University of Chinese Academy of Sciences, Beijing 100049, China*
[c]*School of Pharmaceutical Science and Technology, Hangzhou Institute for Advanced Study, Hangzhou 330106, China*

Received 12 January 2025; received in revised form 15 October 2025; accepted 16 October 2025

KEY WORDS

Large language model;
Representation learning;
Generalization;
Omics integration;
Single-cell;
Perturbation modeling;
Target identification;
Drug discovery

Abstract Traditional drug discovery suffers from low efficiency and high attrition rates, largely due to the complexity and heterogeneity of human diseases. Omics technologies offer a systems-level perspective for uncovering disease mechanisms and identifying therapeutic targets, but present challenges such as high dimensionality, noise, and heterogeneity. Large language models (LLMs), originally developed for natural language processing, are emerging as powerful tools to address these issues by capturing complex patterns and inferring missing information from large, noisy datasets. We present a three-part framework: (1) Analyzing how LLM architectures and learning paradigms handle challenges specific to genomics, transcriptomics, and proteomics data; (2) Detailing LLM applications in key areas: uncovering disease mechanisms, identifying drug targets, predicting drug response, and simulating cellular behavior; (3) Discussing how insights from omics-integrated LLMs can inform the development of drugs targeting specific pathways, moving beyond single targets towards strategies grounded in underlying disease biology. This framework provides both conceptual insights and practical guidance for leveraging LLMs in omics-driven drug discovery and development.

This article is part of a special issue entitled: Machine Learning in Drug Discovery published in Acta Pharmaceutica Sinica B.

*Corresponding authors.
E-mail addresses: s19-tongxiaochu@simm.ac.cn (Xiaochu Tong), myzheng@simm.ac.cn (Mingyue Zheng), lixutong@simm.ac.cn (Xutong Li).
[†]These authors made equal contributions to this work.

Peer review under the responsibility of Chinese Pharmaceutical Association and Institute of Materia Medica, Chinese Academy of Medical Sciences.

https://doi.org/10.1016/j.apsb.2025.10.034

1. Introduction

Traditionally, drug discovery has primarily followed a target-based paradigm, wherein researchers first identify disease-associated molecular targets, followed by high-throughput compound screening and lead optimization[1]. While this strategy has led to numerous therapeutic breakthroughs, it suffers from several limitations, including low efficiency, high attrition rates, and limited applicability to complex diseases. In recent years, the field has been undergoing a gradual shift from single-target strategies toward a more holistic, systems-level understanding of disease mechanisms and therapeutic interventions, emphasizing the identification of key nodes and pathways in molecular and cellular networks[2,3]. Notably, this transformation has been significantly enabled by the emergence of omics technologies[4].

Omics data—characterized by their multidimensionality, high throughput, and systems-level perspective—offer unprecedented opportunities to elucidate disease mechanisms, uncover novel therapeutic targets, and predict drug responses[5–7]. In particular, certain types of omics datasets are generated under defined perturbations, such as small-molecule treatments, gene knockdown or overexpression, enabling systematic and comprehensive capture of the resulting biological responses[8–11]. However, the effective utilization of omics data poses substantial challenges: heterogeneity across data sources, difficulties in cross-scale integration, and considerable noise and redundancy[12]. These issues are further exacerbated at the single-cell level, where extreme data sparsity and variability increase the complexity of analysis[13]. Collectively, such limitations have hindered the full exploitation of omics data in drug development.

Recent advances in large language models (LLMs) have demonstrated remarkable capabilities in pattern recognition, data imputation, and representation learning, offering promising tools for tackling the inherent complexity of omics datasets[14]. A key scientific question now lies in integrating LLM technologies with large-scale omics data to construct foundational models in bioinformatics, thereby accelerating the drug discovery process.

This review provides a systematic overview of the applications of LLMs in omics technologies and their potential for drug discovery from three key perspectives. First, we explore how LLM-based approaches address major challenges in genomics, transcriptomics, and proteomics, emphasizing their capabilities in processing high-dimensional biological data. Second, we highlight recent applications of omics-driven LLMs in four essential areas of drug discovery: disease mechanism elucidation, target identification, drug response prediction, and virtual cell modeling. Third, we discuss how omics data can inform pathway-based therapeutic strategies, enabling data-driven target prioritization and intervention design. Together, this framework aims to provide both conceptual insights and practical guidance for leveraging LLMs in omics-guided drug discovery and development.

2. Unpacking LLMs: from text to omics data

2.1. What makes LLMs powerful: architecture and evolution

In recent years, the rapid advancement of LLMs has brought transformative progress to the field of artificial intelligence (AI), significantly influencing a wide range of real-world applications. Fundamentally, language models are designed to understand contextual information, predict subsequent words, and generate coherent text[15–17]. The term "large language model" typically refers to models trained on large-scale text corpora, characterized by massive parameter counts, high computational demands, broad generalizability, and the capacity to process long-range dependencies in language[18].

From a technical perspective, the remarkable performance of LLMs is largely attributed to innovations in tokenization and the Transformer architecture. Tokenization refers to the process of converting raw text into a sequence of discrete units—known as tokens—which may represent words, subwords, or characters[19]. These tokens serve as the fundamental building blocks that enable the model to process and interpret language. The Transformer architecture leverages a self-attention mechanism to efficiently model dependencies across entire sequences, thereby enhancing contextual understanding[20]. It is typically structured as an encoder–decoder framework, comprising multi-head attention layers that iteratively refine semantic representations. The encoder processes the input sequence into a context-rich representation, while the decoder uses this representation to generate the output step by step. Multi-head attention allows the model to focus on multiple parts of the input simultaneously, capturing complex, long-range dependencies (Fig. 1).

Although all LLMs are built upon the Transformer framework, they have diverged along distinct architectural trajectories. For instance, OpenAI's GPT series employs a decoder-only architecture optimized for autoregressive text generation[21,22]; Google's BERT utilizes a bidirectional encoder to model deep contextual relationships[23]; and T5 maintains the original encoder–decoder design, reformulating all tasks into a unified text-to-text framework[24].

As LLM architectures and modeling capabilities continue to evolve, these models are increasingly being adapted to domain-specific applications, demonstrating strong potential and versatility in biomedical research, omics analysis, and drug discovery.

2.2. Omics as language: data foundations for LLMs

Emergent abilities refer to qualitative shifts in model behavior that arise only when language models surpass certain scale thresholds, enabling them to perform tasks not explicitly encountered during training[25]. These capabilities are generally attributed to the synergistic scaling of model parameters, training data volume, and computational power. Among these factors, the availability of large and diverse training data is particularly crucial, as it directly shapes the knowledge base and generalization capacity of LLMs. In recent years, rapid advances in omics technologies, particularly in single-cell omics, have led to the accumulation of vast, high-dimensional biological datasets[4,26,27]. In this section, we introduce the core characteristics and sources of these data, providing a foundation for evaluating their potential role in LLMs.

Omics encompasses a suite of high-throughput biological disciplines dedicated to the systematic profiling of biomolecules such as DNA, RNA, proteins, and metabolites. Major omics fields—including genomics, transcriptomics, proteomics, epigenomics, and metabolomics—provide multi-layered insights into the regulation and function of biological systems[28–31]. Technological advances in sequencing and mass spectrometry (Fig. 2) have significantly enhanced molecular coverage, resolution, and

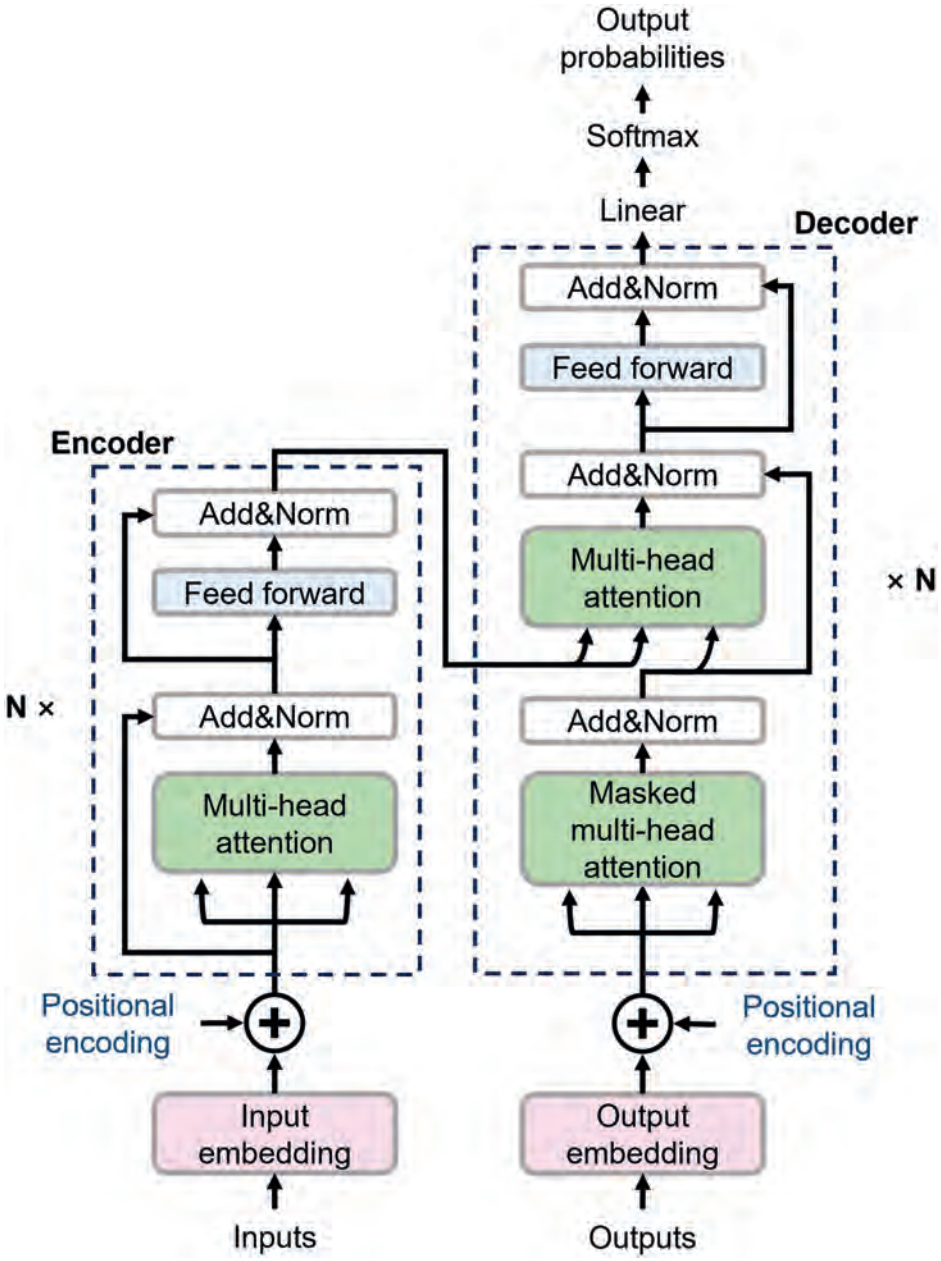

Figure 1 Architecture of transformer.

the capacity to generate rich, multimodal datasets across diverse biological conditions[27,32–37]. Emerging subfields like spatial omics and single-cell omics further enrich this landscape by enabling molecular profiling within native spatial or cellular contexts, which is particularly valuable for understanding tissue heterogeneity and disease microenvironments[38].

Sequencing-based technologies generally serve two primary functions: determining molecular sequences and quantifying expression levels. While sequence data (*e.g.*, DNA, RNA, and protein) are typically linear and structured, gene expression is represented as higher-dimensional matrices that capture variations in gene activity across cells or conditions. These fundamental differences in data structure motivate the development of specialized modeling approaches to effectively leverage omics data within the framework of LLMs (Table 1[39–46]).

2.3. Structuring omics for LLMs: from data complexity to representation design

While inherently rich in biological information due to their high dimensionality, omics datasets often face impediments to deep mining owing to noise and heterogeneity[47,48]. Noise—stemming from operational variability, instrument errors, cross-platform discrepancies, and environmental factors—complicates signal extraction and makes batch correction a pervasive challenge. Beyond batch effects, cross-omics integration further demands careful handling of modality-specific discrepancies[49]. Heterogeneity, while biologically meaningful, often arises from multiple overlapping factors such as genetic background, transient cell states, and regulatory variability, making it difficult to model and interpret systematically. Moreover, while high dimensionality

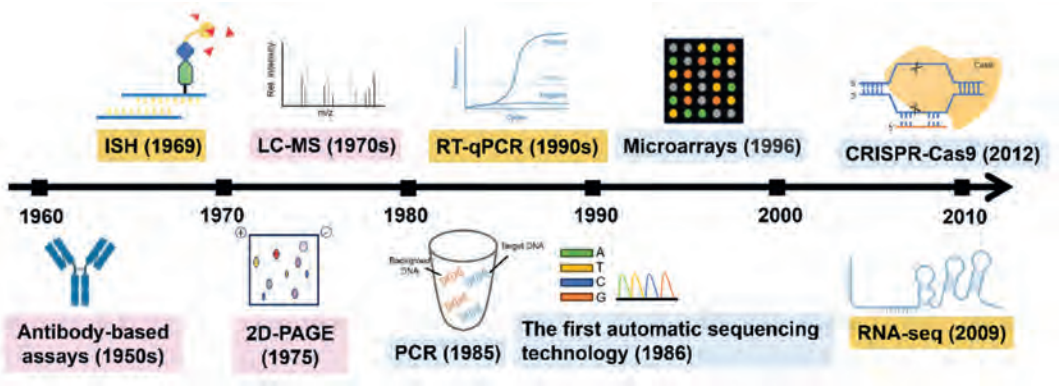

Figure 2 An overview of the development of omics technologies, including genomics, transcriptomics, proteomics (blue: genomics; yellow: transcriptomics; pink: proteomics).

Table 1 Representative omics data for language models.

Modality	Data form	Data source	Tokenization strategies	Example	Ref.
Genomics (DNA)	DNA sequences (A/T/C/G)	Human genome (~2.75 B nucleotide bases)	Nucleotide as token: k-mer	DNABERT (2021)	39
		Human genome (~2.75 B nucleotide bases), multi-species genome references (~32.5 B nucleotide bases)	Nucleotide as token: byte-pair encoding (BPE)	DNABERT-2 (2024)	40
Transcriptomics (mRNA)	Transcript sequences (A/U/C/G)	ncRNA sequences (~23 M)	Nucleotide as token: one-hot encoding	RNA-FM (2022)	41
Protein sequences/ structures	Amino acid sequences, MSAs, 3D structures	UniRef (~65 M protein sequences)	Residue as token	ESM-2 (2024)	42
Single-cell transcriptomics	Cell-by-gene expression matrix	PanglaoDB (~1.1 M human cells, 74 tissues, 209 datasets)	Gene as token: binning for value	scBert (2022)	43
		CELLxGENE (~33 M normal human cells)		scGPT (2023)	44
		Genecorpus-30 M (~30 M human scRNA-seq cells from 561 datasets, multiple public sources)	Gene as token: Ranking for value	Geneformer (2023)	45
Spatial transcriptomics	Cell-by-gene expression matrix + 2D coordinates	~11.4 M cells (~8.7 M scRNA-seq, ~2.7 M SRT) from public datasets (HCA, HTCA, GEO, CosMx SMI)	Cell as tokens	CellPLM (2023)	46

allows for precise and detailed feature characterization, it also contributes to data sparsity. A prime example is single-cell RNA sequencing (scRNA-seq), which exhibits pronounced sparsity with many zero counts, arising from limited capture efficiency and technical variability inherent in single-cell protocols[50,51].

To address these complex challenges, LLMs, adept at uncovering patterns and relationships within large-scale textual datasets, offer a powerful framework for managing high dimensionality, modeling intricate dependencies, and even inferring missing information. However, their effective application to omics requires careful adaptation. In particular, domain-specific tokenization strategies present both a significant challenge and a critical opportunity for effectively modeling diverse omics modalities.

For DNA and RNA sequences, which naturally exhibit text-like properties, common approaches include one-hot encoding of individual nucleotides[41], k-mer segmentation that divides sequences into fixed-length, overlapping substrings[39], and subword methods such as byte pair encoding (BPE) that learn variable-length, data-driven tokens[40]. For protein sequences, models like ESM-2 treat each amino acid as a token, effectively capturing evolutionary constraints encoded in the primary sequence[42]. In the context of single-cell data, where large-scale expression profiles are available, the analogy extends to treating each cell as a "sentence" and each gene as a "word". Cells are represented through their gene expression profiles and the proteins encoded by those genes, requiring tokenization approaches that preserve both gene identity and quantitative expression information. For example, scBERT applies a binning strategy to discretize continuous expression values into categorical tokens, thereby reducing technical noise[43]. scGPT extends this concept by incorporating three types of tokens into its input embeddings: (1) gene names, representing identity; (2) adaptively discretized expression values, tailored to each cell; and (3) condition tokens, capturing meta-information associated with individual genes[44]. Unlike

scBERT's global binning, scGPT's adaptive strategy enables more fine-grained, cell-specific representation. Geneformer takes another approach by ranking genes within each cell according to expression levels, thereby representing the cell as an ordered "sentence" of genes[45]. Beyond gene-level tokenization, some models focus on cell-level representations. For instance, CellPLM treats individual cells as tokens and tissues or spatial regions as sentences, leveraging spatially resolved transcriptomics data to capture intercellular relationships[46]. Overall, these diverse tokenization strategies aim to preserve biologically meaningful structure in single-cell data, thereby enhancing the ability of LLMs to model cellular heterogeneity and support downstream analyses (Tables 1[39−46]).

In summary, the successful adaptation of LLMs to omics data depends on understanding the distinct characteristics of each modality and developing tailored representations, setting the stage for models that can learn meaningful biological patterns.

3. Applications of LLMs in bioinformatics

Understanding biological information flow, from DNA through RNA to functional proteins, requires sophisticated computational approaches to address diverse predictive challenges. LLMs are increasingly pivotal in tackling these challenges, leveraging self-supervised pre-training on vast, unlabeled sequence datasets to learn complex biological patterns and long-range dependencies[52,53]. Compared with earlier deep learning models, LLMs overcome critical limitations including restricted context windows and limited receptive fields, thus enabling them to handle large-scale, long-sequence inputs and to capture a global understanding of long-range interactions across biological sequences[53,54]. Through automated pattern recognition, LLMs develop a semantic understanding of underlying biological relationships, integrating multi-omics data including genomic sequences, protein structures, and functional annotations into unified representation spaces[55]. Subsequent fine-tuning enables breakthroughs across the key stages of the central dogma process.

At the DNA level, foundational genomics research involves predicting how the genome functions and is regulated[34,56]. Core tasks include understanding the impact of genetic variation, analyzing epigenetic marks governing gene expression, and identifying essential regulatory elements, among others (Fig. 3). Progressing to the RNA transcriptome layer involves analyzing the dynamic RNA intermediates derived from DNA. Unlike genomics, transcriptomics presents unique challenges due to its dynamic and condition-specific nature—the same genome can give rise to different transcriptomes under varying environmental conditions, making RNA sequences less conserved than DNA[27]. Transcriptomic LLMs address this variability through transfer learning and self-supervised learning, enabling inference of regulatory mechanisms across diverse environments. Predictive applications encompass RNA modification detection, structural conformation prediction, and functional feature annotation, among others (Fig. 3)[56]. Finally, at the protein level, proteomics grapples with predictions concerning the structure, function, and interactions of functional molecules[57]. This encompasses tasks such as designing novel protein sequences, determining complex tertiary folds and functions, and predicting interactions within large networks (Fig. 3).

Through comprehensive pre-training on biological sequences, these foundation models learn sequence patterns and biological contexts that can be adapted for specific downstream tasks, contributing to biological discovery and enabling their applications across the central dogma. Table 2[39,41,58−62] summarizes the training data, framework, applications, and deficiencies of several representative models.

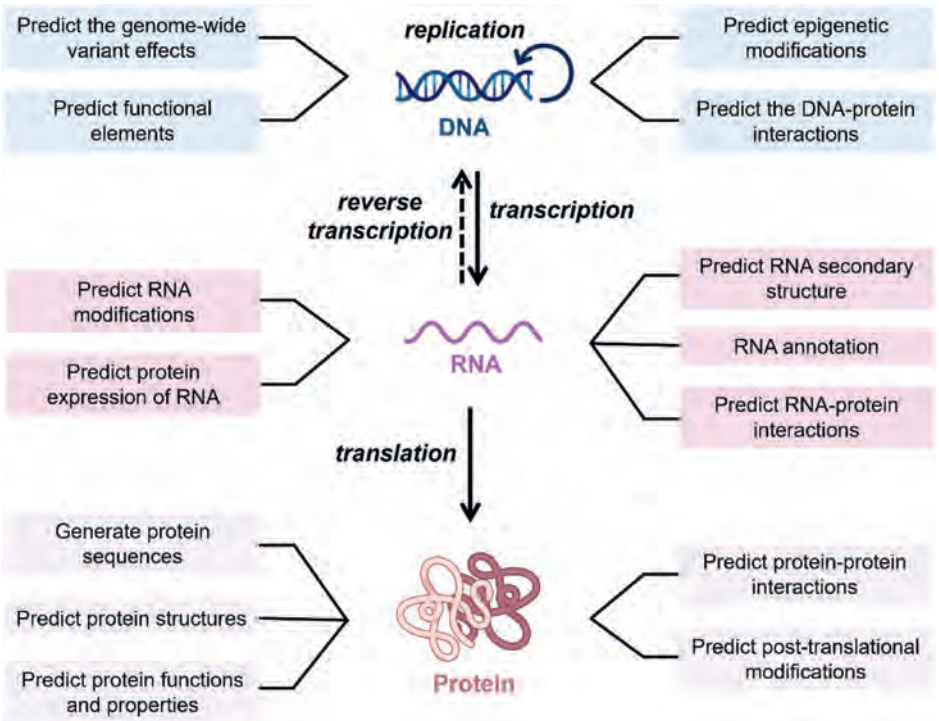

Figure 3 Applications of LLMs in genomics, transcriptomics and proteomics in order of the central dogma.

Table 2 Representative LLMs in bioinformatics.

Model	Dataset	Architectures	Training protocol	Application	Limitation	Ref.
DNABERT	Human genome	12-layer transformer 768 hidden units 12 attention heads k-mer tokenization	120 k steps Batch size 2000 8 × NVIDIA 2080Ti GPUs ∼25 days training	Promoter prediction DNA methylation prediction Viant effect prediction Splice site prediction Gene expression prediction	Generalization across species High computational costs	39
RNA-FM	RNAcentral database (23 million ncRNA sequences)	12-layer transformer 640 hidden units 20 multi-head attention Single nucleotide tokens	MLM self-supervised learning 8 × A100 GPUs One month training	RNA secondary/3D structure prediction Protein−RNA interaction modeling Evolutionary trend inference	Interpretability to specific tasks	41
HyenaDNA	Single human reference genome	Decoder-only hyena architecture 2−8 layers 128−256 width Single nucleotide tokens	10−20 k steps Batch size 64−256 8 × A100 GPUs 80 min training	Promoter prediction Enhancer identification Epigenetic marks prediction Chromatin profiling Species classification	Generalization across species	58
BERT2OME	RMBase database	12-layer transformer 768 hidden units 12 attention heads 2D CNN 2 hidden layers	MLM self-supervised learning Batch size 10−30 25−100 epochs	2′-*O*-methylation modification site prediction across multiple species (human, mouse, yeast)	Generalization across species	59
GenerRNA	RNAcentral database (29.75 million RNA sequences)	24-layer transformer Self-attention mechanism Model dimension 1280 BPE tokenization	MLM self-supervised learning Batch size 128 20 epochs 16 × A100 GPUs One week training	Protein-binding RNA generation Therapeutic RNA design Novel sequence generation with stable secondary structures	Lack of protein-specific fine-tuning datasets High computational costs	60
AlphaGenome	ENCODE (RNA-seq, PRO-cap, DNase-seq, ATAC-seq, TF ChIP-seq, histone ChIP-seq); GTEx (RNA-seq); FANTOM5 (CAGE data); 4D Nucleome portal (genomic contact maps)	U-Net-inspired backbone 9-layer transformer 768-1536 hidden units 8 attention heads Single nucleotide tokens	Batch size 64 15 k steps pre-training 250 k steps distillation 512 TPUv3 cores 64 H100 GPUs	Genome track prediction Variant effect prediction Splicing analysis Gene expression prediction Contact map prediction	Generalization across species Lack of evaluation systems High computational costs	61
ProLLM	Human dataset (4577 proteins, 75875 interactions); SHS27K, SHS148K, STRING datasets; Mol-Instructions dataset	Flan-T5-large backbone 1024-dim ProtTrans embedding ProCoT format Instruction fine-tuning	Fine-tuning on Flan-T5-large Batch size 2 per device A40-48G GPU	Protein−protein interaction prediction Signaling pathway analysis Interaction type classification Chain-of-thought reasoning for protein relationships	High computational costs	62

GPU, graphics processing unit; TPU, tensor processing unit; MLM, masked language model; CNN, convolutional neural network.

3.1. Capturing long-range dependencies: LLMs enabling global modeling of DNA regulatory syntax

Genomic sequence modeling refers to the computational process of learning representations from DNA sequences to elucidate their functional roles in gene regulation and cellular processes[53]. Traditional deep learning methods, such as the convolutional neural network (CNN)-based model DeepBind[63], detect local regulatory patterns using fixed-length scanning filters but have limited ability to capture long-range dependencies critical for understanding regulatory syntax. This constraint hampers predictive accuracy in genome-wide functional annotation tasks.

The emergence of LLMs has revolutionized genomic sequence modeling by overcoming these fundamental limitations through

the capture of global contextual dependencies. A representative example is DNABERT, which adapts BERT architecture to DNA sequences, employing a 12-layer Transformer with multi-head self-attention, which overcomes the restricted receptive fields in traditional CNN/RNN models[39,40]. Following a pre-training and fine-tuning paradigm, DNABERT learns general genomic syntax through masked language modeling (MLM) on 2.75 billion nucleotides of unlabeled human genome sequences and is then adapted to specific tasks with minimal labeled data. By using k-mer tokenization and processing sequences up to 512 tokens, the model enables comprehensive analysis of long-range regulatory interactions. Across diverse genomic tasks, including promoter prediction, transcription factor binding site identification, and regulatory element classification, DNABERT achieves a 33.5% accuracy improvement over CNN-based methods and attains mean accuracy above 0.9 across 690 ENCODE datasets[64]. Its attention mechanisms further provide direct visualization of critical regulatory regions without the need for human guidance.

Despite DNABERT's success, the quadratic computational complexity of attention mechanisms fundamentally limits the modeling of long-range genomic interactions that span hundreds of thousands of nucleotides. HyenaDNA addressed this by replacing attention with Hyena operators, which combine long convolutions and data-controlled gating to reduce complexity from $O(L^2)$ to $O(L \log_2 L)$. As a result, it can model sequences up to 1 million nucleotides at single nucleotide resolution, representing a 500-fold increase in sequence length compared to existing dense attention-based models[58]. Building upon these advances, Caduceus introduced reverse complement equivariance to model DNA's bidirectional nature *via* BiMamba and MambaDNA modules, which process both forward and reverse complement sequences with shared parameters. It achieves superior performance in variant effect prediction and regulatory element discovery while using significantly fewer parameters[65]. These innovations mark a paradigm shift from purely computational solutions toward biologically informed model design, embedding domain-specific constraints directly into model architectures for both computational efficiency and biological relevance.

3.2. Learning generalizable representations: LLMs advancing RNA foundation modeling

Conventional computational approaches for biological sequence analysis rely heavily on handcrafted features and task-specific model architectures, requiring extensive domain knowledge for effective feature engineering. For example, early deep learning models such as iRNA-PseU[66] combine BiLSTM[67] networks with manually engineered nucleotide chemical properties (electron affinity and hydrogen bonding strength) for specific modification site prediction. These methods lack generalization capabilities, often necessitating the training of separate models for different modification types and species, thus limiting their applicability across diverse biological scenarios.

The emergence of RNA foundation models marks a paradigm shift from task-specific architectures to universal representation learning. RNA-FM[41] exemplifies this transformation with a 12-layer BERT-based model trained *via* self-supervised learning using 23 million non-coding RNA sequences from the RNAcentral database. In contrast to traditional methods requiring manual feature engineering, RNA-FM employs MLM to learn contextual representations by predicting randomly masked nucleotides (15% of input tokens), capturing both local sequence patterns and long-range dependencies without annotated data. The resulting embeddings encode multi-scale biological information, including structural similarities, functional relationships, and evolutionary conservation across RNA families. Notably, RNA-FM enables zero-shot prediction of RNA secondary structures and modification sites, demonstrating that pre-training on unlabeled sequences can capture fundamental biophysical principles governing RNA behavior. This enables robust performance across diverse downstream applications with minimal task-specific fine-tuning.

Building upon foundational advances, subsequent research has evolved toward foundation-model-based specialization, leveraging universal representations for efficient task-specific fine-tuning and representing a fundamental departure from traditional isolated approaches. Models such as BERT2OME[59] and GenerRNA[60] exemplify this paradigm by inheriting general sequence understanding while adding specialized modules to address multi-task integration and data synthesis challenges.

BERT2OME[59] integrates pre-trained BERT embeddings with a 2D-CNN architecture for multi-task learning across seven RNA modification types. By treating the embedding matrices as spatial feature maps to extract local dependencies and employing SMOTE to address data imbalance, the model achieves an impressive 99.15% accuracy on human RNA sequences.

GenerRNA[60] introduces a generative pre-training paradigm for RNA design using a 350-million parameter autoregressive Transformer trained on 29.75 million sequences. Unlike discriminative models, this generative paradigm creates novel RNA sequences with desired properties, addressing data scarcity by synthesizing training data for underrepresented RNA families and enabling controlled generation of high-affinity protein-binding RNAs.

3.3. Cross-modality integration: LLM-unified representation powering multi-task intelligence

Traditional deep learning frameworks are inherently limited in handling multimodal biological data, typically processing single data types in isolation. For example, models such as DeepSEA[68] employ CNNs to identify transcription factor binding sites within DNA sequences, but cannot integrate complementary information sources like epigenetic modifications or protein−DNA interaction networks. This modality segregation prevents models from capturing essential cross-modal associations, as protein sequences, gene expression profiles, and pathway annotations are analyzed by independent architectures that fail to leverage their inherent biological interconnections. Consequently, these single-modality approaches encounter performance ceilings in complex tasks that require integrated, multi-source reasoning—particularly when structural, functional, or literature-derived contextual information is essential.

LLMs fundamentally transform genomic analysis by enabling unified embedding spaces that integrate multimodal biological information across diverse regulatory mechanisms. AlphaGenome exemplifies this paradigm shift by simultaneously processing 11 genomic modalities, including gene expression, splicing, chromatin accessibility, histone modifications, transcription factor binding, and contact maps within a single framework[61]. Unlike conventional methods requiring separate specialized models, AlphaGenome introduces a multi-scale U-Net-inspired architecture that integrates 1D sequence embeddings for linear tracks with 2D pairwise embeddings for spatial interactions. This design captures both local sequence motifs and long-range dependencies across megabase-scale contexts. Key innovations of the model

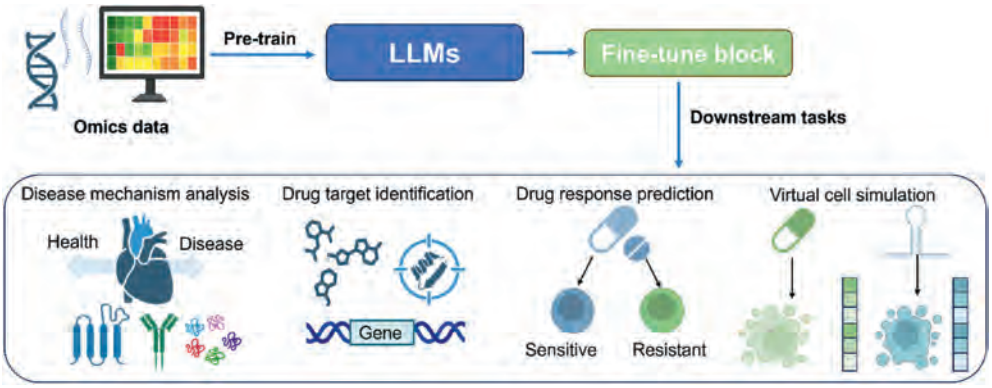

Figure 4 Downstream applications of LLMs in drug discovery (Created with BioRender.com).

include a Transformer-CNN fusion module for multi-scale pattern recognition and a two-stage training paradigm: multi-task pre-training across all 11 modalities, followed by knowledge distillation from 64 teacher models using sequence augmentation. This approach encourages the model to uncover shared regulatory principles that underlie diverse biological processes. Leveraging comprehensive datasets from ENCODE[64], GTEx[69] and FANTOM5[70] consortia, AlphaGenome processes 5930 human and 1128 mouse genome tracks across diverse cell types. This unified architecture achieves state-of-the-art performance on 22 out of 24 genome track prediction tasks and 24 out of 26 variant effect prediction benchmarks, surpassing specialized models like ChromBPNet[71], SpliceAI[72] and Orca[73]. These results demonstrate how LLMs can effectively integrate heterogeneous genomic data through joint representation learning.

ProLLM[62] advances beyond static sequence-function representations by introducing systems-level biological reasoning through its protein chain of thought (ProCoT) mechanism. It reformulates protein−protein interaction (PPI) prediction as a natural language reasoning task, simulating cellular signaling pathways from upstream receptors through intermediate signaling proteins to downstream effectors. ProCoT transforms multi-step protein interactions into sequential reasoning chains, using Transformer architectures to capture indirect protein associations beyond direct physical interactions. To support this process, ProLLM incorporates protein embeddings from ProtTrans[74] and applies instruction fine-tuning on biological datasets, enabling dynamic tracing of complex signal transduction events. It achieves micro-F1 scores of 91.05, 85.32, 87.66, and 89.21 across four PPI datasets, significantly outperforming traditional graph neural network methods. These results underscore the power of biological reasoning in modeling complex protein interaction networks.

4. Applications of LLMs in drug discovery

Drug discovery is typically a prolonged and inefficient process, primarily due to escalating experimental costs and time, as well as a limited understanding of disease-related mechanisms. Unlike conventional approaches that rely on static biomarkers or drug molecular features, LLMs trained on massive single-cell omics datasets can model dynamic, context-dependent gene regulatory relationships, simulate perturbation effects, and infer causal mechanisms at single-cell resolution. From elucidating disease mechanisms and identifying therapeutic targets to predicting drug sensitivity and simulating cellular responses, LLM-powered frameworks provide a unified, versatile, and biologically informed approach. Together, these developments underscore the transformative role of LLMs in reshaping how we model, understand and intervene in complex biological systems. We summarized the applications of LLMs in drug discovery (Fig. 4), along with the training data, framework and deficiencies of representative models (Table 3[44,45,75,76]).

4.1. From static markers to dynamic networks: LLM-enabled discovery of disease mechanisms

Understanding the cellular and molecular basis of disease is fundamental to drug discovery. Complex diseases often involve multiple interacting cell types, but traditional methods—such as histopathological analysis and animal models—lack the resolution needed to pinpoint cell type-specific driver genes within heterogeneous disease microenvironments[77].

Earlier computational approaches typically focused on identifying differentially expressed genes (DEGs) from bulk omics data, followed by pathway enrichment. However, this strategy overlooks the hierarchical structure of gene regulation and fails to capture dynamic, continuous cell state transitions, often conflating causative genes with downstream effects.

LLMs, pre-trained on massive single-cell atlases, can learn gene co-expression patterns and cellular trajectories. When fine-tuned on disease-specific datasets, they enable precise identification of pathogenic cell states and prediction of perturbation outcomes. This capability facilitates virtual screening of candidate driver genes, mapping of dynamic regulatory networks, and in-depth mechanistic insights—while reducing the sample size demands of conventional experimental studies.

Geneformer[45], a context-aware LLM incorporating attention mechanisms, has demonstrated strong performance in network-based biological prediction tasks. Trained on Genecorpus-30M, a large-scale pre-training corpus comprising 29.9 million human single-cell transcriptomes, Geneformer encodes transcriptomic data into rank-based representations. These representations emphasize key genes that define cell states while reducing the impact of noise and technical artifacts. The inputs are processed through six Transformer encoder blocks under an MLM strategy, mapping genes into high-dimensional space and dynamically refining their representations based on the cellular context to capture expression differences across cell states.

Geneformer introduces an innovative perturbation simulation strategy, wherein one or more genes are removed directly from the

Table 3 Representative LLMs in drug discovery.

Model	Dataset	Architectures	Training protocol	Application	Limitation	Ref.
scGPT	CELLxGENE	12-layer transformer 512 hidden units 8 attention heads Binning for value tokenization	Masked learning Batch size 32 6 epochs NVIDIA GPU	Cell type annotation Gene function prediction Perturbation prediction	Performance optimization bottleneck Model scale inefficiency	44
Geneformer	Genecorpus-30 M	6-layer transformer 256 hidden units 4 attention heads Rank value tokenization	Masked learning 12 × V100 GPUs Batch size 12 3 epochs Three days training	Cell type annotation Gene function prediction Gene regulatory networks prediction	Performance optimization bottleneck Generalization across biological contexts	45
scFoundation	Over 50 million human single-cell transcriptomics data	12-layer transformer 768 hidden units 12 attention heads Gene expression tokenization	Read depth-aware modeling 12 × A100 GPUs Two weeks training	Cell type annotation Gene regulatory networks prediction Drug response prediction	Computational cost bottleneck Model scale inefficiency	75
STATE	Tahoe-100 M dataset Parse-PBMC dataset	State embedding: 16-layer transformer 2048 hidden units 16 attention heads State transition: Bidirectional self-attention layer	State embedding 8 × H100 GPUs	Gene profiles prediction	Lack of evaluation systems	76

GPU, graphics processing unit.

transcriptomic encoding to mimic gene repression or activation. By comparing cellular or gene embeddings before and after *in silico* gene deletion, the model accurately assesses the functional impact of target genes and the downstream network effects. When fine-tuned on single-cell RNA-seq data from fetal cardiomyocytes in the Fetal Cell Atlas[78], Geneformer successfully distinguished cardiomyocytes affected by hypertrophic and dilated cardiomyopathy. By quantifying how simulated gene perturbations shifted cellular states toward either disease phenotype, the study identified 478 key genes driving phenotypic transitions and revealed relevant genes and signaling pathways. Among them, gelsolin (GSN) and phospholamban (PLN) induced the most pronounced disease-state shifts upon perturbation. Subsequent CRISPR-mediated knockout experiments confirmed that silencing these genes significantly improved contractile stress responses in cardiac microtissues.

Previous studies have primarily focused on the application of LLMs in deciphering disease mechanisms at the single-gene level. However, gene function *in vivo* is fundamentally governed by complex molecular interaction networks. Cellular states emerge from intricate interplay among biomolecules, making the reconstruction and interpretation of gene regulatory networks (GRNs) central to understanding cellular function. Embedding genes within a network context enriched with molecular functional information provides a valuable dimension to gene annotation and offers a critical foundation for systems biology research[12].

scPRINT[79] is a bidirectional Transformer model specifically designed for genome-scale, cell-specific gene network inference using scRNA-seq data. It is pre-trained on a large dataset of 50 million cells from the CELLxGENE repository, encompassing diverse species, disease types, and ethnic backgrounds. In contrast to Geneformer's MLM-based training, scPRINT introduces a novel bottleneck learning task: the model compresses high-dimensional gene expression profiles into latent embeddings and reconstructs the original expression matrix. This forces the model to learn and retain structural information about gene−gene interactions during the compression process.

In the context of disease mechanism studies, scPRINT can infer and compare GRNs from disease and healthy single-cell atlases, revealing cell state-specific rewiring of regulatory networks. For instance, in a study of benign prostatic hyperplasia (BPH)[80], scPRINT reconstructed GRNs for fibroblasts from both BPH patients and healthy individuals. The model identified prostate-associated gene 4 (*PAGE4*) as a disease-specific hub gene in BPH fibroblasts. The *PAGE4*-centered subnetwork was significantly enriched in pathways related to metal ion exchange, oxidative stress response, and inflammation, thereby uncovering a cascade of inflammation-fibrosis signaling events likely driving BPH progression.

By inferring gene regulatory relationships at single-cell resolution, scPRINT expands gene function annotation from isolated molecular features to dynamic interaction networks. This provides a powerful computational framework for decoding network-level regulatory mechanisms that underpin disease initiation and progression.

To conclude, omics-based LLMs offer a transformative approach to understanding disease-associated cellular and molecular mechanisms, directly pointing to potential therapeutic targets. By enabling high-resolution mechanistic analysis at the single-cell level, these models enhance both the depth and efficiency of complex disease research and lay a solid foundation for the advancement of precision medicine.

4.2. From perturbation signals to cell representation: LLMs accelerating target discovery

Identifying therapeutic targets is critical for drug discovery, yet large-scale experimental validation remains costly, labor-intensive, and time-consuming, with a relatively low success rate. While structure-based AI methods have shown promise, they

are constrained to proteins with known 3D structures and often lack the ability to account for cell-type-specific functional contexts.

LLMs have demonstrated strong capabilities in modeling the intrinsic relationships between gene function and cellular states through large-scale pre-training. One key application of LLMs lies in perturbation prediction—the simulation of gene inhibition or activation and its resulting impact on the cell's transcriptomic landscape. This enables the identification of genes whose perturbation may reverse disease-associated cellular states toward healthy phenotypes. Unlike disease mechanism modeling, which predicts cell state changes in response to known perturbations, target identification emphasizes reverse perturbation prediction, inferring the causal genetic perturbation from an observed cell state. This strategy supports both the discovery of key regulators of cellular lineage progression and the identification of actionable therapeutic targets.

A representative model of this paradigm is scGPT[44], a Transformer-based language model designed for single-cell applications. Built upon stacked Transformer layers with multi-head attention, scGPT accepts numerical embeddings that encode gene identity, expression levels, and experimental conditions. The model is pre-trained in a self-supervised manner on over 33 million normal human cells from the CELLxGENE database[81]. By jointly optimizing cell- and gene-level representations, scGPT overcomes the non-sequential nature of gene expression profiles and exhibits strong generalizability across diverse datasets and tasks.

In reverse perturbation prediction tasks, scGPT has demonstrated remarkable performance. When fine-tuned on the Norman dataset[82], which includes perturbations of 20 genes, the model accurately inferred the genetic perturbation for given cell states. On previously unseen perturbed cells, scGPT achieved a top-1 accuracy of 91.4%, successfully identifying the target gene responsible for the observed transcriptomic changes.

Beyond gene perturbation tasks, LLMs are also reshaping traditional drug target prediction pipelines. These pipelines are typically built on the assumption that drugs sharing similar mechanisms of action induce comparable transcriptional perturbations. Models such as CMap[83], SSGCN[84], and PertKGE[85] rely on analyzing DEGs profiles following drug treatment. By comparing transcriptional responses between treated and control conditions, these models estimate the similarity between novel compounds and known drug or knockdown signatures, thereby inferring potential targets.

Within this framework, LLMs can enhance prediction accuracy by learning more expressive representations of DEG features. Models like Geneformer and scGPT embed gene perturbation profiles into high-dimensional latent spaces that capture gene co-expression patterns and their association with cell states. These embeddings distill complex perturbation signals into compact, informative vectors that can be seamlessly with traditional similarity-based inference frameworks such as CMap[83].

Notably, models like GenePT further advance this field by reducing dependence on large-scale expression datasets. Instead, GenePT extracts gene–cell associations directly from unstructured biomedical literature using natural language processing, enabling the generation of meaningful gene embeddings solely from gene IDs[86]. This knowledge-driven approach maintains the core principle of similarity-based inference while extending applicability to contexts where experimental perturbation data are scarce.

Collectively, LLMs offer a unified and versatile framework for therapeutic target identification by linking gene-level perturbation signatures to drug mechanisms of action. Their ability to simulate and reason over complex cellular states, infer causal relationships, and integrate diverse data sources significantly enhances the precision of target prioritization and accelerates early-stage drug discovery.

4.3. From bulk profiles to single-cell insights: LLMs enabling scalable prediction of drug responses

Accurate prediction of drug response, particularly the half-maximal inhibitory concentration (IC_{50}) across diverse cell lines, is essential for precision medicine, as it reflects variability in drug sensitivity and resistance among different cell types[87]. However, traditional *in vitro* screening and bulk RNA-seq-based models, such as Precily[88], often fail to capture microenvironment-driven and subtype-specific drug responses observed in clinical settings. Although most available datasets rely on bulk RNA-seq, single-cell profiling can uncover resistance heterogeneity across distinct cellular subtypes, offering valuable mechanistic insights[89]. Yet, the limited availability of single-cell drug response data remains a significant bottleneck.

LLMs can help address this limitation by leveraging vast unlabeled single-cell datasets and employing transfer learning to enhance generalization and feature representation. This enables the prediction of drug sensitivity and resistance at single-cell resolution, even in scenarios with scarce labeled data.

A prominent example is scFoundation[75], a Transformer-based model with approximately 100 million parameters, pre-trained on over 50 million scRNA-seq profiles covering 19,264 genes. Its architecture incorporates an embedding module and an asymmetric encoder–decoder structure, optimized for both scalability and effective representation learning. A core innovation of scFoundation is its read-depth-aware modeling strategy, specifically designed to mitigate the technical artifacts arising from variable sequencing depth in scRNA-seq data. Unlike traditional masked modeling strategies that focus only on non-zero values, scFoundation randomly masks both zero and non-zero expression values. This encourages the model to capture biologically meaningful gene–gene relationships while reducing technical biases. As a result, scFoundation produces harmonized cell embeddings across heterogeneous datasets, improving performance in downstream tasks such as cell-type annotation, perturbation prediction, and cross-platform integration.

Importantly, scFoundation's pre-trained representations can also enhance bulk-level drug response models. For instance, DeepCDR, a deep learning framework that uses bulk RNA-seq and drug structural information to predict IC_{50} values, shows improved predictive accuracy when augmented with features extracted from scFoundation. Empirical evaluations demonstrate that incorporating these single-cell-informed features significantly improves Pearson's correlation between predicted and actual IC_{50} values across drugs and cancer types. This cross-resolution transfer strategy exemplifies how LLMs can bridge the gap between bulk and single-cell modalities to improve drug response prediction.

In single-cell drug sensitivity prediction, features derived from scFoundation offer significant advantages. Given the scarcity of labeled single-cell drug response datasets, models like SCAD[90] apply domain adaptation to transfer knowledge from bulk to single-cell settings. When trained on unified embeddings generated by scFoundation from both data types, SCAD consistently outperforms baseline models using raw gene expression,

achieving higher area under the curve (AUC) scores across all tested drugs.

Together, these findings highlight the transformative potential of LLMs in constructing shared embedding spaces that unify heterogeneous omics data. By capturing subtle transcriptional signatures and drug-induced shifts at single-cell resolution, models like scFoundation offer a robust computational framework for modeling intrinsic drug sensitivity, particularly within complex tumor and immune microenvironments. This not only enhances the scalability of drug responses prediction but also lays a foundation for personalized therapy design driven by high-resolution transcriptomic profiling.

4.4. From gene profiles to dynamic simulation: LLM-powered virtual cells for predictive pharmacology

Predicting cellular responses to chemical and genetic perturbations is essential for improving drug development efficiency and reducing clinical trial failure rates[91]. While traditional *in vivo* and *in vitro* models remain useful in early screening stages, they are often costly, physiologically limited, and unable to capture the full extent of cellular heterogeneity—contributing to the high ($\sim$90%) attrition rate in clinical trials.

Several conventional models, including TranSiGen[92], DLEPS[93], and DeepCE[94], have been developed to predict gene expression profiles following drug perturbations. Building on this foundation, LLMs can replace traditional feature extraction components by leveraging large-scale single-cell pre-training to learn gene co-expression patterns and capture cell state dynamics. For example, GeneCompass[95], trained on over 120 million single-cell transcriptomes, effectively generates gene embeddings for bulk RNA-seq perturbation tasks and outperforms conventional methods in gene expression prediction.

LLMs have further extended into single-cell perturbation modeling, giving rise to the concept of the "Virtual Cell". In this framework, models take single-cell transcriptomic profiles as input and use the contextual modeling capabilities of Transformer architectures to simulate cellular transcriptional response to perturbations[96,97]. Typically, the input comprises an initial cell state and perturbation information (*e.g.*, drug identity, dosage, or genetic perturbation), and the output is the predicted downstream transcriptional changes. This approach enables fine-grained assessment of cell state transitions across diverse systems—including stem cells, cancer cells, and immune cells—offering both cell-type specificity and perturbation sensitivity.

A prominent example of this approach is STATE[76], an advanced LLM-based virtual cell model trained on millions of single-cell profiles. Its input consists of a transcriptomic profile representing the initial cell state along with associated perturbation metadata. The model outputs the predicted post-perturbation gene expression profile. STATE comprises two key components: (1) State embedding: a bidirectional Transformer that encodes the cell state and perturbation condition into high-dimensional embeddings. (2) State transition: a lightweight, specialized multilayer perceptron that decodes these embeddings into the perturbed gene expression profile. This asymmetric architecture enables the model to generalize across conditions while preserving biological fidelity.

STATE is pre-trained on two large-scale single-cell datasets: (1) Tahoe-100M[98], comprising 100 million transcriptomes spanning 379 small molecules across 50 cancer cell lines, providing rich data for modeling drug−cell interactions. (2) Parse-PBMC[99], which includes over 10 million peripheral blood mononuclear cells from 1152 samples, supporting generalization to complex immune cell populations.

The model achieves state-of-the-art performance in cross-cell-type gene expression prediction, outperforming baseline models across various datasets and metrics. Notably, it enables perturbation prediction by accurately inferring molecular responses in previously unseen cell types using only pre-trained cell embeddings, without the need for fine-tuning. This capability enables robust simulation of cellular dynamics under diverse physiological and pathological conditions. More broadly, such LLM-powered virtual cell models represent a transformative computational paradigm for disease modeling, target discovery, and drug screening. As multi-omics and perturbation datasets continue to grow, this approach holds promise for evolving into a "digital human" platform for predictive pharmacology, ultimately boosting the efficiency and success rate of drug development.

5. Existing challenges and future research directions

While LLMs show promising capabilities in handling noisy and sparse omics inputs through imputation and representation learning, their performance remains fundamentally constrained by the quality of input data[100]. Technical noise, batch effects, and data sparsity, stemming from limited capture efficiency, complicate feature extraction and compromise predictive robustness[101]. As such, careful data preprocessing, including normalization, batch correction, and quality control, remains essential to mitigate artifacts and support biologically meaningful predictions[102,103]. Data scarcity poses an even greater challenge in low-resource omics domains such as metabolomics, microbiomics, and epigenomics, owing to the limited availability of standardized, high-quality datasets. For example, single-cell metabolomics continues to struggle with low experimental sensitivity[104]. Unlike transcriptomics, where RNA can be amplified, there is no comparable amplification technique exists for cellular metabolites. Consequently, many metabolites remain undetected because their concentrations fall below the detection threshold. Furthermore, annotation is particularly difficult given the structural diversity of metabolites. This profound sparsity not only limits dataset standardization but also increases overfitting risk, impeding LLMs from capturing key biological phenomena such as dynamic metabolic regulation in cancer or intricate host−microbe interactions[31].

Another limitation of omics-based LLMs lies in their high computational demands. Pre-training large models, such as Geneformer, reportedly takes about three days on a distributed setup with 12 high-performance GPUs[45]. This resource intensity stems in part from both the characteristics of the training data and the complexity of the model architecture. On one hand, the quadratic computational complexity of Transformer attention mechanisms poses scalability challenges, especially when processing datasets at the million-cell scale, necessitating architectural innovations to improve efficiency. For example, scPRINT adopts efficient sampling strategies and attention optimization techniques to enable pre-training on datasets with 50 million cells without requiring prohibitively expensive hardware resources[79]. On the other hand, larger context windows are typically pursued to better capture comprehensive regulatory information. However, to maintain training efficiency, models generally avoid using the entire gene expression matrix. Instead, they focus on biologically informative subsets, such as genes with high expression or variability, consistent with feature selection strategies commonly used

in traditional statistical models. As a result, many omics-based LLMs enforce strict input sequence length constraints, such as Geneformer (2048 tokens) and scGPT (around 1200 genes per cell). This necessitates the exclusion of lowly expressed or rare, yet biologically important genes, thereby risking the loss of critical regulatory signals[44,45]. Nevertheless, expanding context windows does not consistently yield performance improvements in biological tasks, potentially because many regulatory motifs often acting over short local ranges[53].

Interpretability and generalization further constrain the broader adoption of LLMs in omics research. Most models update all parameters during training, making it difficult to trace the contribution of specific inputs to the final predictions[105]. while attention mechanisms do not imply causal attribution and provide only limited interpretability in biological contexts[106]. Generalization across cell types, disease states, and species also remains a challenge, with models trained in one context often performing poorly in others. For example, in one evaluation using CRISPRi-perturbed K562 cell data to assess model fidelity[107], the LLM-based models Geneformer and scGPT underperformed compared to conventional deep learning models. Despite the rapid advancement of omics-based LLMs, the field still lacks standardized evaluation frameworks. This absence hampers model benchmarking, limits reproducibility, and makes it difficult for researchers to make informed choices in model selection and development.

As experimental techniques continue to evolve, growing biological data drives the emergence of new omics disciplines like microbiomics[108] and epigenomics[109], revealing novel mechanisms through computational analysis. Omics data can be viewed as the "language" of biology, while LLMs act as interpreters, enabling deeper exploration of complex biological systems. Meanwhile, spatial omics technologies are achieving unprecedented resolution, approaching whole-transcriptome coverage for RNA and extending toward whole-proteome measurements for proteins[110]. These technologies are moving into clinical applications for personalized medicine through patient tissue analysis[111]. Integrating LLMs with spatial omics represents a promising new frontier, for example, models like CellPLM[46] combine gene expression data with spatial context to uncover cellular interactions and disease mechanisms. However, effectively bridging the modality gap and addressing spatial data sparsity will require the development of specialized architectures and biologically informed pre-training strategies[112]. To overcome these challenges, key advancement areas for future research include spatial multi-omics integration, computational cell biology, intercellular communication, synthetic biology and spatial bioinformatics tools[113].

At present, directly applying LLMs to *de novo* drug design remains challenging, as it requires integrated analysis of chemical structure, pharmacokinetics, safety, and manufacturability. This necessitates interdisciplinary collaboration across biology, chemistry, pharmacology, toxicology and clinical medicine throughout screening, optimization, preclinical evaluation, and clinical trials.

Acknowledgments

This study was supported by the Strategic Priority Research Program of the Chinese Academy of Sciences (No. XDB0830000, China), National Natural Science Foundation of China (Nos. 82204278, T2225002, and 82273855), SIMM-SHUTCM Traditional Chinese Medicine Innovation Joint Research Program (No. E2G805H, China), Shanghai Municipal Science and Technology Major Project, National Key Research and Development Program of China (Nos. 2023YFC2305904 and 2022YFC3400504), Key Technologies R&D Program of Guangdong Province (No. 2023B1111030004, China), Shanghai Sailing Program (No. 24YF2755600, China), and the China Postdoctoral Science Foundation (No. 2024M763421).

Author contributions

Xia Sheng, Xiaoya Zhang and Yuxin Xing: Writing—Original draft preparation. Yuqi Shi: Formal analysis. Chuanlong Zeng: Formal analysis. Xiaochu Tong: Formal analysis and Supervision. Mingyue Zheng: Conceptualization and Supervision. Xutong Li: Writing—Review & Editing, Conceptualization and Supervision.

Conflicts of interest

The authors have no conflicts of interest to declare.

References

1. Swinney DC, Anthony J. How were new medicines discovered?. *Nat Rev Drug Discov* 2011;**10**:507−19.
2. Sadri A. Is target-based drug discovery efficient? Discovery and "off-target" mechanisms of all drugs. *J Med Chem* 2023;**66**:12651−77.
3. Vincent F, Nueda A, Lee J, Schenone M, Prunotto M, Mercola M. Phenotypic drug discovery: recent successes, lessons learned and new directions. *Nat Rev Drug Discov* 2022;**21**:899−914.
4. Schuster SC. Next-generation sequencing transforms today's biology. *Nat Methods* 2008;**5**:16−8.
5. Xiang Y, Li XD, Gao Q, Xia JF, Yue ZY. ExplainMIX: explaining drug response prediction in directed graph neural networks with multi-omics fusion. *IEEE J Biomed Health Inform* 2025;**29**: 5339−49.
6. Karczewski KJ, Snyder MP. Integrative omics for health and disease. *Nat Rev Genet* 2018;**19**:299−310.
7. Montaner J, Ramiro L, Simats A, Tiedt S, Makris K, Jickling GC, et al. Multilevel omics for the discovery of biomarkers and therapeutic targets for stroke. *Nat Rev Neurol* 2020;**16**:247−64.
8. Peidli S, Green TD, Shen CY, Gross T, Min J, Garda S, et al. scPerturb: harmonized single-cell perturbation data. *Nat Methods* 2024;**21**:531−40.
9. Gavriilidis GI, Vasileiou V, Orfanou A, Ishaque N, Psomopoulos F. A mini-review on perturbation modelling across single-cell omic modalities. *Comput Struct Biotechnol J* 2024;**23**:1886−96.
10. Subramanian A, Narayan R, Corsello SM, Peck DD, Natoli TE, Lu X, et al. A next generation connectivity map: L1000 platform and the first 1,000,000 profiles. *Cell* 2017;**171**:1437−52.e17.
11. Wei ZT, Si DM, Duan B, Gao YC, Yu Q, Zhang ZB, et al. PerturBase: a comprehensive database for single-cell perturbation data analysis and visualization. *Nucleic Acids Res* 2024;**53**:D1099−111.
12. Lähnemann D, Köster J, Szczurek E, McCarthy DJ, Hicks SC, Robinson MD, et al. Eleven grand challenges in single-cell data science. *Genome Biol* 2020;**21**:31.
13. Linderman GC, Zhao J, Roulis M, Bielecki P, Flavell RA, Nadler B, et al. Zero-preserving imputation of single-cell RNA-seq data. *Nat Commun* 2022;**13**:192.
14. Shi XJ, Wang JC, Chung GJ, Julian D, Qiao LP. Data imputation based on retrieval-augmented generation. *Appl Sci* 2025;**15**:7371.
15. Hochreiter S, Schmidhuber J. Long short-term memory. *Neural Comput* 1997;**9**:1735−80.
16. Chelba C, Norouzi M, Bengio S. N-gram language modeling using recurrent neural network estimation. *arXiv preprint* 2017: 1703.10724.

17. Schmidt RM. Recurrent neural networks (RNNs): a gentle introduction and overview. *arXiv preprint* 2019:1912.05911.
18. Liu YH, He H, Han TL, Zhang X, Liu MY, Tian JM, et al. Understanding LLMs: a comprehensive overview from training to inference. *Neurocomputing* 2025;**620**:129190.
19. Mielke SJ, Alyafeai Z, Salesky E, Raffel C, Dey M, Gallé M, et al. Between words and characters: a brief history of open-vocabulary modeling and tokenization in NLP. *arXiv preprint* 2021:2112.10508.
20. Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez AN, et al. Attention is all you need. *arXiv preprint* 2023:1706.03762v7.
21. Brown TB, Mann B, Ryder N, Subbiah M, Kaplan J, Dhariwal P, et al. Language models are few-shot learners. *arXiv preprint* 2020. 2005.14165.
22. Radford A, Wu J, Child R, Luan D, Amodei D, Sutskever I. Language models are unsupervised multitask learners. *OpenAl blog* 2019;**1**:9.
23. Devlin J, Chang MW, Lee K, Toutanova K. BERT: pre-training of deep bidirectional transformers for language understanding. *ACL Anthology* 2019:4171−86.
24. Raffel C, Shazeer N, Roberts A, Lee K, Narang S, Matena M, et al. Exploring the limits of transfer learning with a unified text-to-text transformer. *J Mach Learn Res* 2023;**21**:1−67.
25. Berti L, Giorgi F, Kasneci G. Emergent abilities in large language models: a survey. *arXiv preprint* 2025:2503.05788.
26. Haque A, Engel J, Teichmann SA, Lönnberg T. A practical guide to single-cell RNA-sequencing for biomedical research and clinical applications. *Genome Med* 2017;**9**:75.
27. Wang Z, Gerstein M, Snyder M. RNA-Seq: a revolutionary tool for transcriptomics. *Nat Rev Genet* 2009;**10**:57−63.
28. Feinberg AP. Epigenomics reveals a functional genome anatomy and a new approach to common disease. *Nat Biotechnol* 2010;**28**:1049−52.
29. Gygi SP, Rochon Y, Franza BR, Aebersold R. Correlation between protein and mRNA abundance in yeast. *Mol Cell Biol* 1999;**19**: 1720−30.
30. Marguerat S, Schmidt A, Codlin S, Chen W, Aebersold R, Bähler J. Quantitative analysis of fission yeast transcriptomes and proteomes in proliferating and quiescent Cells. *Cell* 2012;**151**:671−83.
31. Danzi F, Pacchiana R, Mafficini A, Scupoli MT, Scarpa A, Donadelli M, et al. To metabolomics and beyond: a technological portfolio to investigate cancer metabolism. *Signal Transduct Targeted Ther* 2023;**8**:1−22.
32. Tsang HF, Xue VW, Koh SP, Chiu YM, Ng LPW, Wong SCC. NanoString, a novel digital color-coded barcode technology: current and future applications in molecular diagnostics. *Expert Rev Mol Diagn* 2017;**17**:95−103.
33. Sanger F, Nicklen S, Coulson AR. DNA sequencing with chain-terminating inhibitors. *Proc Natl Acad Sci USA* 1977;**74**:5463−7.
34. Schena M, Shalon D, Davis RW, Brown PO. Quantitative monitoring of gene expression patterns with a complementary DNA microarray. *Science* 1995;**270**:467−70.
35. Ran FA, Hsu PD, Wright J, Agarwala V, Scott DA, Zhang F. Genome engineering using the CRISPR-Cas9 system. *Nat Protoc* 2013;**8**: 2281−308.
36. Higuchi R, Dollinger G, Walsh PS, Griffith R. Simultaneous amplification and detection of specific DNA sequences. *Nat Biotechnol* 1992;**10**:413−7.
37. Gall JG, Pardue ML. Formation and detection of RNA−DNA hybrid molecules in cytological preparations. *Proc Natl Acad Sci USA* 1969; **63**:378−83.
38. Williams CG, Lee HJ, Asatsuma T, Vento-Tormo R, Haque A. An introduction to spatial transcriptomics for biomedical research. *Genome Med* 2022;**14**:68.
39. Ji YR, Zhou ZH, Liu H, Davuluri RV. DNABERT: pre-trained bidirectional encoder representations from transformers model for DNA-language in genome. *Bioinformatics* 2021;**37**:2112−20.
40. Zhou ZH, Ji YR, Li WJ, Dutta P, Davuluri R, Liu H. DNABERT-2: efficient foundation model and benchmark for multi-species genome. *arXiv preprint* 2024:2306.15006v2.
41. Chen JY, Hu ZH, Sun SQ, Tan QX, Wang YX, Yu QZ, et al. Interpretable RNA foundation model from unannotated data for highly accurate RNA structure and function predictions. *arXiv preprint* 2022. 2204.00300.
42. Lin ZM, Akin H, Rao RS, Hie B, Zhu ZK, Lu WT, et al. Evolutionary-scale prediction of atomic-level protein structure with a language model. *Science* 2023;**379**:1123−30.
43. Yang F, Wang WC, Wang F, Fang Y, Tang DY, Huang JZ, et al. scBERT as a large-scale pretrained deep language model for cell type annotation of single-cell RNA-seq data. *Nat Mach Intell* 2022;**4**: 852−66.
44. Cui HT, Wang C, Maan H, Pang K, Luo FN, Duan N, et al. scGPT: toward building a foundation model for single-cell multi-omics using generative AI. *Nat Methods* 2024;**21**:1470−80.
45. Theodoris CV, Xiao L, Chopra A, Chaffin MD, Al Sayed ZR, Hill MC, et al. Transfer learning enables predictions in network biology. *Nature* 2023;**618**:616−24.
46. Wen HZ, Tang WZ, Dai XN, Ding JY, Jin W, Xie YY, et al. CellPLM: pre-training of cell language model beyond single cells. *bioRxiv* 2023. 2023.10.03.560734.
47. Grün D, Kester L, van Oudenaarden A. Validation of noise models for single-cell transcriptomics. *Nat Methods* 2014;**11**:637−40.
48. Li W, Yang F, Wang F, Rong Y, Liu LJ, Wu BZ, et al. scPROTEIN: a versatile deep graph contrastive learning framework for single-cell proteomics embedding. *Nat Methods* 2024;**21**:623−34.
49. Baysoy A, Bai ZL, Satija R, Fan R. The technological landscape and applications of single-cell multi-omics. *Nat Rev Mol Cell Biol* 2023; **24**:695−713.
50. Risso D, Perraudeau F, Gribkova S, Dudoit S, Vert JP. A general and flexible method for signal extraction from single-cell RNA-seq data. *Nat Commun* 2018;**9**:284.
51. Hafemeister C, Satija R. Normalization and variance stabilization of single-cell RNA-seq data using regularized negative binomial regression. *Genome Biol* 2019;**20**:296.
52. Yue TW, Wang YX, Zhang LX, Gu CM, Xue HR, Wang WP, et al. Deep learning for genomics: from early neural nets to modern large language models. *Int J Mol Sci* 2023;**24**:15858.
53. Consens ME, Dufault C, Wainberg M, Forster D, Karimzadeh M, Goodarzi H, et al. To transformers and beyond: large language models for the genome. *arXiv preprint* 2023:2311.07621.
54. Liu JJ, Yang MY, Yu YK, Xu HX, Li K, Zhou XB. Large language models in bioinformatics: applications and perspectives. *arXiv preprint* 2024:2401.04155.
55. Benegas G, Batra SS, Song YS. DNA language models are powerful predictors of genome-wide variant effects. *bioRxiv* 2022. 2022.08.22.504706.
56. Manzoni C, Kia DA, Vandrovcova J, Hardy J, Wood NW, Lewis PA, et al. Genome, transcriptome and proteome: the rise of omics data and their integration in biomedical sciences. *Briefings Bioinf* 2018; **19**:286−302.
57. Al-Amrani S, Al-Jabri Z, Al-Zaabi A, Alshekaili J, Al-Khabori M. Proteomics: concepts and applications in human medicine. *World J Biol Chem* 2021;**12**:57−69.
58. Nguyen E, Poli M, Faizi M, Thomas A, Birch-Sykes C, Wornow M, et al. HyenaDNA: long-range genomic sequence modeling at single nucleotide resolution. *arXiv preprint* 2023:2306.15794.
59. Soylu NN, Sefer E. BERT2OME: prediction of 2′-*O*-methylation modifications from RNA sequence by transformer architecture based on BERT. *IEEE ACM Trans Comput Biol Bioinf* 2023;**20**:2177−89.
60. Zhao YC, Oono K, Takizawa H, Kotera M. GenerRNA: a generative pre-trained language model for *de novo* RNA design. *PLoS One* 2024;**19**:e0310814.
61. Avsec Ż, Latysheva N, Cheng J, Novati G, Taylor KR, Ward T, et al. AlphaGenome: advancing regulatory variant effect prediction with a unified DNA sequence model. *bioRxiv* 2025. 2025.06.25.661532.
62. Jin MY, Xue HC, Wang ZT, Kang BM, Ye RS, Zhou KX, et al. ProLLM: protein chain-of-thoughts enhanced LLM for protein−protein interaction prediction. *arXiv preprint* 2024;**2405**:06649.

63. Alipanahi B, Delong A, Weirauch MT, Frey BJ. Predicting the sequence specificities of DNA- and RNA-binding proteins by deep learning. *Nat Biotechnol* 2015;**33**:831−8.
64. Feingold E, Good P, Guyer M, Kamholz S, Liefer L, Wetterstrand K, et al. The ENCODE (ENCyclopedia of DNA elements) project. *Science* 2004;**306**:636−40.
65. Schiff Y, Kao CH, Gokaslan A, Dao T, Gu A, Kuleshov V. Caduceus: bi-directional equivariant long-range DNA sequence modeling. *arXiv preprint* 2024:2403.03234.
66. Chen W, Tang H, Ye J, Lin H, Chou KC. iRNA-PseU: identifying RNA pseudouridine sites. *Mol Ther Nucleic Acids* 2016;**5**:e332.
67. Zhou P, Shi W, Tian J, Qi ZY, Li BC, Hao HW, et al. *Attention-based bidirectional long short-term memory networks for relation classification*. *2016 short papers*. ACL; 2016. p. 207−12.
68. Zhou J, Troyanskaya OG. Predicting effects of noncoding variants with deep learning-based sequence model. *Nat Methods* 2015;**12**:931−4.
69. The GTEx Consortium, Ardlie KG, Deluca DS, Segrè AV, Sullivan TJ, Young TR, et al. The genotype-tissue expression (GTEx) pilot analysis: multitissue gene regulation in humans. *Science* 2015;**348**: 648−60.
70. The FANTOM Consortium and the RIKEN PMI and CLST (DGT). A promoter-level mammalian expression atlas. *Nature* 2014;**507**: 462−70.
71. Pampari A, Shcherbina A, Kvon EZ, Kosicki M, Nair S, Kundu S, et al. ChromBPNet: bias factorized, base-resolution deep learning models of chromatin accessibility reveal cis-regulatory sequence syntax, transcription factor footprints and regulatory variants. *bioRxiv* 2024. 2024.12.25.630221.
72. Jaganathan K, Panagiotopoulou KS, McRae JF, Darbandi SF, Knowles D, Li YI, et al. Predicting splicing from primary sequence with deep learning. *Cell* 2019;**176**:535−48.e24.
73. Zhou J. Sequence-based modeling of three-dimensional genome architecture from kilobase to chromosome scale. *Nat Genet* 2022;**54**: 725−34.
74. Elnaggar A, Heinzinger M, Dallago C, Rehawi G, Wang Y, Jones L, et al. ProtTrans: toward understanding the language of life through self-supervised learning. *IEEE Trans Pattern Anal Mach Intell* 2022; **44**:7112−27.
75. Hao MS, Gong J, Zeng X, Liu CM, Guo YC, Cheng XY, et al. Large-scale foundation model on single-cell transcriptomics. *Nat Methods* 2024;**21**:1481−91.
76. Adduri AK, Gautam D, Bevilacqua B, Imran A, Shah R, Naghipourfar M, et al. Predicting cellular responses to perturbation across diverse contexts with STATE. *bioRxiv* 2025. 2025.06.26.661135.
77. Van de Sande B, Lee JS, Mutasa-Gottgens E, Naughton B, Bacon W, Manning J, et al. Applications of single-cell RNA sequencing in drug discovery and development. *Nat Rev Drug Discov* 2023;**22**:496−520.
78. Cao JY, O'Day DR, Pliner HA, Kingsley PD, Deng M, Daza RM, et al. A human cell atlas of fetal gene expression. *Science* 2020;**370**: eaba7721.
79. Kalfon J, Samaran J, Peyré G, Cantini L. scPRINT: pre-training on 50 million cells allows robust gene network predictions. *Nat Commun* 2025;**16**:3607.
80. Joseph DB, Henry GH, Malewska A, Reese JC, Mauck RJ, Gahan JC, et al. Single-cell analysis of mouse and human prostate reveals novel fibroblasts with specialized distribution and microenvironment interactions. *J Pathol* 2021;**255**:141−54.
81. Megill C, Martin B, Weaver C, Bell S, Prins L, Badajoz S, et al. Cellxgene: a performant, scalable exploration platform for high dimensional sparse matrices. *bioRxiv* 2021. 2021.04.05.438318.
82. Norman TM, Horlbeck MA, Replogle JM, Ge AY, Xu A, Jost M, et al. Exploring genetic interaction manifolds constructed from rich single-cell phenotypes. *Science* 2019;**365**:786−93.
83. Lamb J, Crawford ED, Peck D, Modell JW, Blat IC, Wrobel MJ, et al. The Connectivity Map: using gene-expression signatures to connect small molecules, genes, and disease. *Science* 2006;**313**:1929−35.
84. Zhong FS, Wu XL, Yang RR, Li XT, Wang DY, Fu ZY, et al. Drug target inference by mining transcriptional data using a novel graph convolutional network framework. *Protein Cell* 2022;**13**: 281−301.
85. Ni SK, Kong XT, Zhang YY, Chen ZY, Wang ZK, Fu ZY, et al. Identifying compound−protein interactions with knowledge graph embedding of perturbation transcriptomics. *Cell Genom* 2024;**4**:100655.
86. Chen YQ, Zou J. Simple and effective embedding model for single-cell biology built from ChatGPT. *Nat Biomed Eng* 2025;**9**:483−93.
87. Unger FT, Witte I, David KA. Prediction of individual response to anticancer therapy: historical and future perspectives. *Cell Mol Life Sci* 2015;**72**:729−57.
88. Chawla S, Rockstroh A, Lehman M, Ratther E, Jain A, Anand A, et al. Gene expression based inference of cancer drug sensitivity. *Nat Commun* 2022;**13**:5680.
89. Vasudevan S, Adejumobi IA, Alkhatib H, Roy Chowdhury S, Stefansky S, Rubinstein AM, et al. Drug-induced resistance and phenotypic switch in triple-negative breast cancer can be controlled *via* resolution and targeting of individualized signaling signatures. *Cancers* 2021;**13**:5009.
90. Zheng ZT, Chen JY, Chen XJ, Huang L, Xie WD, Lin QZ, et al. Enabling single-cell drug response annotations from bulk RNA-seq using SCAD. *Adv Sci* 2023;**10**:2204113.
91. Bontonou M, Haget A, Boulougouri M, Audit B, Borgnat P, Arbona JM. A comparative analysis of gene expression profiling by statistical and machine learning approaches. *Bioinform Adv* 2024;**5**: vbae199.
92. Tong XC, Qu N, Kong XT, Ni SK, Zhou JY, Wang K, et al. Deep representation learning of chemical-induced transcriptional profile for phenotype-based drug discovery. *Nat Commun* 2024;**15**:5378.
93. Zhu J, Wang JX, Wang X, Gao MJ, Guo BB, Gao MM, et al. Prediction of drug efficacy from transcriptional profiles with deep learning. *Nat Biotechnol* 2021;**39**:1444−52.
94. Pham TH, Qiu Y, Zeng JC, Xie L, Zhang P. A deep learning framework for high-throughput mechanism-driven phenotype compound screening and its application to COVID-19 drug repurposing. *Nat Mach Intell* 2021;**3**:247−57.
95. Yang XD, Liu GL, Feng GH, Bu DC, Wang PF, Jiang J, et al. GeneCompass: deciphering universal gene regulatory mechanisms with a knowledge-informed cross-species foundation model. *Cell Res* 2024;**34**:830−45.
96. Bunne C, Roohani Y, Rosen Y, Gupta A, Zhang XK, Roed M, et al. How to build the virtual cell with artificial intelligence: priorities and opportunities. *Cell* 2024;**187**:7045−63.
97. Roohani YH, Hua TJ, Tung PY, Bounds LR, Yu FB, Dobin A, et al. Virtual cell challenge: toward a turing test for the virtual cell. *Cell* 2025;**188**:3370−4.
98. Zhang J, Ubas AA, De Borja R, Svensson V, Thomas N, Thakar N, et al. *Tahoe-100M*: a giga-scale single-cell perturbation atlas for context-dependent gene function and cellular modeling. *bioRxiv* 2025:2025.02.20.639398.
99. Büttner M, Ostner J, Müller CL, Theis FJ, Schubert B. scCODA is a Bayesian model for compositional single-cell data analysis. *Nat Commun* 2021;**12**:6876.
100. Szałata A, Hrovatin K, Becker S, Tejada-Lapuerta A, Cui HT, Wang B, et al. Transformers in single-cell omics: a review and new perspectives. *Nat Methods* 2024;**21**:1430−43.
101. Sun FY, Li HY, Sun DQ, Fu SL, Gu L, Shao X, et al. Single-cell omics: experimental workflow, data analyses and applications. *Sci China Life Sci* 2025;**68**:5.
102. Gulati GS, D'Silva JP, Liu YH, Wang LH, Newman AM. Profiling cell identity and tissue architecture with single-cell and spatial transcriptomics. *Nat Rev Mol Cell Biol* 2025;**26**:11−31.
103. Hemstrom W, Grummer JA, Luikart G, Christie MR. Next-generation data filtering in the genomics era. *Nat Rev Genet* 2024;**25**:750−67.
104. Seydel C. Single-cell metabolomics hits its stride. *Nat Methods* 2021; **18**:1452−6.

105. Ma Q, Jiang Y, Cheng H, Xu D. Harnessing the deep learning power of foundation models in single-cell omics. *Nat Rev Mol Cell Biol* 2024;**25**:593−4.
106. Chefer H, Gur S, Wolf L. Transformer interpretability beyond attention visualization. *Proc IEEE Comput Soc Conf Comput Vis Pattern Recognit* 2021:782−91.
107. Liu X, Boylan J, Bouiller T, Solovyeva E, Ataman B, Hoersch S, et al. Evaluating foundation models for *in-silico* perturbation. *bioRxiv* 2025. 2025.05.11.653338.
108. Kumar PS. Microbiomics: were we all wrong before?. *Periodontol* 2000 2021;**85**:8−11.
109. Wang KC, Chang HY. Epigenomics: technologies and applications. *Circ Res* 2018;**122**:1191−9.
110. Dezem FS, Arjumand W, DuBose H, Morosini NS, Plummer J. Spatially resolved single-cell omics: methods, challenges, and future perspectives. *Annu Rev Biomed Data Sci* 2024;**7**:131−53.
111. Vandereyken K, Sifrim A, Thienpont B, Voet T. Methods and applications for single-cell and spatial multi-omics. *Nat Rev Genet* 2023;**24**:494−515.
112. Heumos L, Schaar AC, Lance C, Litinetskaya A, Drost F, Zappia L, et al. Best practices for single-cell analysis across modalities. *Nat Rev Genet* 2023;**24**:550−72.
113. Choi H, Park J, Kim S, Kim J, Lee D, Bae S, et al. CELLama: foundation model for single cell and spatial transcriptomics by cell embedding leveraging language model abilities. *bioRxiv* 2024. 2024.05.08.593094.