Computational and Structural Biotechnology Journal 27 (2025) 265–277

Contents lists available at ScienceDirect

Computational and Structural Biotechnology Journal

journal homepage: www.elsevier.com/locate/csbj

Review Article

AI-driven multi-omics integration for multi-scale predictive modeling of genotype-environment-phenotype relationships

You Wu [a,*], Lei Xie [a,b,c,d,*]

[a] Ph.D. Program in Computer Science, The Graduate Center, The City University of New York, New York, NY, USA
[b] Ph.D. Program in Biology and Biochemistry, The Graduate Center, The City University of New York, New York, NY, USA
[c] Department of Computer Science, Hunter College, The City University of New York, New York, NY, USA
[d] Helen & Robert Appel Alzheimer's Disease Research Institute, Feil Family Brain & Mind Research Institute, Weill Cornell Medicine, Cornell University, New York, NY, USA

ARTICLE INFO

Keywords:
Machine learning
Deep learning
Single cell
Omics data
Drug discovery
Precision medicine
Complex disease

ABSTRACT

Despite the wealth of single-cell multi-omics data, it remains challenging to predict the consequences of novel genetic and chemical perturbations in the human body. It requires knowledge of molecular interactions at all biological levels, encompassing disease models and humans. Current machine learning methods primarily establish statistical correlations between genotypes and phenotypes but struggle to identify physiologically significant causal factors, limiting their predictive power. Key challenges in predictive modeling include scarcity of labeled data, generalization across different domains, and disentangling causation from correlation. In light of recent advances in multi-omics data integration, we propose a new artificial intelligence (AI)-powered biology-inspired multi-scale modeling framework to tackle these issues. This framework will integrate multi-omics data across biological levels, organism hierarchies, and species to predict genotype-environment-phenotype relationships under various conditions. AI models inspired by biology may identify novel molecular targets, biomarkers, pharmaceutical agents, and personalized medicines for presently unmet medical needs.

Contents

* Corresponding authors at: Ph.D. Program in Computer Science, The Graduate Center, The City University of New York, New York, NY, USA.
E-mail addresses: ywu1@gradcenter.cuny.edu (Y. Wu), lei.xie@hunter.cuny.edu (L. Xie).

https://doi.org/10.1016/j.csbj.2024.12.030
Received 31 October 2024; Received in revised form 22 December 2024; Accepted 26 December 2024

Available online 2 January 2025

1. Introduction

A fundamental challenge in biology is predicting phenotypes, considering the complex interactions between genotypes and environmental influences and perturbations [1]. Organismal phenotypes encompass observable physical characteristics (e.g., eye color), behavioral patterns (e.g., memory), physiological functions (e.g., blood pressure), and clinical manifestations (e.g., pain). However, an organism's phenotype does not directly emerge from its genotype. Several intermediate phenotypes, known as endophenotypes [2], delineate molecular attributes at an intermediate level of organization, complexity, or scale between the molecular/genetic level and the organismal phenotype. Endophenotypes typically include RNA expression, protein expression and post-translational modifications, metabolite concentrations, and similar molecular markers. To establish linkages between genotype, environment, and phenotype, it is essential to utilize endophenotypes as a means of connecting an organism's genetic foundation to its observable traits.

The latest advances in sequencing and high-throughput technologies have generated vast amounts of multi-omics data, including genomics, epigenomics, transcriptomics, proteomics, metabolomics, lipidomics, glycomics, cytomics/cellomics, microbiomics, metagenomics, radiomics, interactomics, and chemical genomics [3]. With the exception of genomics and epigenomics data that characterize genotypes, and microbiomics, metagenomics, and chemical genomics data that provide environmental information, most omics data reveals the molecular landscape of distinct endophenotypes at various levels. These omics data are crucial in linking genetic information to phenotypic outcomes and predicting phenotype responses to environments. Ultimately, endophenotypes can serve as biomarkers and offer specific targets linked to disease causes, thereby facilitating the development of effective and safe therapeutic interventions.

While each omics type provides a unique perspective on molecular processes within cells, tissues, or organisms, it is essential to integrate all layers of omics data to fully comprehend the complexity and interdependencies of biological systems [4]. First, rooted in the central dogma of molecular biology, it is necessary to connect multiple levels of omics data—from DNAs and RNAs to proteins and phenotypic outcomes—to understand how genetic information is converted into functional molecules and ultimately, phenotypes. Second, integrating data across multiple omics levels enables the identification of key regulatory elements that act as critical control points in cellular pathways, revealing the complex interactions and feedback loops governing cellular processes. Finally, individual omics datasets provide only partial information about a biological system. Their integration will enhance the predictive power of computational models aimed at establishing connections between genetics and phenotype.

The human body comprises a diverse array of cell types. These cells are organized in a hierarchical structure: cells combine to form tissues, tissues form organs, and organs collaborate to create a functional organism. Cells communicate through chemical signals such as hormones and neurotransmitters. Recent advances in single-cell and spatial omics techniques now make it possible to observe and quantify heterogeneous cellular processes and cell-cell communications across an organism's hierarchical levels at single-cell resolution [5–8]. Spatial single-cell omics data will be crucial in linking molecular events to organism phenotypes [9–12]. Therefore, it is critical to integrate omics data across biological scales—from cell to tissue to organ to organism.

Beyond integrating omics data across biological levels and organismal scales, it is imperative to also integrate data across different species [13–15]. Omics studies in model systems are essential for advancing biological understanding. Model organisms have long been instrumental in investigating gene functions, regulatory mechanisms, cellular processes, tissue formation, organ development, and genetic factors influencing complex behaviors. Genetically engineered models are invaluable tools for understanding molecular disease mechanisms, evaluating potential treatments, and assessing therapeutic interventions' safety and efficacy. Recent advances in functional genomics, such as CRISPR-Cas9 and perturb-seq, now enable large-scale assessments of gene functions and dissection of gene regulatory networks using model organisms. As multi-omics data from model organisms becomes increasingly accessible, innovative methods are needed to transfer this knowledge to human contexts, thereby advancing fundamental and translational biomedical sciences.

Cross-level, cross-scale, and cross-species multi-omics data integration, along with predictive modeling of genotype-environment-phenotype relationships, will not only generate new insights into life's fundamental principles but also drive the identification of novel molecular targets, biomarkers, pharmaceutical agents, and personalized medicines for unmet medical needs. The target-based drug discovery and development approach, which emerged following the human genome revolution and now dominates the pharmaceutical industry, is widely recognized as time-consuming, expensive, and often unproductive. A recent survey indicates that over 90% of approved medications originated from phenotype-based drug discovery and development [16]. Perturbation functional omics profiling provides a quantitative, mechanistic, and high-throughput phenotype readout for compound screening, thereby significantly enhancing the potential of phenotype-driven drug discovery [17].

In summary, elucidating the genetic and molecular foundations of complex human traits and disorders, and predicting organismal phenotypes under diverse genotypic and environmental interactions, requires integrating multi-omics data across modalities, biological levels, and species (Fig. 1). This review paper first summarizes available perturbation omics data and examines recent machine learning advances, with a focus on deep learning techniques for multi-omics data integration. Due to the exponential growth of deep learning literature, the paper coverage is necessarily selective but representative of current field trends. The methodology for paper selection is detailed in the supplemental material. The paper then critically examines limitations in current methodologies and proposes two solutions to address existing challenges: biology-inspired, AI-driven framework for multi-omics integration and multi-scale predictive modeling. This framework aims to predict human phenotypic responses to unprecedented perturbations. By bridging computational and biological approaches, it holds promise for illuminating fundamental life principles and discovering new molecular targets, biomarkers, pharmacological agents, and personalized therapies for currently intractable diseases.

2. Data resources to support the predictive modeling

Predictive modeling of phenotypes from genotypes under perturbations needs labeled data. Recent advances in perturbation omics techniques have generated extensive datasets by deliberately manipulating biological systems using methods like CRISPR gene editing, RNA interference (RNAi), and small molecule treatments. These interventions are

Table 1
Perturbation data resource, *linked data resource.

Source	Perturbation Type	Molecular Profiling	Assay Readout	Datasets Included
TCGA [18]	Drug	Genomic, transcriptomic, epigenomic, proteomic	Clinical and survival data	33 tissue types
LINCS Data Portal [19,20]	Drug, CRISPR-Cas9, ShRNA	Perturbed transcriptomic, proteomic	Transcriptomic, proteomic, kinase binding, cell viability, cell growth inhibition, apoptosis, morphology	LINCS 1000, LINCS proteomic, ChEMBL*, Tox21*, Cell Painting morphological profiling assay*
DepMap [21]	CRISPR-Cas9, RNAi screen, drug	Genomic, transcriptomic, proteomic	Perturbed genomic, transcriptomic, proteomic, drug sensitivity, drug response	CCLE, GDSC, CTRP
scPerturb [28]	CRISPR-cas9, CRISPRi, CRISPRa, TCR stim, cytokines, drug	Single-cell RNA sequencing (scRNA-seq), proteomic, epigenomic	Perturbed scRNA-seq, proteomic, chromatin accessibility	Sci-plex, 44 public single-cell perturbation datasets
PharmacoDB [30]	Drug	Genomic, transcriptomic, proteomic	Drug sensitivity, drug response	CCLE, GDSC, NCI-60, PRISM, FIMM, GTRP, GRAY, gCSI
ProteomicsDB [31]	Drug	Proteomics, transcriptomics	Posttranslational modifications (PTMs), perturbed proteomics, phenomics	DecryptE, DecryptM, GeneCards*, UniProt*, OmniPathDB* and Gene Information eXtension (GIX)*

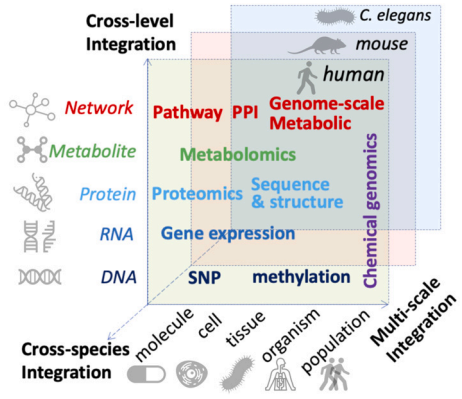

Fig. 1. Illustration of the three axes of multi-omics data integration: (1) Cross-level Integration, representing the integration of omics data across molecular layers (DNA, RNA, protein, metabolites, and networks) to enable modeling molecular interplays; (2) Cross-species Integration, capturing understandings across model organisms (e.g., *C. elegans*, mouse, and human) to improve translational research; and (3) Multi-scale Integration, spanning molecular to population-level data, across scales from single molecules to whole organisms and populations to connect molecular phenotypes to organismal phenotypes.

then systematically measured using various molecular profiling techniques, which provide detailed 'readouts' or labels that capture the system's response to each perturbation. Although these data are highly biased to certain biological conditions (cell types, diseases, etc.) and perturbation types, they are the starting point for machine learning. Several representative data sets are listed in Table 1 and summarized below.

TCGA (The Cancer Genome Atlas) [18] is a comprehensive resource that has molecularly characterized thousands of primary cancers and matched normal samples across diverse cancer types. By integrating data on genetic mutations, gene expression, methylation, and protein profiles, TCGA provides a robust framework for understanding the molecular mechanisms of cancer, aiding in the identification of biomarkers and therapeutic targets.

LINCS (Library of Integrated Network-based Cellular Signatures), part of the Connectivity Map (CMAP) project, aims to elucidate cellular responses to various perturbations, including small molecule treatments and genetic modifications. Using high-throughput techniques, LINCS generates extensive datasets on gene expression and protein levels. The LINCS Data Portal provides access to data from multiple sources [19,20].

DepMap (Dependency Map) [21] is a pioneering initiative that systematically identifies the genetic and chemical dependencies of cancer cells. Through high-throughput CRISPR-Cas9, RNAi, and chemical screens, it maps essential genes and pathways critical for cancer cell survival. This comprehensive resource integrates data from several platforms, including the Cancer Cell Line Encyclopedia (CCLE) [22,23], Genomics of Drug Sensitivity in Cancer (GDSC) [24,25], and the Cancer Therapeutics Response Portal (CTRP) [26,27].

scPerturb [28] focuses on single-cell perturbation studies, providing detailed insights into how individual cells respond to genetic modifications and other perturbations. By employing advanced single-cell RNA sequencing techniques, scPerturb captures the heterogeneity and dynamic responses of single cells, enabling researchers to unravel gene functions, regulatory networks, and the impact of genetic changes at unprecedented resolution. It offers an integrative dataset from 44 published works, incorporating various methods and Sci-Plex [29].

PharmacoDB [30] is an integrative database that consolidates pharmacogenomic data from multiple high-throughput drug screening studies. It provides a platform for exploring drug responses across diverse cancer cell lines, facilitating the identification of drug efficacy, resistance mechanisms, and potential biomarkers. PharmacoDB supports personalized medicine by linking molecular profiles with drug sensitivity data, advancing the development of tailored therapeutic strategies.

ProteomicsDB [31] is a comprehensive database that integrates human proteomic data from numerous high-throughput proteomics experiments. It provides detailed information on protein expression, post-translational modifications (PTMs), and protein interactions. The platform also highlights recent studies on decrypting the molecular basis of cellular drug phenotypes (DecryptE [32]) and analyzing drug actions and protein modifications through dose- and time-resolved proteomics (DecryptM [33]).

3. State-of-the-art of machine learning methods for multi-omics data integration and predictive modeling

3.1. Unsupervised learning

One of the major technical challenges in multi-omics data integration is handling data distribution shifts. These shifts primarily arise from two sources: technical confounders, such as batch effects, and biological confounders (e.g., sex, age, disease state, etc.). Traditional statistical methods have laid the foundation for multi-omics data integration. These methods include a variety of techniques, such as correlation-based analysis (e.g., BindSC [59], Seurat v3 [60], Scanorama [61], and MaxFuse [62]), matrix factorization (e.g., iNMF [63] and LIGER [64]), Bayesian-based methods (e.g., MOFA+ [65]), nearest neighbor-based approaches (e.g., fastMNN [66] and Seurat v4 [67]), and dictionary learning (e.g., Seurat v5 [68]).

Another traditional yet powerful class of approaches is kernel methods, which project data into high-dimensional feature spaces to uncover complex, non-linear relationships. For instance, multi-kernel linear mixed models with adaptive lasso (MKpLMM) have demonstrated

Table 2
Representative state-of-the-art computational methods for multi-omics data integration toward predictive modeling of genotype-environment-phenotype relationships.

Learning	Methods	Representative Papers	Modality	Notes
Unsupervised	Autoencoder	ScVI [34]	scRNA-seq	Effective in removing batch effects; however, it is constrained to analyzing only single modality data
		scANVI [35]	scRNA-seq	Facilitates label transfer with uncertainty measures in semi-supervised learning; limited to a single modality
		TotalVI [36]	scRNA-seq, surface protein	Learns a joint probabilistic representation of both RNA and proteins; but requires paired measurements and does not align domains across different experiments
		Cobolt [37]	mRNA-seq, scRNA-seq, ATAC-seq, scATAC-seq	Offers guided multimodal integration for paired RNA-seq and ATAC-seq data, but the assumptions of a multinomial distribution might ignore the biological context of different modalities
		MultiVI [38]	scRNA-seq, scATAC-seq, surface protein	Guides multimodal integration accounting for modality-specific noise; uses a symmetric approach for joint representation, though affected by data sparsity
		scMVP [39]	scRNA-seq, scATAC-seq	Provides non-symmetric multimodal integration with multi-head attention and cycle-GAN; but requires paired sample data
		GLUE [40]	scRNA-seq, scATAC-seq	Triple-omics integration while simultaneously inferring regulatory interactions; adversarial training may lack stability
		Biolord [41]	scRNA-seq, drug, dosage, cell line	Encodes cellular identity attributes separately for better representation; needs exploration of unknown attributes to improve generalizability
		ChemCPA [42]	Bulk & scRNA-seq, drug, dosage	Incorporates compound structure and bulk RNA-seq data with adversarial training to adapt to single-cell data; effective for unseen compounds but needs evaluation on unseen cell lines
	Transformer	scGPT [43]	scRNA-seq, scATAC-seq, surface protein, Perturb-seq	Foundation model trained on over 10M cells, capable of learning cell-specific information; requires paired data and limited reliability in zero-shot settings
		GeneCompass [14]	Cross-species, scRNA-seq, perturb-seq, LINCS1000	Foundation model trained over 12OM cells cross-species incorporating prior knowledge; confined to transcriptomic data
		Prophet [44]	Cell state, treatment, phenotypic readout	Strong generalizability to unseen cell states and interventions; limited to non-true OOD scenarios, fixed representations risk error propagation
		SATURN [45]	Cross-species, scRNA-seq, protein sequence	Enables cross-species analysis by merging protein language models with scRNA data; challenges exist due to the absence of direct orthologs and it requires paired data
	Other techniques	scCLIP [46]	scRNA-seq, scATAC-seq	Employs contrastive learning for multimodal single-cell data; paired sample data is mandatory
		MatchCLOT [47]	scRNA-seq, scATAC-seq, surface protein abundance	Combines contrastive learning with optimal transport; reliant on paired sample data
Supervised	Multimodal	Yang et al. [48]	Image, RNA-seq, ATAC-seq, Hi-c	Integrates various data types for cancer models; each model is specific to one type of cancer and requires paired data
		Faisal et al. [49]	H&E WSIs and molecular profile features	Correlates histopathological images with molecular profiles; demands paired data and is specific to individual cancer models
		DSIR [50]	DNA methylation, mRNA and miRNA expression	Utilizes a similarity matrix for cancer subtyping; dependent on paired data and tailored to individual cancers
		DLSF [51]	DNA methylation, mRNA and miRNA expression	Applies a cycle autoencoder to extract a consistent sample manifold at the multi-omics level; also requires paired data for each cancer model
		MOMA [52]	DNA methylation, mRNA and miRNA expression	Processes genes and methylation data using a geometric approach; models need to be individually trained for each cancer type and paired data is needed
	Knowledge graph and other techniques	Lee et al. [53]	Bulk & scRNA-seq	Develops patient-specific cell-cell communication networks to predict immune checkpoint inhibitors efficacy and uncover key pathways; yet, it simplifies complex network relationships
		BioBridge [54]	Protein, molecule, disease, biological process, molecular function, and cellular component	Leverages knowledge graphs to transition between unimodal foundations without fine-tuning; lacks quantitative evidence for molecular generation tasks
		One for all [55]	Literature category description, molecule property description, relation type description	Constructs text-attributed graphs for diverse cross-domain associations; it does not meet the state-of-the-art performance for individual tasks
		GEARS [56]	Gene-gene interaction, scRNA-seq	Integrates GNN with a gene-gene interaction knowledge graph; limited to the same cell type and experimental condition, with confounding factors from combinatorial perturbational data
		TxGNN [57]	Biological process, protein, disease, phenotype, anatomy, molecular function, drug, cellular component, pathway, exposure	Identifies therapeutic candidates in a zero-shot setting and includes expert-validated model interpretation; noisy and incomplete medical knowledge graphs limit predictive power, and contradictory relationships between entities are overlooked
		PinnacleAI [58]	Protein-protein interaction, scRNA across various cell types and tissues	Generates contextualized representations across protein, cell type, and tissue hierarchies with multi-level attention mechanisms; limited to cell types or tissues in the training set, making it less effective for predicting diseases not represented in healthy subjects

their efficacy in high-dimensional multi-omics prediction tasks. These methods can identify predictive regions as well as predictive layers of omics data through a data-driven approach [69]. Kernel fusion techniques, which combine multiple omics layers using kernel-based representations, have proven effective in tasks such as disease subtype classification [70]. Additionally, pathway-induced multiple kernel learning (PIMKL) leverages biological pathways to construct specialized kernels for robust predictive modeling [71]. Comprehensive reviews further highlight the versatility and adaptability of kernel methods in addressing challenges associated with omics data integration [72].

Our focus, however, is on deep representation learning methods, which have shown great promise in addressing these challenges (Table 2). Representative techniques include autoencoders, transformers, and contrastive learning. The strength of these methods lies in their ability to operate without requiring labeled phenotypic data, which is often scarce and infeasible to obtain.

3.1.1. Autoencoder

Deep generative models, particularly Variational Autoencoders (VAEs), are at the forefront of analyzing complex, high-dimensional single-cell sequencing data. VAEs employ an encoder to interpret input data and a decoder to reconstruct it, learning a latent distribution. The objective that it optimizes is to mirror the input while minimizing the Kullback-Leibler divergence between the latent embedding's prior and posterior distributions.

scVI [34] models gene expression in scRNA-seq data using VAE with a zero-inflated negative binomial distribution, conditioned on batch annotations and two unobserved variables: a cell-specific scaling factor and a latent biological variable. Neural networks map these latent variables to the distribution parameters, producing batch-corrected, normalized transcript estimates for differential expression and imputation. A separate neural network, trained via variational inference and stochastic optimization, approximates the posterior distribution of latent variables, ensuring scalable and accurate analysis of single-cell RNA-seq data.

The same group further developed scANVI [35], which integrates semi-supervised learning with cell type annotations. It can be useful for transfer labels while measuring uncertainty, especially when dealing with complex label structures such as hierarchical cell types. However, both models are limited on RNA-seq data as a single modality.

TotalVI [36] took advantage of the CITE-seq technique, which can simultaneously measure the abundance of the proteins on the cell surface, to provide the opportunity for multifaceted analysis of both RNA-seq and the functional information in proteins. It uses VAEs to learn a joint probabilistic representation of the paired measurements that counts for batch effects for both modalities. The RNA modeling strategy is similar to scVI [34]. The protein modeling explicitly has modality-specific technical factors such as a protein background, which enable a denoised view of data. However this method requires paired measured samples, nor there is domain alignment consideration.

More recent tool Cobolt [37] introduces a symmetric multi-modal VAE network for multi-omics data integration with a Product of Experts model (PoE) model [73]. PoE combines the variational posteriors of the multiple modalities (the experts) by taking their product and normalizing the result. It was trained on paired multi-omics data to guide the integration of unpaired data, resulting in a joint representation of single-cell RNA-seq and ATAC-seq datasets, which can be beneficial for various downstream tasks. Despite its guidance on the unpaired datasets, this method assumed a multinomial distribution for both modalities which may cause potential information loss.

In contrast, MultiVI [38] employs a modality-specific noise system suited to both gene expression and chromatin accessibility, with negative binomial distribution and Bernoulli distribution respectively. In contrast to Cobolt's PoE technique, MultiVI utilizes a distributional mean and penalization strategy for a more optimized integration of latent embeddings. Moreover, its ability to incorporate cell surface protein abundance broadens its scope, allowing for a richer understanding of cellular properties.

The strengths of both MultiVI and Cobolt, which implemented symmetric multimodal VAE for joint modality representations, are tempered by the challenges of extreme sparsity and random noise in the datasets. These factors can confound the biological variance, posing obstacles to downstream analysis and scalability of the model. Addressing this, scMVP [39], employs a non-symmetric framework that enables the construction of a unified latent space for scRNA-seq and scATAC-seq data. This is achieved via a clustering consistency-enforced multi-view VAE, which is further enhanced by multi-head self-attention mechanisms and a cycle-GAN module, thereby increasing the robustness across both modalities. However, it again requires simultaneous multi-modality measurements with individual cells to function effectively.

To address the challenge of information loss when integrating data across different modalities, GLUE [40] employs a modality-specific graph VAE to refine the feature transformation process by modeling regulatory relationships between chromatin regions and genes. It learns not only local but also global information. With a scalable adversarial alignment, GLUE also enables the integration of three modalities such as gene expression, chromatin accessibility, and DNA methylations.

Biolord [41] is a deep generative method designed to predict cellular responses to unseen drugs and genetic perturbations. It uses an autoencoder to separately encode multiple attributes of cellular identity, along with a single encoding for unknown attributes. This setup defines a decomposed latent space, serving as the input for the generative module to provide measurement predictions. The authors claim this design disentangles the representation with respect to known attributes. However, further exploration of the representation of unknown attributes would enhance the model's generalizability.

Hetzel et al. introduced ChemCPA [42] a model that incorporates knowledge about compound structures and transfers bulk RNA-seq data into both identical and different gene sets between source (bulk) and target (single-cell) datasets. It uses an encoder-decoder architecture with adversarial training, allowing the model to disentangle representations of various attributes and study the effects on specific sources. Although the model was evaluated on unseen compounds, it would be more interesting if it could also work on unseen cell lines.

3.1.2. Transformer

In research areas such as natural language processing (NLP) and computer vision (CV), Transformer as highlighted by the attention mechanism has gained significant attention in recent years, as evidenced by its successful deployment in foundation models. Pioneering models such as BERT [74], GPT [75,76], PaLM [77,78], and LLaMA [79] have set benchmarks in NLP as well as DALL-E [80] in CV have made significant contributions to various downstream tasks. In a biological context, similar to how words construct a sentence, genes construct cells. Analogous to how natural language acts as a foundational layer for interpreting human behavior, the transcriptome similarly serves as a fundamental layer for unraveling the intricacies of gene regulatory mechanisms in biology. Studies have utilized single-cell transcriptomic data to construct pre-trained foundation models, such as scGPT [43], Genefomer [81] and scFoundation [82]. The representative work scGPT constructed the first foundation model through pretraining on over 10 million cells with a 12-layered transformer architecture. It also supports multiple omics data integration from paired data sources. The utilization of the self-attention approach over genes enables the encoding of gene-gene interaction, and the cell conditional tokenization also allows the model to learn cell-specific information, such as different batches and sequencing modalities. However, this technique is constrained by paired data, and exhibits limited reliability in zero-shot settings [83].

While foundation models have achieved notable successes in a variety of downstream tasks, their potential has not yet been leveraged for cross-species data integration. However, the conserved nature of gene regulatory mechanisms across different species presents an out-

standing opportunity to delve into the complexities of gene regulation through such integrative analysis. Bridging the cross-species analytical gap, GeneCompass [14] emerges as an innovative foundation model, extensively pre-trained on a vast dataset comprising over 120 million single-cell transcriptomes from human and mouse origins. It integrates gene IDs, expression values, and prior knowledge together as gene tokens, implementing a 12-layer transformer model for encoding. It also facilitates a variety of downstream tasks through supervised learning, encompassing gene regulatory network elucidation, predictions of drug effects, gene dosage implications, and cellular responses to perturbations. However, GeneCompass is limited to transcriptomics data.

The same group recently developed another method, Prophet [44], for cellular phenotype prediction, which integrates cell states, treatments, and phenotypic readouts. Each component is tokenized into a joint embedding space using different feature encoding strategies, and a transformer with a regression head is applied to learn the relationships between these components. While the authors demonstrated the model's generalizability to unseen cell states and interventions, it does not fully address an out-of-distribution (OOD) scenario where the distribution between training and testing data is significantly different. Additionally, the representations of cell states and interventions are fixed, meaning potential errors could propagate into the training process and distort the predictions.

3.1.3. Other techniques (contrastive learning etc.)

SATURN [45] stands out as the first model that combines protein embeddings, generated using large protein language model ESM2 [84], with gene expression from scRNA-seq datasets. Overcoming the challenges of absent direct one-to-one orthologs, it couples protein embeddings with gene expression, employing soft clustering to form 'macrogene' groups. This approach allows the model to learn universal cell embeddings that bridge differences between individual single-cell experiments even when they have different genes. It combines training with conditional autoencoders with ZINB loss inspired by Lopez et al. [34], and other learning metrics by forcing the different cells within the same dataset far apart using weakly supervised learning and similar cells across the dataset closer to each other in an unsupervised manner. But it requires paired information.

scCLIP [46] introduces a novel application of transformers to scATAC-seq data, drawing inspiration from the contrastive learning principles of CLIP [85], it trains a pair of transformer-based encoders on multimodal single-cell data, utilizing a contrastive loss function for optimization. The result is scCLIP's adeptness at integrating multimodal data into a singular, unified embedding space, with the scalability to accommodate extensive tissue and organismal data from large-scale atlas projects.

Recent applications of optimal transport (OT) in single-cell data analysis have enabled the identification of cellular dynamics and the alignment of multi-omics datasets. MatchCLOT [47] leverages these advancements by training two modality-specific encoders to project single-cell multimodal measurements onto a unified latent space. A novel OT algorithm is then employed for the soft-matching of cells between modalities, using batch labels to narrow the search space and mitigate distribution shifts.

3.2. Supervised learning

Supervised learning techniques, which combine diverse data types like imaging, genomic, and transcriptomic data to improve prediction tasks, have gained considerable attention in recent years. These methods often utilize modality-specific networks (e.g., convolutional neural networks for image data, fully connected networks for sequencing data) to capture local features, while joint representations are learned through shared latent spaces that promote cross-modal integration. Despite their strengths, these techniques typically require paired datasets from multiple modalities, which can limit their scalability and generalizability. Here, we introduce two major approaches: multi-modal supervised learning and knowledge graph.

3.2.1. Multi-modal supervised learning

Yang et al. [48] propose a method using autoencoder across different modalities to achieve integration, each modality is encoded using a local network, such as a convolutional network for image data, fully connected network for sequence data (RNA-seq and ATAC-seq), graph convolutional network for Hi-C. The joint representations are learned from the shared latent space, facilitating the translation between different modalities via a combination of encoders and decoders.

Faisal et al. [49] adopt a deep learning-based multimodal fusion algorithm to integrate H&E whole slide images (WSIs) and molecular profile features, including Copy-Number of Variation (CNV), RNA-seq, and Mutation Status (MUT). Their method is particularly rigorous for its comprehensive application in survival prediction and patient risk stratification, enhanced by a focus on interpretability through the analysis of feature importance and gene attributions.

Deep Subspace Integration Representation (DSIR) [50] represents another technique for multi-modality integration, utilizing deep subspace learning to simultaneously learn the local and global structures. By constructing a consensus similarity matrix, DSIR finetunes its model for cancer subtype identification through spectral clustering.

Similarly, DLSF [51] also obtains the self-representation coefficient matrix for disease subtype identification, what it differs from DSIR is the exploration of the shared global similarity structure, because DLSF uses cycle autoencoders with a shared self-expressive layer to adaptively extract a consistent sample manifold a multi-omics level.

Moreover, a geometrical approach Module-based Omics Data Integration MOMA [52] vectorizes genes and modules, using the vector sum of genes within a module to represent it. The incorporation of an attention mechanism as a mediator allows the model to identify the most related modules among multiple omics data types, by training with various tasks of predicting phenotypes.

For all the multi-modal techniques mentioned above, despite their potential for cross-modal integration, those approaches require paired data from the various modalities and are tailored to individual cancer types, limiting their generalizability.

3.2.2. Knowledge graph and other techniques

Graph (network) representation has been widely applied in systems biology to represent biological organizations and interactions [86]. It is successful in integrating diverse types of biological and chemical data for representing genotype-environment-phenotype relationships [87]. Compared with multi-modal supervised learning, graph learning directly encodes complex interactions between entities and captures semantic relationships underlying data. This allows for the seamless integration of information from diverse sources, the deduction of new information based on existing knowledge, and a deeper understanding of context and interconnections between entities.

Lee et al. [53] propose a machine learning model to predict cancer response to immune checkpoint inhibitors (ICIs). The network is constructed on cell-cell communication with cell types as nodes and communication strength as edges, which is deconvoluted from the patient's bulk tumor transcriptomics data. The model can also identify key communication pathways that are consistent with single-cell level information. However, the graph is shallowly designed and more sophisticated deep learning models could be utilized to reveal complex relationships.

BioBridge [54] is representative of the integration of multimodal foundation models. To overcome the singularity of foundation models by applying knowledge graphs to learn the transformation between one unimodal foundation model and another, and only the bridge module needs training while all the base foundation models are kept fixed, resulting in great computational efficiency. A various ranges of predic-

tion tasks can be performed via BioBridge including cross-modality retrieval tasks, semantic similarity inference, protein-protein interaction, and cross-species protein-phenotype matching. But it lacks quantitative evidence for molecular generation tasks.

The OFA [55]approach suggests using text-attributed graphs to represent the diverse cross-domain attributes and connections in a graph to combine various types of graph data. This method involves converting these descriptions into feature vectors in the same embedding space using language models, regardless of their original domain. Additionally, the method introduces "nodes-of-interest" to standardize how we approach different graph-related tasks using a single task. OFA also uses a unique method called graph prompting by adding special structures to the graph that act like prompts, allowing the model to perform a wide range of tasks without fine-tuning. The model is designed to handle various fields, such as citation networks, molecular structures, and knowledge bases. Despite the strengths of this method, the performance for individual tasks seems suboptimal.

Integrating deep learning with a knowledge graph of gene-gene interactions, GEARS [56] predicts transcriptional responses to both single and multigene perturbations using single-cell RNA sequencing data from perturbational screens. It employs a Graph Neural Network (GNN) to study genetic relationships and perturbational expression changes, enabling predictions for gene combinations not experimentally perturbed. However, the model is limited to the same cell type or experimental condition, and its reliance on combinatorial perturbational data introduces confounding factors that need further addressing.

More recently, TxGNN [57] has been introduced as a graph foundation model for drug repurposing. It explicitly identifies therapeutic candidates under a zero-shot setting, by implementing metric learning to transfer knowledge from well-studied disease to incurable diseases with no treatment. Biomedical knowledge is encoded with GNN, and a decoder incorporated with auxiliary information from similarity-based metric learning is further used to address the representation of the disease that may be sparsely annotated. The authors also include a model interpreter that is further validated by human experts. However, medical knowledge graphs are noisy and often incomplete, limiting model's predictive power, more information of molecular interactions may be further addressed. Additionally, the contradictory relationships between various entities in the knowledge graph are overseen.

PinnacleAI [58] stands out as a context-specific model for protein representation learning, it combines the information from multiple hierarchies including protein-protein interaction, cell type-to-cell type interactions and tissue-tissue interactions. The model has protein-, cell type- and tissue-level attention mechanisms that enable the algorithm to generate contextualized representations of proteins, cell types and tissues in a single unified embedding space. But this work again is limited to the cell types or tissues in the training set, failing to predict the specific diseases may not be represented in healthy human subjects.

4. Challenges in machine learning techniques

Despite significant progress in applying machine learning to the integration of multi-omics data and predictive modeling of genotype-environment-phenotype relationships, several challenges persist. These include the need for biologically informed representation learning, scarcity and ambiguity of labeled data, inability to generalize out-of-distribution, and dealing with incomplete and noisy graphs.

4.1. Need for biologically informed representation learning

A fundamental hurdle arises from the multi-level hierarchical organization of biological systems, as discussed in the Introduction section. On one hand, multiple statistically insignificant variations at a lower level can collectively result in significant changes at a higher level (e.g., gene expression) [88]. Hence, a network biology approach is imperative to enhance biological signals [89]. On the other hand, many genotypes exert a pleiotropic effect on complex diseases and traits [90]. Consequently, a higher-level endophenotype demonstrates greater discriminatory power concerning the organismal phenotype than a lower-level one. Therefore, a cross-level modeling approach is necessary to simulate the asymmetrical information transmission process between genotype and phenotype [91]. This, in turn, will enhance model interpretability and facilitate the elucidation of molecular underpinnings of phenotypes [92,93].

4.2. Scarcity and ambiguity of labeled data

The scarcity of labeled data significantly hinders the application of machine learning in the predictive modeling of genotype-environment-phenotype relationships through multi-omics data. Current multi-modal learning often necessitates paired omics data with shared labels, a challenge exacerbated by the infrequent availability of such labeled data in many instances. For example, transcriptomics and proteomics data from the brain tissues of Alzheimer's disease patients can only be obtained from post-mortem persons. Consequently, constructing a practical machine learning model for living patients relies on genomics or brain imaging data, despite transcriptomics and proteomics data exhibiting stronger predictive power for phenotypic responses to drug treatments and other environmental influences than genomics and brain imaging data.

The issue of *phenotype label ambiguity* is a concern that has not received sufficient attention in machine learning. Recent efforts, including Human Phenotype Ontology (HOP) [94] and Phenotype and Trait Ontology (PATO) [95], pave the way to address this problem. HPO is a standardized, comprehensive vocabulary that describes human phenotypic abnormalities encountered in genetic disorders. It provides a systematic way to characterize and classify observable traits, symptoms, and clinical features associated with human diseases. PATO provides a standardized vocabulary for describing phenotypic qualities in a manner that can be consistently applied across different species. However, additional efforts are needed to incorporate ontologies into machine learning models.

4.3. Inability to generalize out-of-distribution

A more pressing data issue emerges with an out-of-distribution (OOD) scenario, where new unseen cases differ significantly from the data used to train the model [96]. Technological limitations and human biases have illuminated only a fraction of the vast biological and chemical universe. For instance, among over 20,000 human genes, only proteins encoded by hundreds of genes have known small molecule ligands, without accounting for isoforms, protein complexes, mutation states, and conformations. Despite an estimated 10^{60} small organic molecules in the chemical space, only approximately 10^{8} have known bioactivities. Single-cell profiling techniques have generated omics data for numerous cell types, but only around 100 of them have controlled perturbations and functional genomics readouts. The combined space of chemicals, biomolecules, and endo- and organismal phenotypes is staggeringly vast [97].

Another significant issue arises due to a notable distribution shift from *in vitro* to *in vivo* settings. This shift often results in disease models failing to accurately reflect the efficacy and toxicity of drugs in humans. There is a critical need for a computational approach that can effectively disentangle confounding factors while preserving unique features. Existing methods that fail to adequately address confounding factors often overlook their connection to clinical outcomes. A more systematic approach is required to address this challenge.

To address the OOD problem, it becomes imperative to quantify the prediction uncertainty of new cases [98,99]. Uncertainty quantification is particularly critical in high-stakes applications like drug discovery and precision medicine. Given the resource-intensive nature of drug discovery, uncertainty quantification aids in decision-making by offering

insights into the confidence levels associated with predictions. In precision medicine, where erroneous predictions about drug efficacy or safety can have severe consequences, uncertainty quantification is essential for assessing the risks associated with model predictions.

4.4. Incomplete and noisy graphs

In the realm of predictive modeling for genotype-environment-phenotype relationships, two key issues within graph learning remain inadequately addressed: the incorporation of novel nodes lacking previously recognized connections in an established graph model and the identification of dubious or conflicting relationships.

The construction of a high-quality graph model for a biological system is a labor-intensive, domain-specific task that often demands manual data curation. Furthermore, the graph model may fall short in capturing implicit knowledge and intricate patterns not explicitly represented in the data, restricting its ability to unveil novel discoveries. This limitation is particularly critical in biology, where a vast number of biological and chemical entities remain uncharted, lacking any annotations. These unannotated nodes become isolated in the graph model, impeding inference for them. For instance, a drug-like chemical compound lacking significant structural similarity to existing drugs and without known protein targets becomes an isolated node in a drug-gene-disease graph. It becomes impractical and unreliable to infer its associations with diseases.

Various machine learning-based automatic processes have been developed to enhance graph models, such as predicting gene-disease associations through Natural Language Processing [14,43,45,100], and drug-target interaction predictions [101–103]. However, these predicted relationships may be inaccurate, resulting the introduction of false positives and conflicting relationships. Few attention has been paid to addressing the issue of dubious relationships in knowledge graphs, especially when it is generated from biomedical publications, many of which cannot be reproduced [104–106].

5. AI-powered knowledge-enriched multi-scale genotype-environment-phenotype predictive modeling

Recent advances in deep learning, coupled with the growing accessibility of multi-omics data, have opened avenues for predicting emergent phenotypes through novel perturbations under diverse genotypes. Leveraging these developments, we propose two complementary approaches and their combinations: (1) biology-inspired end-to-end multi-modal multi-task deep learning, (2)physics-informed context-specific multi-scale knowledge graphs.

5.1. Biology-inspired end-to-end multi-modal multi-task deep learning

Compared to classical machine learning, one of the unique features of deep neural networks is their capacity for end-to-end learning. End-to-end learning tackles a complex task from inception to completion, as opposed to dividing the task into smaller sub-tasks and addressing them independently. In the context of predictive modeling for genotype-environment-phenotype relationships, the conventional strategy requires paired data for all the modalities. In contrast, we propose an end-to-end deep neural network that explicitly models asymmetric information flows from DNAs to RNAs to proteins to metabolites and ultimately to the organismal phenotype, following the central dogma of molecular biology, as illustrated in Fig. 2. A foundation model for each data modality can be pre-trained and fine-tuned using modality-specific unlabeled and labeled data. When paired data across two biological levels is available, the models from different levels can be connected through contrastive learning [91], transfer learning [107], or other techniques [108]. With labeled organismal phenotype data, all modalities are interconnected and fine-tuned from genotypes to phenotypes. Environmental factors can be applied to any level, contingent on the nature of influences and perturbations — examples include CRISPR-Cas9 on DNA, RNAi on RNA, and small molecule inhibitors on proteins. Utilizing a fully-trained end-to-end model, it becomes feasible to incorporate endophenotype information, even if it cannot be directly obtained (such as brain tissue proteomics for a living AD patient), thereby improving predictions of organismal phenotypes from a genotype.

The biology-inspired end-to-end model can address the OOD and label scarcity problem from various perspectives. The pre-trained foundation model has exhibited notable generalization capabilities. For instance, the protein language model has proven successful in tasks such as protein structure predictions [109], protein design [110], and predicting protein-chemical interactions [101]. Contrastive learning has proven successful in integrating multi-omics data, as demonstrated in the previous section. Notably, several proof-of-concept studies have shown the promise of end-to-end models that adhere to the multi-level organization of a biological system. For example, the Cross-Level Information Transmission (CLEIT) network employs transcriptomics endophenotypes as an intermediate layer to connect genomic mutations with cellular phenotypes through contrastive learning [91]. This approach enhances phenotype predictions from genotypic data. Leveraging transfer learning, TransPro predicts proteomics profiles induced by unobserved chemicals based on transcriptomics data [107]. It is observed that predicting organismal phenotypes via predicted and imputed proteomics signatures by TransPro is more accurate than relying on experimentally determined transcriptomics or proteomics data, which often suffer from noise and sparsity. Combining contrastive learning with multi-task learning guided by clinical features, Guided-Stab achieved survival prediction by cancer transcriptomics [108]. An end-to-end model, which links genotypes to phenotypes by integrating multiple endophenotypes based on their biological relationships, is anticipated to offer a robust tool for establishing genotype-environment-phenotype relationships.

5.2. Personalized physics-informed multi-scale knowledge graph

Considering the elevated incidence of false negatives and false positives in relationships, as well as the presence of coarse-grained and ambiguous phenotypes in current biological network models, we propose three solutions to harness the potential of graph learning for predictive modeling of genotype-environment-phenotype relationships. These solutions comprise (1) the explicit representation of physical interactions within molecular networks, (2) the construction of context-dependent networks with fine-grained phenotypes, and (3) the development of multi-scale network models.

Genotype-phenotype relationships in many existing network models, such as gene-disease networks, primarily rely on statistical correlations derived from Genome-Wide Association Studies (GWAS). Without insight into the underlying molecular interactions, determining the molecular drivers responsible for a phenotype and predicting phenotypic responses to novel perturbations becomes challenging. By incorporating quantitative details of molecular interactions into the network, it becomes possible to rationalize how molecular changes may impact phenotypes. For example, mutations in DNA sequence can alter regulatory DNA-protein, regulatory RNA-protein, or protein-protein interactions, subsequently influencing the binding affinity or kinetics of these interactions, leading to changes in gene expression, signaling transduction, or metabolism. Illustrated in Fig. 3, representing experimentally determined DNA/RNA-protein, protein-protein, chemical-protein, and other interactions in a network model with weighted and signed edges encoding the degree (or certainty) and direction of interaction changes allows for more confident inference of genotype-phenotype relationships. High-throughput techniques have emerged to explore novel molecular interactions [111,112]. New machine learning methods, e.g., model-agnostic semi-supervised meta-learning, can efficiently explore understudied interactions [113]. Transfer learning enables predicting functional activities of ligand binding, i.e., antagonist vs agonist [114].

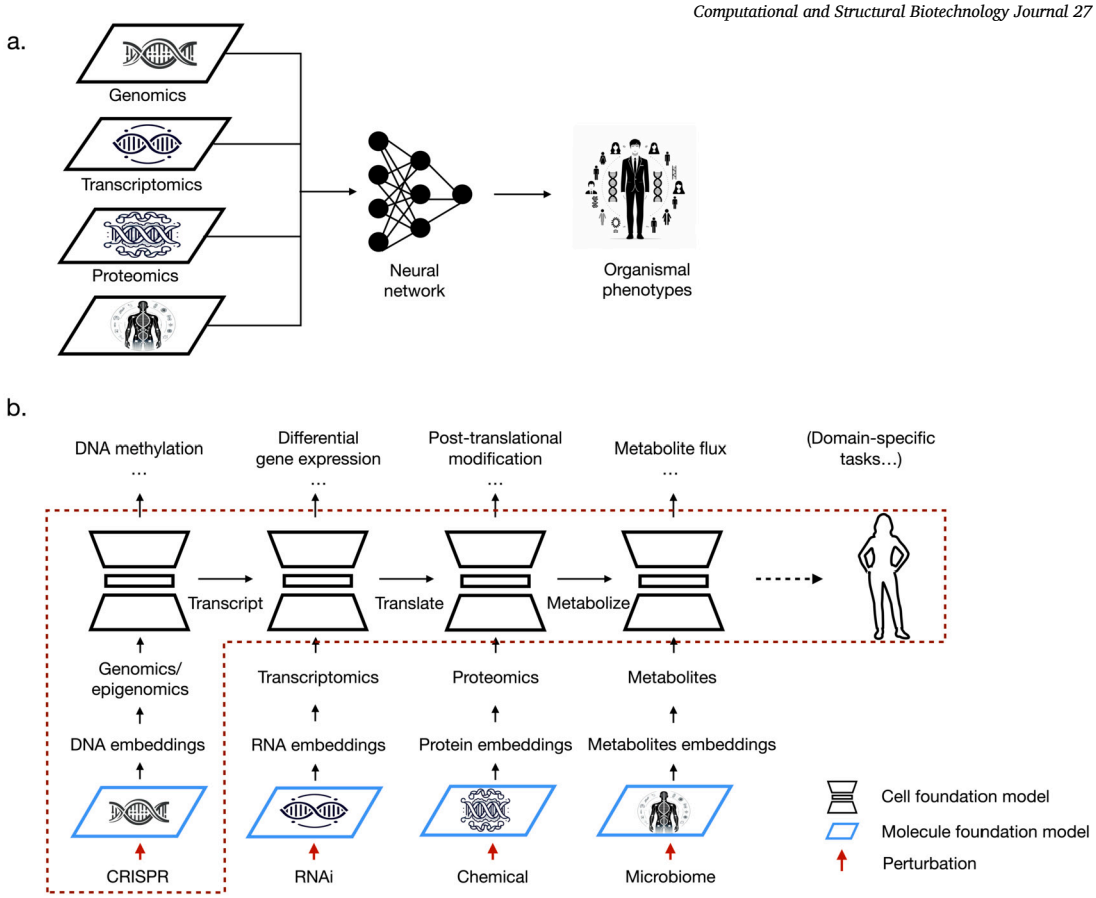

Fig. 2. Illustration of multi-modal supervised learning. (a) A conventional strategy that requires paired data for all the modalities simultaneously. (b) An end-to-end deep neural network explicitly models asymmetric information flows from DNAs to RNAs to proteins to metabolites and ultimately to the organismal phenotype, where intermediate endophenotype data may be missing. Molecule foundation models generate embeddings for perturbations such as CRISPR, RNAi, chemicals, and microbiome data, while cell foundation models extract key information from each omics layer (genomics, transcriptomics, proteomics, and metabolomics) using unlabeled data. These models are fine-tuned with domain-specific tasks, such as predicting methylation from genomics data, differential gene expression from transcriptomics data, and post-translational modifications from proteomics data. The model is further optimized using paired flows between modalities (e.g., DNA-RNA, RNA-protein) to capture complex relationships. Once fully trained, it can predict phenotypes from any given modality, even in the absence of intermediate data, by leveraging learned endophenotypes.

Many existing network models are canonical aggregations across different conditions. For instance, in a gene-disease network, "Alzheimer's disease" (AD) is often depicted by a single node, and the gene-gene interaction network remains constant across all diseases. However, AD has several subtypes resulting from different etiologies (e.g., APOE4 vs. TREM2). Similarly, the gene-gene interaction network undergoes rewiring dependent on biological contexts (such as cell types, disease stages, and species). This coarse-grained representation falls short of capturing the complexities of biology. We propose to decompose the aggregated network model into an interconnected multiplex network model. Each plex in the network represents a subtype or an individual. In the case of a gene-disease network, using disentangled embeddings of disease biomarkers (e.g., brain imaging for AD), a subtype of AD or an individual patient (i.e., phenotype) can be represented by a class-specific embedding and a subtype/individual-specific embedding, which can be derived from patient-level data like medical imaging and electronic health records. Subtype/individual-specific gene-gene interaction networks can be derived from gene embeddings learned from a large language model [43,81]. It is anticipated that such a fine-grained network model will be more potent in predictive modeling of genotype-environment-phenotype relationships compared to a coarse-grained aggregated model.

The inherent complexity and hierarchical organization of a biological system naturally lend themselves to representation on a multi-scale. For instance, a tissue can be portrayed through a cell-cell interacting network, and each cell can be captured by a cell type-specific gene-gene interacting network. Algorithmically, a multi-scale cell-cell interacting network can be conceptualized as a network of networks. While the network of networks concept has found widespread application in modeling areas such as the internet, smart cities, social networks, supply chains, telecommunications, cloud computing, and financial systems [115], its utilization in systems biology remains relatively limited [116]. Given the abundance of single-cell and spatial omics data, there is a compelling opportunity to explore the application of the network of networks paradigm for omics data integration and analysis in systems biology.

5.3. Integration of machine learning models, knowledge graphs, and generative AI

The proposed machine learning and knowledge graph approaches mentioned above are complementary. Integrating these two approaches will further enhance the predictive power of genotype-environment-phenotype relationships. Although the machine learning model excels at discerning subtle patterns from raw data and augmenting missing links within a knowledge graph, it may lack a comprehensive understanding of the global contexts of these patterns. Conversely, a knowledge graph can consolidate patterns into a cohesive network within a broader context. Inference of missing links from a knowledge graph can both validate and refute predictions made by a machine learning model.

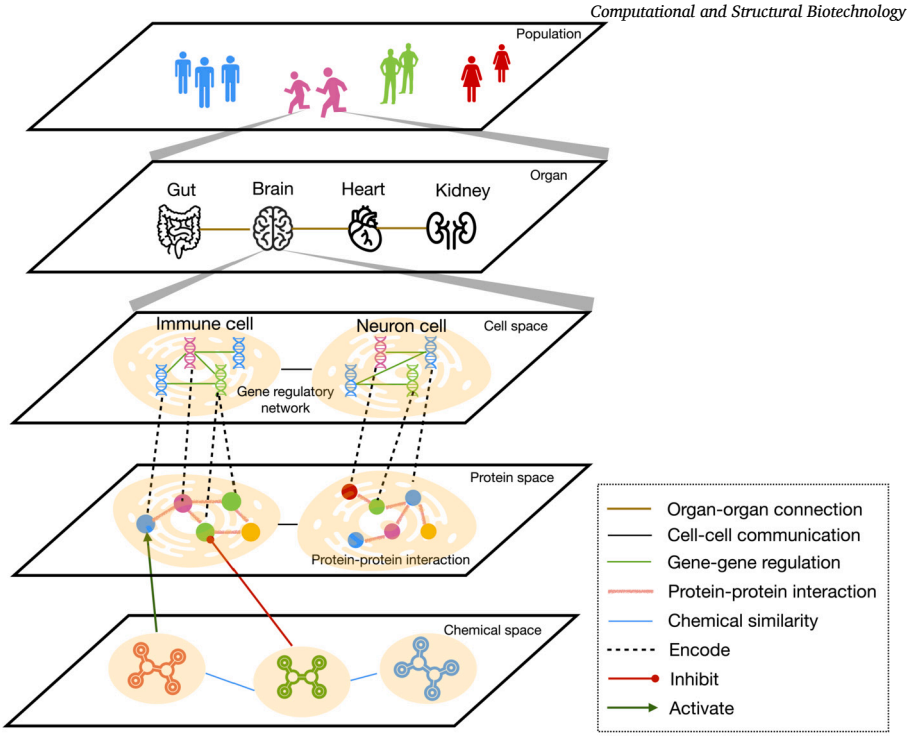

Fig. 3. Illustration of personalized physics-informed multi-scale knowledge graph. It represents a multi-scale network model capturing experimentally determined DNA/RNA-protein, protein-protein, chemical-protein, and other biological interactions. A drug or chemical can activate or inhibit a protein, which is represented by positive or negative edges, thereby influencing protein-protein interactions. Each protein encodes specific genes within different cell types through gene regulatory networks, propagating effects across cellular, organ, and population levels. Weighted and signed edges in the network encode the degree (or certainty) and direction of interaction changes, allowing more confident inference of genotype-phenotype relationships.

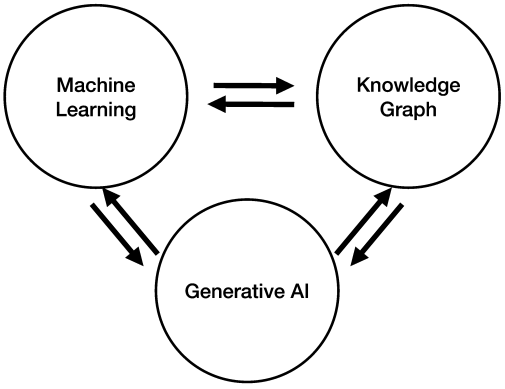

Fig. 4. Integration of machine learning models, mechanistic models, knowledge graphs, and generative AI.

Both machine learning models and knowledge graphs, which focus on predictive analytics, can benefit from integration with generative AI models. On one hand, a generative model can enhance predictive models in several ways. Generative models have the capability to generate synthetic data samples that closely resemble real data. These synthetic samples can effectively augment the training dataset of predictive models, particularly in scenarios where real data is limited. Furthermore, generative models can learn the underlying distribution of observed data, enabling them to identify outliers or OOD cases effectively. Additionally, they can be utilized to impute missing values by generating plausible values conditioned on the observed data. On the other hand, machine learning models can enhance personalization and mitigate hallucination in generative models through techniques such as reinforcement learning, attention mechanisms, conditional generation, active learning, and others Fig. 4.

6. Conclusion

The fusion of multi-omics data and AI techniques marks a significant advancement in comprehending complex biological systems and predicting outcomes across diverse environments and perturbations. In this paper, we have explored the interleaved interactions between genotype, environment, and phenotype, highlighting the pivotal role of endophenotypes as intermediate markers linking genetic makeup to observable traits. Central to our discussion is the integration of multi-omics data, spanning various biological levels from single cells to whole organisms, and encompassing different data modalities and species. We have addressed the shortcomings of current machine learning methods, particularly in accurately predicting relationships between genotype, environment, and phenotype. Our proposed framework, inspired by biology and driven by AI, aims to untangle the complexities of living organisms and lay the groundwork for personalized medicine.

It is important to underscore that AI alone cannot accomplish our objectives. A comprehensive representation of human biology and physiology needs a digital twin that captures micro and macro dynamics of the human body and its interactions with the environment [117–119]. This necessitates the integration of AI with mechanism-based modeling, a promising technique for addressing challenges in machine learning. For example, constraint-based metabolic network modeling can predict organismal phenotypes directly, such as growth rates under diverse conditions. Unlike “black box” machine learning models, mechanism-based models explicitly represent system processes and interactions, offering insights into underlying principles. Leveraging existing knowledge, they can make predictions even with limited data, exhibiting greater generalizability across scenarios. Their transparency facilitates interpretation and understanding of influencing factors, crucial for applications like biomedicine. Additionally, the seamless integration of prior knowledge enhances prediction accuracy and relevance. In conclusion, a biology-inspired AI model, coupled with mechanism-based modeling, holds considerable promise for advancing our understanding

of genotype-environmental-phenotype relationships and informing critical decision-making.

CRediT authorship contribution statement

You Wu: Writing – review & editing, Writing – original draft, Visualization, Validation, Supervision, Software, Resources, Project administration, Methodology, Investigation, Formal analysis, Data curation, Conceptualization. **Lei Xie:** Writing – review & editing, Writing – original draft, Supervision, Resources, Project administration, Methodology, Investigation, Formal analysis, Conceptualization.

Declaration of competing interest

The authors whose names are listed immediately below certify that they have no affiliations with or involvement in any organization or entity with any financial interest (such as honoraria; educational grants; participation in speakers' bureaus; membership, employment, consultancies, stock ownership, or other equity interest; and expert testimony or patent-licensing arrangements), or non-financial interest (such as personal or professional relationships, affiliations, knowledge or beliefs) in the subject matter or materials discussed in this manuscript.

You Wu
Lei Xie

Appendix A. Supplementary material

Supplementary material related to this article can be found online at https://doi.org/10.1016/j.csbj.2024.12.030.

References

[1] Via S, Lande R. Genotype-environment interaction and the evolution of phenotypic plasticity. Evolution 1985;39(3):505–22.
[2] Gottesman II, Gould TD. The endophenotype concept in psychiatry: etymology and strategic intentions. Am J Psychiatr 2003;160(4):636–45.
[3] Hasin Y, Seldin M, Lusis A. Multi-omics approaches to disease. Genome Biol 2017;18:1–15.
[4] Graw S, Chappell K, Washam CL, Gies A, Bird J, Robeson MS, et al. Multi-omics data integration considerations and study design for biological systems and disease. Mol Omics 2021;17(2):170–85.
[5] Cheng C, Chen W, Jin H, Chen X. A review of single-cell rna-seq annotation, integration, and cell–cell communication. Cells 2023;12(15):1970.
[6] 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(8):550–72.
[7] Baysoy A, Bai Z, Satija R, Fan R. The technological landscape and applications of single-cell multi-omics. Nat Rev Mol Cell Biol 2023;24(10):695–713.
[8] Vandereyken K, Sifrim A, Thienpont B, Voet T. Methods and applications for single-cell and spatial multi-omics. Nat Rev Genet 2023;24(8):494–515.
[9] Walsh LA, Quail DF. Decoding the tumor microenvironment with spatial technologies. Nat Immunol 2023;24(12):1982–93.
[10] Bressan D, Battistoni G, Hannon GJ. The dawn of spatial omics. Science 2023;381(6657):eabq4964.
[11] Palla G, Fischer DS, Regev A, Theis FJ. Spatial components of molecular tissue biology. Nat Biotechnol 2022;40(3):308–18.
[12] Dries R, Chen J, Del Rossi N, Khan MM, Sistig A, Yuan G-C. Advances in spatial transcriptomic data analysis. Genome Res 2021;31(10):1706–18.
[13] Song Y, Miao Z, Brazma A, Papatheodorou I. Benchmarking strategies for cross-species integration of single-cell rna sequencing data. Nat Commun 2023;14(1):6495.
[14] Yang X, Liu G, Feng G, Bu D, Wang P, Jiang J, et al. GeneCompass: Deciphering universal gene regulatory mechanisms with a knowledge-informed cross-species foundation model. Cell Research 2024;34:830–45.
[15] Rosen Y, Brbić M, Roohani Y, Swanson K, Li Z, Leskovec J. Toward universal cell embeddings: integrating single-cell rna-seq datasets across species with Saturn. Nat Methods 2024:1–9.
[16] Moffat JG, Vincent F, Lee JA, Eder J, Prunotto M. Opportunities and challenges in phenotypic drug discovery: an industry perspective. Nat Rev Drug Discov 2017;16(8):531–43.
[17] 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(12):899–914.
[18] Tomczak K, Czerwińska P, Wiznerowicz M. Review the cancer genome atlas (tcga): an immeasurable source of knowledge. Contemp Oncol/Współczesna Onkol 2015;2015(1):68–77.
[19] Koleti A, Terryn R, Stathias V, Chung C, Cooper DJ, Turner JP, et al. Data portal for the library of integrated network-based cellular signatures (lincs) program: integrated access to diverse large-scale cellular perturbation response data. Nucleic Acids Res 2018;46(D1):D558–66.
[20] Stathias V, Turner J, Koleti A, Vidovic D, Cooper D, Fazel-Najafabadi M, et al. Lincs data portal 2.0: next generation access point for perturbation-response signatures. Nucleic Acids Res 2020;48(D1):D431–9.
[21] Tsherniak A, Vazquez F, Montgomery PG, Weir BA, Kryukov G, Cowley GS, et al. Defining a cancer dependency map. Cell 2017;170(3):564–76.
[22] Barretina J, Caponigro G, Stransky N, Venkatesan K, Margolin AA, Kim S, et al. The cancer cell line encyclopedia enables predictive modelling of anticancer drug sensitivity. Nature 2012;483(o). 7391, pp. 603–607.
[23] Ghandi M, Huang FW, Jané-Valbuena J, Kryukov GV, Lo CC, McDonald III ER, et al. Next-generation characterization of the cancer cell line encyclopedia. Nature 2019;569(o). 7757, pp. 503–508.
[24] Garnett MJ, Edelman EJ, Heidorn SJ, Greenman CD, Dastur A, Lau KW, et al. Systematic identification of genomic markers of drug sensitivity in cancer cells. Nature 2012;483(o). 7391, pp. 570–575.
[25] Iorio F, Knijnenburg TA, Vis DJ, Bignell GR, Menden MP, Schubert M, et al. A landscape of pharmacogenomic interactions in cancer. Cell 2016;166(3):740–54.
[26] Basu A, Bodycombe NE, Cheah JH, Price EV, Liu K, Schaefer GI, et al. An interactive resource to identify cancer genetic and lineage dependencies targeted by small molecules. Cell 2013;154(5):1151–61.
[27] Seashore-Ludlow B, Rees MG, Cheah JH, Cokol M, Price EV, Coletti ME, et al. Harnessing connectivity in a large-scale small-molecule sensitivity dataset. Cancer Discov 2015;5(11):1210–23.
[28] Peidli S, Green TD, Shen C, Gross T, Min J, Garda S, et al. scperturb: harmonized single-cell perturbation data. Nat Methods 2024:1–10.
[29] Srivatsan SR, McFaline-Figueroa JL, Ramani V, Saunders L, Cao J, Packer J, et al. Massively multiplex chemical transcriptomics at single-cell resolution. Science 2020;367(6473):45–51.
[30] Smirnov P, Kofia V, Maru A, Freeman M, Ho C, El-Hachem N, et al. Pharmacodb: an integrative database for mining in vitro anticancer drug screening studies. Nucleic Acids Res 2018;46(D1):D994–1002.
[31] Lautenbacher L, Samaras P, Muller J, Grafberger A, Shraideh M, Rank J, et al. Proteomicsdb: toward a fair open-source resource for life-science research. Nucleic Acids Res 2022;50(D1):D1541–52.
[32] Eckert S, Berner N, Kramer K, Schneider A, Müller J, Lechner S, et al. Decrypting the molecular basis of cellular drug phenotypes by dose-resolved expression proteomics. Nat Biotechnol 2024:1–10.
[33] Zecha J, Bayer FP, Wiechmann S, Woortman J, Berner N, Müller J, et al. Decrypting drug actions and protein modifications by dose- and time-resolved proteomics. Science 2023;380(6640):93–101.
[34] Lopez R, Regier J, Cole MB, Jordan MI, Yosef N. Deep generative modeling for single-cell transcriptomics. Nat Methods 2018;15(12):1053–8.
[35] Xu C, Lopez R, Mehlman E, Regier J, Jordan MI, Yosef N. Probabilistic harmonization and annotation of single-cell transcriptomics data with deep generative models. Mol Syst Biol 2021;17(1):e9620.
[36] Gayoso A, Steier Z, Lopez R, Regier J, Nazor KL, Streets A, et al. Joint probabilistic modeling of single-cell multi-omic data with totalvi. Nat Methods 2021;18(3):272–82.
[37] Gong B, Zhou Y, Purdom E. Cobolt: integrative analysis of multimodal single-cell sequencing data. Genome Biol 2021;22:1–21.
[38] Ashuach T, Gabitto MI, Koodli RV, Saldi G-A, Jordan MI, Yosef N. Multivi: deep generative model for the integration of multimodal data. Nat Methods 2023;20(8):1222–31.
[39] Li G, Fu S, Wang S, Zhu C, Duan B, Tang C, et al. A deep generative model for multi-view profiling of single-cell rna-seq and atac-seq data. Genome Biol 2022;23(1):20.
[40] Cao Z-J, Gao G. Multi-omics single-cell data integration and regulatory inference with graph-linked embedding. Nat Biotechnol 2022;40(10):1458–66.
[41] Piran Z, Cohen N, Hoshen Y, Nitzan M. Disentanglement of single-cell data with biolord. Nat Biotechnol 2024:1–6.
[42] Hetzel L, Boehm S, Kilbertus N, Günnemann S, Theis F, et al. Predicting cellular responses to novel drug perturbations at a single-cell resolution. Adv Neural Inf Process Syst 2022;35:711–26722.
[43] Cui H, Wang C, Maan H, Pang K, Luo F, Duan N, et al. scgpt: toward building a foundation model for single-cell multi-omics using generative AI. Nat Methods 2024:1–11.
[44] Ji Y, Tejada-Lapuerta A, Schmacke NA, Zheng Z, Zhang X, Khan S, et al. Scalable and universal prediction of cellular phenotypes. bioRxiv, 2024.
[45] Rosen Y, Brbić M, Roohani Y, Swanson K, Li Z, Leskovec J. Toward universal cell embeddings: integrating single-cell rna-seq datasets across species with Saturn. Nat Methods 2024:1–9.
[46] Xiong L, Chen T, Kellis M. scclip: multi-modal single-cell contrastive learning integration pre-training. In: NeurIPS 2023 AI for science workshop; 2023.
[47] Gossi F, Pati P, Martinelli A, Rapsomaniki MA. Matchclot: single-cell modality matching with contrastive learning and optimal transport. bioRxiv, 2022.

[48] Yang KD, Belyaeva A, Venkatachalapathy S, Damodaran K, Katcoff A, Radhakrishnan A, et al. Multi-domain translation between single-cell imaging and sequencing data using autoencoders. Nat Commun 2021;12(1):31.
[49] Chen RJ, Lu MY, Williamson DF, Chen TY, Lipkova J, Noor Z, et al. Pan-cancer integrative histology-genomic analysis via multimodal deep learning. Cancer Cell 2022;40(8):865–78.
[50] Yang B, Yang Y, Su X. Deep structure integrative representation of multi-omics data for cancer subtyping. Bioinformatics 2022;38(13):3337–42.
[51] Zhang C, Chen Y, Zeng T, Zhang C, Chen L. Deep latent space fusion for adaptive representation of heterogeneous multi-omics data. Brief Bioinform 2022;23(2):bbab600.
[52] Moon S, Lee H. Moma: a multi-task attention learning algorithm for multi-omics data interpretation and classification. Bioinformatics 2022;38(8):2287–96.
[53] Lee J, Kim D, Kong J, Ha D, Kim I, Park M, et al. Cell-cell communication network-based interpretable machine learning predicts cancer patient response to immune checkpoint inhibitors. Science Advances 2024;10(5):eadj0785.
[54] Wang Z, Wang Z, Srinivasan B, Ioannidis VN, Rangwala H, Anubhai R. Biobridge: bridging biomedical foundation models via knowledge graph. arXiv preprint.. Available from: arXiv:2310.03320, 2023.
[55] Liu H, Feng J, Kong L, Liang N, Tao D, Chen Y, et al. One for all: towards training one graph model for all classification tasks. arXiv preprint. arXiv:2310.00149, 2023.
[56] Roohani Y, Huang K, Leskovec J. Predicting transcriptional outcomes of novel multigene perturbations with gears. Nat Biotechnol 2023:1–9.
[57] Huang K, Chandak P, Wang Q, Havaldar S, Vaid A, Leskovec J, et al. A foundation model for clinician-centered drug repurposing. Nat Med 2024:1–13.
[58] Li MM, Huang Y, Sumathipala M, Liang MQ, Valdeolivas A, Ananthakrishnan AN, et al. Contextual AI models for single-cell protein biology. Nat Methods 2024;21(8):1546–57.
[59] Dou J, Liang S, Mohanty V, Miao Q, Huang Y, Liang Q, et al. Bi-order multimodal integration of single-cell data. Genome Biol 2022;23(1):112.
[60] Stuart T, Butler A, Hoffman P, Hafemeister C, Papalexi E, Mauck WM, et al. Comprehensive integration of single-cell data. Cell 2019;177(7):1888–902.
[61] Hie B, Bryson B, Berger B. Efficient integration of heterogeneous single-cell transcriptomes using scanorama. Nat Biotechnol 2019;37(6):685–91.
[62] Chen S, Zhu B, Huang S, Hickey JW, Lin KZ, Snyder M, et al. Integration of spatial and single-cell data across modalities with weakly linked features. Nat Biotechnol 2023:1–11.
[63] Kriebel AR, Welch JD. Uinmf performs mosaic integration of single-cell multi-omic datasets using nonnegative matrix factorization. Nat Commun 2022;13(1):780.
[64] Liu J, Gao C, Sodicoff J, Kozareva V, Macosko EZ, Welch JD. Jointly defining cell types from multiple single-cell datasets using liger. Nat Protoc 2020;15(11):3632–62.
[65] Argelaguet R, Arnol D, Bredikhin D, Deloro Y, Velten B, Marioni JC, et al. Mofa+: a statistical framework for comprehensive integration of multi-modal single-cell data. Genome Biol 2020;21:1–17.
[66] Haghverdi L, Lun AT, Morgan MD, Marioni JC. Batch effects in single-cell rna-sequencing data are corrected by matching mutual nearest neighbors. Nat Biotechnol 2018;36(5):421–7.
[67] Hao Y, Hao S, Andersen-Nissen E, Mauck WM, Zheng S, Butler A, et al. Integrated analysis of multimodal single-cell data. Cell 2021;184(13):3573–87.
[68] Hao Y, Stuart T, Kowalski MH, Choudhary S, Hoffman P, Hartman A, et al. Dictionary learning for integrative, multimodal and scalable single-cell analysis. Nat Biotechnol 2023:1–12.
[69] Li J, Lu Q, Wen Y. Multi-kernel linear mixed model with adaptive lasso for prediction analysis on high-dimensional multi-omics data. Bioinformatics 2020;36(6):1785–94.
[70] Yang H, Cao H, He T, Wang T, Cui Y. Multilevel heterogeneous omics data integration with kernel fusion. Brief Bioinform 2020;21(1):156–70.
[71] Manica M, Cadow J, Mathis R, Rodríguez Martínez M. Pimkl: pathway-induced multiple kernel learning. npj Syst Biol Appl 2019;5(1):8.
[72] Mariette J, Vialaneix N. Kernels for omics. In: Biological data integration: computer and statistical approaches; 2024. p. 151–93.
[73] Hinton GE. Training products of experts by minimizing contrastive divergence. Neural Comput 2002;14(8):1771–800.
[74] Devlin J, Chang M-W, Lee K, Toutanova K. Bert: pre-training of deep bidirectional transformers for language understanding. arXiv preprint. arXiv:1810.04805, 2018.
[75] Brown T, Mann B, Ryder N, Subbiah M, Kaplan JD, Dhariwal P, et al. Language models are few-shot learners. Adv Neural Inf Process Syst 2020;33:1877–901.
[76] Achiam J, Adler S, Agarwal S, Ahmad L, Akkaya I, Aleman FL, et al. Gpt-4 technical report. arXiv preprint. arXiv:2303.08774, 2023.
[77] Anil R, Dai AM, Firat O, Johnson M, Lepikhin D, Passos A, et al. Palm 2 technical report. arXiv preprint. arXiv:2305.10403, 2023.
[78] Chowdhery A, Narang S, Devlin J, Bosma M, Mishra G, Roberts A, et al. Palm: scaling language modeling with pathways. J Mach Learn Res 2023;24(240):1–113.
[79] Touvron H, Lavril T, Izacard G, Martinet X, Lachaux M-A, Lacroix T, et al. Llama: open and efficient foundation language models. arXiv preprint. arXiv:2302.13971, 2023.
[80] Ramesh A, Pavlov M, Goh G, Gray S, Voss C, Radford A, et al. Zero-shot text-to-image generation. In: Meila M, Zhang T, editors. Proceedings of the 38th international conference on machine learning. Proceedings of machine learning research, vol. 139. 2021. p. 8821–31. Available from: https://proceedings.mlr.press/v139/ramesh21a.html.
[81] 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(o). 7965, pp. 616–624.
[82] Hao M, Gong J, Zeng X, Liu C, Guo Y, Cheng X, et al. Large scale foundation model on single-cell transcriptomics. bioRxiv. 2023.
[83] Kedzierska KZ, Crawford L, Amini AP, Lu AX. Assessing the limits of zero-shot foundation models in single-cell biology. bioRxiv. 2023. p. 2023–310.
[84] Lin Z, Akin H, Rao R, Hie B, Zhu Z, Lu W, et al. Evolutionary-scale prediction of atomic-level protein structure with a language model. Science 2023;379(6637):1123–30.
[85] Radford A, Kim JW, Hallacy C, Ramesh A, Goh G, Agarwal S, et al. Learning transferable visual models from natural language supervision. In: International conference on machine learning. PMLR; 2021. p. 8748–63.
[86] Li MM, Huang K, Zitnik M. Graph representation learning in biomedicine and healthcare. Nat Biommed Eng 2022;6(12):1353–69.
[87] Lee B, Zhang S, Poleksic A, Xie L. Heterogeneous multi-layered network model for omics data integration and analysis. Front Genet 2020;10:501269.
[88] Barghi N, Hermisson J, Schlötterer C. Polygenic adaptation: a unifying framework to understand positive selection. Nat Rev Genet 06 2020;21:1–13.
[89] Cowen L, Ideker T, Raphael BJ, Sharan R. Network propagation: a universal amplifier of genetic associations. Nat Rev Genet 2017;18(9):551–62.
[90] Sivakumaran S, Agakov F, Theodoratou E, Prendergast JG, Zgaga L, Manolio T, et al. Abundant pleiotropy in human complex diseases and traits. Am J Hum Genet 2011;89(5):607–18.
[91] He D, Xie L. A cross-level information transmission network for hierarchical omics data integration and phenotype prediction from a new genotype. Bioinformatics 2022;38(1):204–10.
[92] Yang JH, Wright SN, Hamblin M, McCloskey D, Alcantar MA, Schrübbers L, et al. A white-box machine learning approach for revealing antibiotic mechanisms of action. Cell 2019;177(6):1649–61.
[93] Liu Q, Xie L. Transynergy: mechanism-driven interpretable deep neural network for the synergistic prediction and pathway deconvolution of drug combinations. PLoS Comput Biol 2021;17(2):e1008653.
[94] Robinson PN, Köhler S, Bauer S, Seelow D, Horn D, Mundlos S. The human phenotype ontology: a tool for annotating and analyzing human hereditary disease. Am J Hum Genet 2008;83(5):610–5.
[95] Gkoutos GV, Schofield PN, Hoehndorf R. The anatomy of phenotype ontologies: principles, properties and applications. Brief Bioinform 2018;19(5):1008–21.
[96] Liu J, Shen Z, He Y, Zhang X, Xu R, Yu H, et al. Towards out-of-distribution generalization: a survey. arXiv preprint. arXiv:2108.13624, 2021.
[97] Lunke S, Bouffler SE, Patel CV, Sandaradura SA, Wilson M, Pinner J, et al. Integrated multi-omics for rapid rare disease diagnosis on a national scale. Nat Med 2023;29(7):1681–91.
[98] Gawlikowski J, Tassi CRN, Ali M, Lee J, Humt M, Feng J, et al. A survey of uncertainty in deep neural networks. Artif Intell Rev 2023;56(Suppl 1):1513–89.
[99] Seoni S, Jahmunah V, Salvi M, Barua PD, Molinari F, Acharya UR. Application of uncertainty quantification to artificial intelligence in healthcare: a review of last decade (2013–2023). Comput Biol Med 2023:107441.
[100] Wu Y, Liu Q, Qiu Y, Xie L. Deep learning prediction of chemical-induced dose-dependent and context-specific multiplex phenotype responses and its application to personalized Alzheimer's disease drug repurposing. PLoS Comput Biol 2022;18(8):e1010367.
[101] Cai T, Xie L, Zhang S, Chen M, He D, Badkul A, et al. End-to-end sequence-structure-function meta-learning predicts genome-wide chemical-protein interactions for dark proteins. PLoS Comput Biol 2023;19(1):e1010851.
[102] Chen L, Tan X, Wang D, Zhong F, Liu X, Yang T, et al. Transformercpi: improving compound–protein interaction prediction by sequence-based deep learning with self-attention mechanism and label reversal experiments. Bioinformatics 2020;36(16):4406–14.
[103] Li M, Lu Z, Wu Y, Li Y. Bacpi: a bi-directional attention neural network for compound–protein interaction and binding affinity prediction. Bioinformatics 2022;38(7):1995–2002.
[104] Prinz F, Schlange T, Asadullah K. Believe it or not: how much can we rely on published data on potential drug targets? Nat Rev Drug Discov 2011;10(9):712–.
[105] Begley CG, Ellis LM. Raise standards for preclinical cancer research. Nature 2012;483(7391):531–3.
[106] Brito JJ, Li J, Moore JH, Greene CS, Nogoy NA, Garmire LX, et al. Recommendations to enhance rigor and reproducibility in biomedical research. GigaScience 2020;9(6). giaa056.
[107] Wu Y, Liu Q, Xie L. Hierarchical multi-omics data integration and modeling predict cell-specific chemical proteomics and drug responses. Cell Rep Methods 2023;3(4).
[108] Wu Y, Bazgir O, Lee Y, Biancalani T, Lu J, Hajiramezanali E. Multitask-guided self-supervised tabular learning for patient-specific survival prediction. In: Machine learning in computational biology; 2024. p. 10–22.
[109] Jumper J, Evans R, Pritzel A, Green T, Figurnov M, Ronneberger O, et al. Highly accurate protein structure prediction with alphafold. Nature 2021;596(7873):583–9.
[110] Madani A, Krause B, Greene ER, Subramanian S, Mohr BP, Holton JM, et al. Large language models generate functional protein sequences across diverse families. Nat Biotechnol 2023;41(8):1099–106.

[111] Reinecke M, Brear P, Vornholz L, Berger B-T, Seefried F, Wilhelm S, et al. Chemical proteomics reveals the target landscape of 1,000 kinase inhibitors. Nat Chem Biol 2023:1–9.
[112] Nechay M, Kleiner RE. High-throughput approaches to profile rna-protein interactions. Curr Opin Chem Biol 2020;54:37–44.
[113] Wu Y, Xie L, Liu Y, Xie L. Semi-supervised meta-learning elucidates understudied molecular interactions. Commun Biol 2024;7(1):1104. https://www.nature.com/articles/s42003-024-06797-z.
[114] Cai T, Abbu KA, Liu Y, Xie L. Deepreal: a deep learning powered multi-scale modeling framework for predicting out-of-distribution ligand-induced gpcr activity. Bioinformatics 2022;38(9):2561–70.
[115] Kenett DY, Perc M, Boccaletti S. Networks of networks–an introduction. Chaos Solitons Fractals 2015;80:1–6.
[116] Schuster M, Greenberg EP. A network of networks: quorum-sensing gene regulation in pseudomonas aeruginosa. Int J Med Microbiol 2006;296(2–3):73–81.
[117] Tang C, Yi W, Occhipinti E, Dai Y, Gao S, Occhipinti LG. A roadmap for the development of human body digital twins. Nat Rev Electr Eng 2024:1–9.
[118] National Academies of Sciences, Engineering, and Medicine and others. Foundational research gaps and future directions for digital twins; 2023.
[119] Katsoulakis E, Wang Q, Wu H, Shahriyari L, Fletcher R, Liu J, et al. Digital twins for health: a scoping review. npj Digit Med 2024;7(1):77.