REVIEW

National Science Review
XX: output, Year
doi: https://doi.org/10.1093/nsr/XXXX

INFORMATION SCIENCE

Foundation Models in Bioinformatics

Fei Guo[1], Renchu Guan[2], Yaohang Li[3], Qi Liu[4], Xiaowo Wang[5], Can Yang[6] and Jianxin Wang[1,*]

ABSTRACT

With the adoption of Foundation Models (FMs), Artificial Intelligence (AI) has become increasingly significant in bioinformatics and has successfully addressed many historical challenges, such as pre-training frameworks, model evaluation, and interpretability. FMs demonstrate notable proficiency in managing large-scale, unlabeled datasets, because experimental procedures are costly and labor-intensive. In various downstream tasks, FMs have consistently achieved noteworthy results, demonstrating high levels of accuracy in representing biological entities. A new era in computational biology has been ushered in by the application of FMs, focusing on both general and specific biological issues. In this review, we introduce recent advancements in bioinformatics FMs that employed in a variety of downstream tasks, including genomics, transcriptomics, proteomics, drug discovery, and single cell analysis. Our aim is to assist scientists in selecting appropriate FMs in bioinformatics, according to four model types: language FMs, vision FMs, graph FMs, and multimodal FMs. In addition to understanding molecular landscapes, AI technology can establish the theoretical and practical foundation for continued innovation in molecular biology.

Keywords: Foundation Model, Bioinformatics, Genomics, Transcriptomics, Proteomics, Drug discovery, Single cell analysis

[1] School of Computer Science and Engineering, Central South Univerisity, Changsha 410083, China;
[2] College of Computer Science and Technology, Jilin University, Changchun 130012, China;
[3] Department of Computer Science, Old Dominion University, Norfolk 23529, USA;
[4] School of Life Sciences and Technology, Tongji University, Shanghai 200092, China;
[5] Department of Automation, Tsinghua University, Beijing 100084, China;
[6] Department of Mathematics, The Hong Kong University of Science and Technology, Hong Kong 999077, China

[*] **Corresponding authors.**
Email: jxwang@mail.csu.edu.cn.

Received: XX XX Year;
Revised: XX XX Year;
Accepted: XX XX Year

INTRODUCTION

Foundation models represent large-scale artificial intelligence systems that undergo extensive pre-training on vast datasets, thereby enabling their application to a diverse array of downstream tasks. FMs are built by training neural networks on labeled and unlabeled data, enabling them to discern fundamental patterns and generalize knowledge to novel tasks. Before the emergence of foundation models, most AI systems were constructed using more traditional methodologies, which relied heavily on explicit human engineering and predefined rules rather than learning directly from data. The emergence of large-scale Pre-Trained Models (PTMs) has fundamentally transformed the landscape of artificial intelligence. The field is currently undergoing a paradigm shift, propelled by the development of models trained on extensive datasets that can be applied across a diverse array of downstream applications. Foundation models present significant opportunities and inherent risks, arising from their capabilities and underlying technical principles, as well as their applications and societal implications [1]. As computational power and data availability continue to expand, significant breakthroughs are being achieved in four key areas: the design of effective architectures, the utilization of rich contextual information, the enhancement of computational efficiency, and the execution of interpretative analysis. The development of FMs underscores the pivotal role of PTMs within the spectrum of AI technology.

As with pre-training architectures, a lot of large-scale foundation models are categorized into four various types of AI models, including language FMs, vision FMs, graph FMs, and multimodal FMs. Language FMs: Word2Vec [2] is an early PTM for converting words into distributed representations; Transformers [3] deal with sequential data, training Large Language Models (LLMs) than Recurrent Neural Networks (RNNs); BERT [4] and GPT [5] are transformer-based PTMs that differ from word-level PTMs. Vision FMs: AlexNet [6] is a Convolutional Neural Networks (CNN) that has significantly advanced Computer Vision (CV);

ORIGINAL UNEDITED MANUSCRIPT

Downloaded from https://academic.oup.com/nsr/advance-article/doi/10.1093/nsr/nwaf028/7979309 by guest on 08 February 2025

© The Author(s) 2025. Published by Oxford University Press on behalf of China Science Publishing & Media Ltd. This is an Open Access article distributed under the terms of the Creative Commons Attribution License (http://creativecommons.org/licenses/by/4.0/), which permits unrestricted reuse, distribution, and reproduction in any medium, provided the original work is properly cited.

ResNet [7] introduces shortcut connections with residual layers trained on ImageNet; Segment Anything Model (SAM) [8] is a promptable segmentation method that segments everything everywhere. Graph FMs: Graph Neural Networks (GNNs) are information processing architectures for emerging and homogenizing tasks; MPNN [9] and GIN [10] employ global and local temporal message-passing mechanisms; Graphormer [11] represents structural relationships between nodes using spatial encoding; GraphRAG [12] is a structured, hierarchical framework for Retrieval Augmented Generation (RAG). Multimodal FMs: ViT [13] outperforms conventional supervised CNN in preliminary studies; CLIP [14] constructs a transformer-based multimodal PTM that demonstrates promising results.

Recently, a number of foundation models have been successfully applied to bioinformatics problems, such as biomarker discovery, enzyme design, antibody-antigen recognition, drug discovery, omics analysis, and disease diagnosis. The objective of this study is to provide an analysis of bioinformatics FMs that can be trained both supervised and unsupervised learning models for applications such as core biological problems and integrated biological issues. With AI technology, it is possible to understand the molecular landscape, as well as aspects of human physiology and molecular biology. Several prominent foundation models are used to gain a deeper understanding of high throughput biological data, followed by a discussion of how prediction and generation models have been applied at various downstream tasks in bioinformatics, as shown in Figure 1.

Current surveys examine bioinformatics FMs from three perspectives. Firstly, several review articles summarize the large language models applied to bioinformatics tasks. Gao [15] outlined transformer-based, bioinformatics-tailored foundation models, directly applied to biological sequence data and serializable data. Heider [16] discussed large language models utilized to identify patterns in bioinformatics and analyzed their potential to revolutionize and accelerate multi-omics and personalized medicine discoveries. Moreover, a few survey papers enumerate the specific models for solving bioinformatics problems. Cheng [17] summarized diffusion modeling frameworks used in computational biology to generate proteins, drugs, and models of protein-ligand interactions. Furthermore, some review literature summarizes many traditional models in the field of bioinformatics and medicine. Li [18] summed up current trends in deep learning models to examine specific biological challenges, evaluating their applications to sequence analysis, structure prediction, and function annotation. Rajpurkar [19] listed generalist medical AI models combining electronic health records, genomics, clinical texts, and medical modalities. Despite this, most current surveys focus almost exclusively on one category of large-scale models or some traditional models that are applied to bioinformatics without taking into account various foundation models.

This review offers new insights into three primary objectives of foundation models in bioinformatics. First, we introduce recent improvements in bioinformatics foundation models as versatile tools. A comprehensive understanding of bioinformatics applications is provided by focusing on four types of foundation models, such as language FMs, vision FMs, graph FMs, and multimodal FMs. Second, we examine bioinformatics FMs for five downstream tasks, including genomics, transcriptomics, proteomics, drug discovery, and single cell analysis. Our discussion focuses on biological databases, training strategies, hyperparameter sizes, and biological applications. Finally, we discuss our perspective on the promising trajectory of bioinformatics FMs, drawing on our experiences with model pre-training frameworks, benchmarking selections, whiteboxes and interpretability, and evaluating model hallucinations.

EVOLUTION OF BIOINFORMATICS FM

The application of FMs in bioinformatics, which grew in popularity along with deep learning, was triggered by the introduction of large-scale pre-trained models. As a result of these efforts in bioinformatics, foundation models (language FMs, vision FMs, graph FMs, and multimodal FMs) have shown promising results for applications in biology (genomics, transcriptomics, proteomics, drug discovery, and single-cell analysis). The underlying bioinformatics foundation models need to be systematically reviewed, with a particular focus on various deep learning architectures. The evolution of FMs in bioinformatics is shown in Figure 2.

The initial bioinformatics models primarily focused on specific prediction tasks, such as sequence classification (e.g., DNA/RNA/protein sequence annotation) and secondary structure prediction. These models were typically task-specific and often developed from the ground up using domain-specific datasets. The subsequent phase in bioinformatics feature modeling involved the adoption of general pre-training

ORIGINAL UNEDITED MANUSCRIPT

Downloaded from https://academic.oup.com/nsr/advance-article/doi/10.1093/nsr/nwaf028/7979309 by guest on 08 February 2025

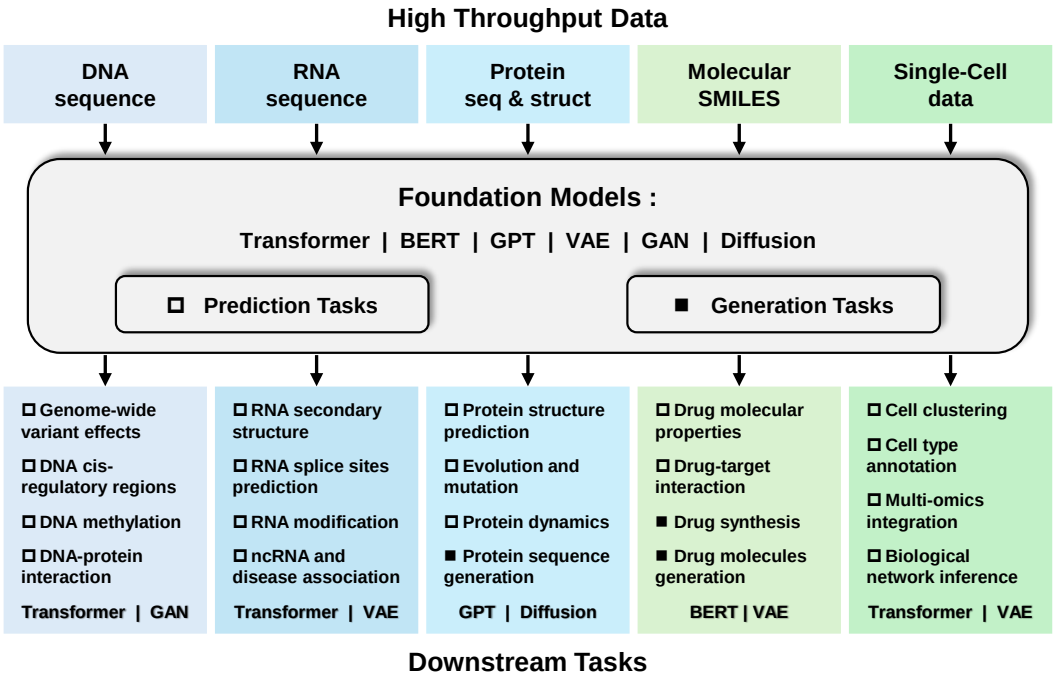

Figure 1. Foundation models in bioinformatics.

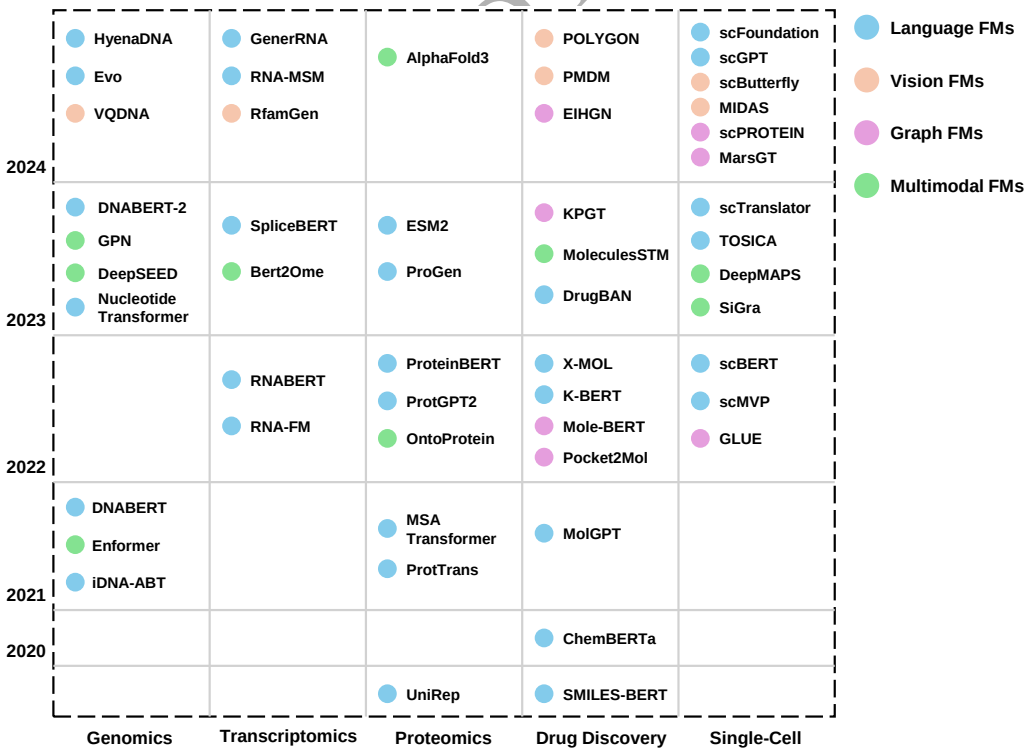

Figure 2. Evolution of FMs in bioinformatics.

Downloaded from https://academic.oup.com/nsr/advance-article/doi/10.1093/nsr/nwaf028/7979309 by guest on 08 February 2025

strategies similar to those employed in natural language processing, such as BERT. Models like DNABERT [20] were created to leverage large-scale pre-training on genomic data, enabling them to capture broad, transferable features inherent in biological sequences. As feature models evolved, they began incorporating multi-task learning and transfer learning techniques, which allowed them to excel across a diverse array of biological tasks including protein folding, genomic sequence analysis, and drug discovery. The emergence of models like AlphaFold [21] for protein structure prediction exemplifies this era; these large-scale pre-trained models could be fine-tuned for highly specialized tasks, resulting in groundbreaking advancements within structural bioinformatics. The latest generation of bioinformatics foundation models is characterized by their multimodal capabilities, which facilitate the integration and reasoning over various types of biological data—including genomic, transcriptomic, proteomic, and even clinical information. Models such as GLUE [22], originally developed for natural language processing tasks but subsequently adapted for bioinformatics applications like literature mining, and multitask deep learning frameworks designed for multi-omics data exemplify the growing trend towards multimodal learning.

By revealing the evolution process of bioinformatics FM, it is possible to understand in depth how the revised model overcomes the limitations and shortcomings of the primary model. Taking advantage of the latest bioinformatics FM, one can achieve unprecedented accuracy, realize an integrated AI model, and perform richer downstream analysis. Taking the classic biological problem "protein three-dimensional structure reconstruction" as a representative demonstration, DeepMind has developed three iterations of an artificial intelligence system over the past five years. AlphaFold [21] was developed to predict protein structure models in 2020, whose prediction is far better than any other method. As a result of AlphaFold's success, deep learning has enormous potential in the field of protein structure. In order to realize this process, AlphaFold follows two steps. First, AlphaFold is to predict the distance and rotation distribution of residues by utilizing 220 deep residual convolution blocks. After that, the gradient descent on the potential-specific potential predicted in the first step is used to conclude the protein structure. Due to its two-step modeling, AlphaFold loses valuable information, especially in regard to the dependence between long-range residue pairs. Furthermore, AlphaFold2 [23] achieves high-precision predictions through two main components: 'EvoFormer' and 'Structure module'. An innovative mechanism for exchanging information with the Multiple Sequence Alignment (MSA) is used in the Evoformer block, which employs a number of novel attention-based components. A significant advancement in the field is the direct inference of paired representations encapsulating both spatial and evolutionary relationships. In contrast to AlphaFold1, the structure module uses invariant point attention to predict 3D coordinates directly. Both AlphaFold1 and AlphaFold2 have attained outstanding performance in protein structure prediction, but their predictive prowess is augmented by the richness and accuracy of the MSA data they utilized. Recently, AlphaFold3 [24] has been meticulously developed and boasts remarkable capabilities in reconstructing the structure of complex biomolecular assemblies, including amino acids, nucleic acids, small molecules, and others. In structural biology, this is a significant breakthrough, offering scientists unprecedented insights into the complex interactions and architectures of biological molecules. A modification of AlphaFold2's architecture, AlphaFold3 uses only four blocks to interact with MSA information and pair representations, reducing dependence on MSA. Generated AI models are used to generate protein structure modules based on these learned representations. Furthermore, AlphaFold3 incorporates a diffusion module that estimates raw atomic coordinates directly, which works with amino-acid-specific frames and torsion angles for side chains. It eliminates the steps involved with traditional frame- and angle-based representations, making it possible to predict molecular structures more directly and holistically. In this process, the stages of bioinformatics FMs are outlined from the earliest presentations to current versions.

BIOINFORMATICS FM IN GENOMICS

Using Transformer to decode DNA's language has gained attention for deciphering biological functions through universal genetic codes, which explain DNA's translation into proteins. DNABERT [20] captures global and transferrable insights into genomic DNA sequences by using a transformer. Using Nucleotide Transformer [25], foundational language models can be built and pre-trained across genomic datasets. In DNABERT-2 [26], byte pair encodings are modified for improved computational efficiency,

ORIGINAL UNEDITED MANUSCRIPT

Downloaded from https://academic.oup.com/nsr/advance-article/doi/10.1093/nsr/nwaf028/7979309 by guest on 08 February 2025

and input length constraints are addressed using multiple strategies. To adapt to novel tasks, HyenaDNA [27] leverages longer context length and a sequence length scheduling technique. From molecular level to genomic scale, Evo [28] is a long-context foundation model that facilitates both predictive and generative tasks. VQDNA [29] has redefined genome tokenization into a holistic system based on data patterns using VQ-VAE for learning genome vocabulary. Pre-trained FMs have been trained on multi-species datasets, and used to predict promoters, enhancers, transcription factor binding sites, and cis-regulatory elements. Bioinformatics FMs in genomics are shown in Table 1. The long sequence length of biological sequences presents many challenges during the training process, making these models unable to address some biological problems. As large-scale sequence modeling, advances in biology and genomics have been made rapidly. Caduceus [30] represents the inaugural RC equivariant bi-directional long-range DNA foundation model, demonstrating superior performance to its predecessors in the realm of long-range models.

Genome-wide variant effects prediction

Mutations in DNA sequences play a significant role in contributing to species diversity. Genome-Wide Association Studies (GWAS) provides an essential amount of biological insight across a wide range of species. AI architectures have evolved to accommodate the complexity of genomic data and the nuances of high-dimensional modalities available for measuring the genome. DeepSEA [31] learns noncoding variant effects on DNA sequence alone, outperforming supervised deep learning models. In the last decade, CNN has dominated deep learning models of DNA sequence. Genomic Pre-trained Network (GPN) [32] investigates effects of genome-wide variants by training models on DNA sequences. Unlike conventional GWAS methods, GPN demonstrates exceptional proficiency in forecasting the impact of rare variants. A number of foundational DNA sequence language models, including DNABERT, DNABERT-2, and Nucleotide Transformer, also predict variants from DNA sequences. Collectively, these advances enhance our understanding of how DNA sequence mutations produce biological diversity.

DNA cis-regulatory regions prediction

In the regulation of gene expression, cis-regulatory sequences, including enhancers and promoters, play a pivotal role and design tissue-specific elements. To comprehend their functions and their associations with diseases, identifying these sequences in DNA is an essential challenge. Enformer [33] predicts gene expression and promoter-enhancer interactions by utilizing a large receptive field, for identifying cis-regulatory regions and offering valuable insights into their functions. A new transfer learning approach based on DNABERT, iEnhancer-BERT [34], facilitates enhancer prediction by utilizing an innovative DNABERT algorithm. In contrast to conventional fine-tuning methods, iEnhancer-BERT applies a CNN layer to classify output from the transformer encoder layers. Thus, biological sequences are now being recognized as the natural language of computational modeling. Furthermore, DeepSEED [35] integrates expert knowledge with learning methodologies to design synthesized promoters that are effective for synthetic promotion.

DNA methylation identification

A fundamental biological process is DNA methylation, which regulates gene expression epigenetically. Various medical conditions are associated with this process, which can also serve as a marker for metagenomic binning. AI models have advanced our understanding of DNA methylation in a variety of biological processes. Currently, iDNA-ABT [36], iDNA-ABF [37], and ccsmeth [38] serve as versatile predictors for a range of methylation predictions, including 6-methyadenine (6mA), 5-hydroxymethylcytosine (5hmC), and 4-methylcytosine (4mC). In iDNA-ABT, Transductive Information Maximization (TIM) is used in conjunction with adaptive embedding, but its potential for detecting DNA methylation patterns is yet to be explored. iDNA-ABF employs a multi-scale architecture instead of a single tokenizer. Based on tokenization, BERT encoders are able to extract diverse embeddings to produce the final evolutionary output. Furthermore, ccsmeth detects haplotype-aware methylation using nanopore sequencing data and PacBio CCS sequencing data, and makes use of the symmetry and aggregation characteristics of 5mC sites for prediction.

ORIGINAL UNEDITED MANUSCRIPT

Downloaded from https://academic.oup.com/nsr/advance-article/doi/10.1093/nsr/nwaf028/7979309 by guest on 08 February 2025

Table 1. Bioinformatics FMs in genomics.

Model	Architecture	Description
- LANGUAGE FMS -		
DNABERT *Bioinform, 2021*	• Transformer (Encoder-only) • Pretrain + Fine-tuning • 110 million parameters	- Promoter regions prediction - TF binding sites identification - Canonical splice sites recognition - Functional genetic variants identification
iDNA-ABT *Bioinform, 2021*	• Transformer (Encoder-only) • No Pretrain • 1.6 million parameters	- 4mC and non-4mC - 5hmC and non-5hmC - 6mA and non-6mA
iDNA-ABF *Genome Biol, 2022*	• Transformer (Encoder-only) • Pretrain + Fine-tuning • 110 million parameters	- 4mC and non-4mC - 5mC and non-5mC - 6mA and non-6mA
DNABERT-2 *ICLR, 2023*	• Transformer (Encoder-only) • Pretrain + Fine-tuning • 117 million parameters	- Core/Proximal promoter prediction - Epigenetic marks prediction - TF binding sites prediction - Splice sites prediction
Nucleotide Transformer *bioRxiv, 2023*	• Transformer (Encoder-only) • Pretrain + Fine-tuning • 500 million parameters	- Enhancers prediction - Promoters prediction - Epigenetic marks predition - Splice sites prediction
HyenaDNA *NeurIPS, 2024*	• Hyena (Decoder-only) • Pretrain + Fune-tuning • 1.6 million parameters	- Enhancers prediction - Promoters prediction - Epigenetic marks predition - Splice sites prediction
Evo *bioRxiv, 2024*	• StripedHyena (Decoder-only) • Pretrain + Fine-tuning • 7 billion parameters	- Fitness effects prediction - Essential genes prediction - prokaryotic promoter-RBS pairs
- VISION FMS -		
VQDNA *ICML, 2024*	• VQ-VAE • Pretrain + Fine-tuning • 110 million parameters	- Promoter detection - Core promoter detection - Splice site prediction - TF binding site prediction
- MULTIMODAL FMS -		
Enformer *Nat Methods, 2021*	• CNN + Transformer • No Pretrain • 228 million parameters	- Activating/repressive mutations - Conserved and non-conserved enhancers - Arbitrary sequences predictions
GPN *PNAS, 2023*	• CNN + Transformer • Pretrain + Fine-tuning • 65 million parameters	- Genome-wide variant effects - Genome structure - DNA motifs
DeepSEED *Nat Commun, 2023*	• cGAN + DenseNet-LSTM • Pretrain + Fine-tuning • 130 million parameters	- DNA sequence generation - Constitutive promoters design - Dox-inducible promoters design - IPTG-inducible promoters design

ORIGINAL UNEDITED MANUSCRIPT

Downloaded from https://academic.oup.com/nsr/advance-article/doi/10.1093/nsr/nwaf028/7979309 by guest on 08 February 2025

BIOINFORMATICS FM IN TRANSCRIPTOMICS

The advancement of BERT-based language models tailored for RNA sequences exhibiting reduced conservation has facilitated the emergence of significant RNA foundation models, such as RNA-FM [39] and RNA-MSM [40]. RNA-FM predicts 2D/3D structures based on self-supervised learning, capturing a variety of structural information that provides a comprehensive understanding of RNA sequence features. RNA-MSM utilizes homologous sequences from RNAcmap, which excels at mapping base pairing probabilities and solvent accessibility to 2D base pairing probabilities. Furthermore, several RNA generative models have recently been proposed by generative AI technology, such as RfamGen [41] and GenerRNA [42]. In RfamGen, alignment information and consensus secondary structure data are explicitly integrated into deep generative models to facilitate the design of RNA family sequences. GenerRNA represents a large-scale model that can be employed for the automation of RNA design. Various RNA sequence, structure and function tasks can be fine-tuned using PTMs. Bioinformatics FMs in transcriptomics are shown in Table 2.

RNA secondary structure prediction

In molecular biology, RNA secondary structure prediction is a significant challenge that requires improving structure prediction models and better understanding RNA folding. A transformer model, tokens and position embeddings, and pre-trained tasks are all integral parts of RNABERT [43]. RNABERT predicts secondary structures, classifies RNA families, and annotates uncharacterized transcripts, thereby elucidating RNA structural properties.

RNA splice sites prediction

Eukaryotic organisms rely on RNA splicing for post-transcriptional gene expression. Through the development of PTMs known as SpliceBERT [44], researchers have made significant advances in sequence-based modeling of RNA splicing. Apart from capturing RNA splicing dynamics, SpliceBERT also enables the identification of splice-disrupting variants, which can be prioritized according to their impact on output. Therefore, researchers are able to gain insight into genetic variations influencing RNA splicing, facilitating effective identification and prioritization of potentially significant variations.

RNA modification detection

Biological processes rely on modifications to RNA during post-transcription. In gene expression regulation, N7-methylguanosine (m7G) and 2'-O-methylation (Nm) RNA modifications represent widespread post-transcriptional modifications in various cellular processes. Using transformer architecture and stacking ensemble techniques, BERT-m7G [45] is a transformative computational tool for precisely pinpointing m7G sites, which is advantageous over labor-intensive experimental approaches. BERT-m7G enables us to uncover post-transcriptional modifications and better understand how m7G affects gene expression. Bert2Ome [46] provides profound insights into the underlying biological mechanisms by directly inferring 2'-O-methylation modification sites. Bert2Ome uses an integrated BERT-based model and CNN to investigate intricate relationships between modifications and RNA sequence content.

BIOINFORMATICS FM IN PROTEOMICS

Proteins have a pivotal role in constructing and maintaining vital processes in life. Protein research has experienced a substantial surge in data accumulation as the field has advanced. Structures of proteins determine how they interact with other molecules and how they function. LLMs provide an effective means of extracting pertinent and valuable information from extensive data sets. ProteinBERT [47] excels at predicting major post-translational modifications, which can be attributed to the incorporation of GO annotation prediction tasks. ProteinBERT has outperformed other deep learning models with larger parameters on various benchmarks covering diverse protein properties. The earliest protein pre-trained method that integrates external knowledge graphs is OntoProtein [48]. Aside from inheriting the strong ability of pre-trained protein language models, the knowledge embedding object also extracts biology knowledge from the knowledge graph. OntoProtein uses generative models to streamline protein downstream tasks. Bioinformatics FMs in proteomics are shown in Table 3. As part of the evaluation of deep learning models in protein science [49], numerous applications and performance characteristics of proteomics FMs are presented, including protein structure classification and enzyme function prediction. Additionally, Critical Assessment of Protein Structure Prediction (CASP) aims to objectively test the structure prediction methods of research groups from around

ORIGINAL UNEDITED MANUSCRIPT

Downloaded from https://academic.oup.com/nsr/advance-article/doi/10.1093/nsr/nwaf028/7979309 by guest on 08 February 2025

Downloaded from https://academic.oup.com/nsr/advance-article/doi/10.1093/nsr/nwaf028/7979309 by guest on 08 February 2025

Table 2. Bioinformatics FMs in transcriptomics.

Model	Architecture	Description
- LANGUAGE FMS -		
RNABERT *NAR Genom Bioinform, 2022*	• Transformer • Pretrain + Finetune	- RNA families classification - Novel transcripts annotation - RNA secondary structure prediction
RNA-FM *arXiv, 2022*	• Transformer • Pretrain + Finetune • 23 million parameters	- Gene expression regulation - SARS-CoV-2 genome evolution - RNA secondary structure prediction
SpliceBERT *bioRxiv, 2023*	• Transformer • Pretrain + Finetune • 19.4 million parameters	- Variant effects on splicing - Cross-species splice site prediction - Human genome branchpoint prediction
RNA-MSM *Nucleic Acids Res, 2024*	• Transformer • Pretrain + Finetune	- RNA solvent accessibility - RNA secondary structure prediction
GenerRNA *bioRxiv, 2024*	• Transformer (Decoder-only) • Pretrain + Finetune	- De novo RNA generation - RNA generation with specific properties
- VISION FMS -		
RfamGen *Nat Methods, 2024*	• VAE + Covariance Model • No Pretrain	- Functional RNA family generation - RNA family sequences representation
- MULTIMODAL FMS -		
Bert2Ome *TCBB, 2023*	• CNN + Transformer • Pretrain + Finetune • 110 million parameters	- 2-O-methylation sites prediction

ORIGINAL UNEDITED MANUSCRIPT

the world. It is feasible for CASPers to evaluate where future efforts could be most effectively directed by categorizing various themes.

Protein structure prediction

Functionality and interaction of proteins are closely related to their structure. Deep learning has gradually improved prediction accuracy and computational speed in predicting protein structures. MSA Transformer [50] constructs a protein language model from MSA. Masked Language Model (MLM) objectives are used to build the model across many protein families. According to the experience with BERT, when it comes to predicting secondary structure or contact, it appears that a model with more parameters is easier to use. ProtTrans [51] appears to be the only model with more parameters than most other models. Furthermore, ProtTrans has made tremendous progress in the prediction of per-residue structure. TAPE [52] establishes a standardized evaluation system for protein transfer learning. Five distinct problems are included in the task set, including protein structure prediction, fluorescence landscape prediction, stability landscape prediction, and protein design. With up to 15 billion parameters, ESM2 [53] trained transformer protein language models for widespread protein downstream applications. A protein structure predictor, ESMFold, developed later by the ESM2 team, demonstrates accuracy that is nearly comparable to alignment-based approaches, while significantly improving processing speed. As the model was scaled up, insights regarding the atomic-level structure began to emerge. PeSTo [54] is a parameter-free geometric deep learning approach designed to identify proteins binding to others. Recently, AlphaFold3 [24] has been developed and can accurately predict protein complexes with less emphasis on co-evolutionary information.

Protein sequence generation

Protein generation is widely applied to drug development and protein engineering. In order to form stable three-dimensional structures, it is hoped that the generated sequences may have good foldability. In addition, it is expected that the desired proteins have specific functional properties, such as enzyme activity. In the field of protein generation, the advancement of LLMs and the incorporation of conditional models has significantly progressed. With ProtGPT2 [55], protein amino acid propensities are generated according to natural principles, modeled after the impressive achievements of Transformer-based language models. Several globular characteristics that correspond to natural proteins are observed in ProtGPT2-generated proteins, according to analyses involving disorder and secondary structure prediction. ZymCTRL language model [56] generates artificial enzymes conditionally upon prompts from the Enzyme Commission. Generated sequences are globular, ordered, and distanced from known protein spaces, and they perform their intended functions. A new algorithm, ProGen [57], integrates UniprotKB Keywords into conditional tags and generates proteins with desired structural properties.

Protein evolution and mutation detection

Protein sequences and structures undergo changes during biological evolution. In order to produce functional diversity in proteins, evolution and mutation play a vital role. It has been suggested that protein language models can effectively predict evolutionary changes and mutations. With DeepSequence [58], a probabilistic model is learned across protein families, and it is superior to existing methods that use evolutionary data to predict mutation effects. It captures conservation in biological data and uses Evidence Lower Bound to score mutations. A new model, UniRep [59], is developed using Long Short-Term Memory (LSTM) to detect remote homologies and mutation effects. EVOLVEpro [60] outperforms existing methodologies, achieving improvements of up to 100-fold in targeted properties across six proteins within the domains of RNA production, genome editing, and antibody binding applications. These findings underscore the benefits of few-shot active learning with minimal experimental data compared to zero-shot predictions.

BIOINFORMATICS FM IN DRUG DISCOVERY

For computer-aided drug discovery, expert knowledge algorithms are used to screen drug molecules, their lead compounds, and their interactions with target molecules. A new approach to molecular fingerprinting, SMILES-BERT [61], departs from knowledge-based molecular fingerprints as input. To represent molecules, SMILES sequences are encoded based on a BERT-based model. Compared to previous models reliant on molecular fingerprints, this approach produced superior results across multiple downstream predictions of molecular properties. Through the Baidu PaddlePaddle platform, X-

ORIGINAL UNEDITED MANUSCRIPT

Downloaded from https://academic.oup.com/nsr/advance-article/doi/10.1093/nsr/nwaf028/7979309 by guest on 08 February 2025

Table 3. Bioinformatics FMs in proteomics.

Model	Architecture	Description
- LANGUAGE FMS -		
UniRep *Nat Methods, 2019*	• mLSTM • Pretrain + Finetune • 18.2 million parameters	- Protein engineering - Remote homologies detection - Mutation effects Identification
MSA Transformer *ICML, 2021*	• Transformer • Pretrain + Finetune • 100 million parameters	- Contact prediction - Secondary structure prediction
ProtTrans *IEEE TPAMI, 2021*	• Transformer • Pretrain + Finetune • 11 billion parameters	- Protein subcellular localization - Secondary structure prediction
ProteinBERT *Bioinform, 2022*	• Transformer • Pretrain + Finetune • 16 million parameters	- Evolutionary: remote homology - Engineering: fluorescence, stability - Secondary structure prediction
ProtGPT2 *Nat Commun, 2022*	• Autoregressive Transformer • Pretrain + Finetune • 738 million parameters	- Protein sequence generation
ZymCTRL *NeurIPS, 2022*	• Transformer • Pretrain + Finetune • 700 million parameters	- Enzymes generation
ESM2 *Science, 2023*	• Transformer • Pretrain + Finetune • 15 billion parameters	- Contact prediction - Protein-protein interactions - Evolutionary: remote homology - Engineering: fluorescence, stability - Secondary structure prediction
ProGen *Nat Biotechnol, 2023*	• Autoregressive Transformer • Pretrain + Finetune • 1.2 billion parameters	- Protein sequence design
- MULTIMODAL FMS -		
OntoProtein *ICLR, 2022*	• BERT + Knowledge Graph • Pretrain + Finetune	- Contact prediction - Protein-protein interactions - Evolutionary: remote homology - Engineering: fluorescence, stability - Protein function prediction: GO - Secondary structure prediction
AlphaFold3 *Nature, 2024*	• Diffusion • No Pretrain	- Protein-ligand interactions - Protein-nucleic acid interactions - Antibody-antigen prediction - Protein structure prediction

ORIGINAL UNEDITED MANUSCRIPT

Downloaded from https://academic.oup.com/nsr/advance-article/doi/10.1093/nsr/nwaf028/7979309 by guest on 08 February 2025

MOL [62] uses a pre-training model for molecular understanding SMILES, fine-tuning downstream molecular analysis tasks, such as predicting molecular properties, analyzing chemical reactions, predicting drug-drug interactions, and optimizing molecules. Bioinformatics FMs in drug discovery are shown in Table 4. For the evaluation of drug FMs, ADMETlab 2.0 [63] has been developed as a web-based system for ADMET, enhancing early drug-likeness evaluation and accelerating drug discovery. A total of 288,967 entries are contained in the ADMET database, where four functions are available for users to easily analyze six types of drug-likeness, predict 31 ADMET endpoints, and perform systematic evaluations and database/similarity searching. Various aspects of physicochemical, medicinal, and ADME properties, as well as toxicity endpoints and toxicophore rules, are evaluated by ADMET for drug discovery. These metrics include 17 physicochemical properties, 13 medicinal chemistry properties, 23 ADME properties, and 8 toxicophore rules.

Drug-like molecular properties prediction

In drug discovery, PTMs determine molecular properties in downstream tasks, such as Absorption, Distribution, Metabolism, Excretion and Toxicology (ADMET) and Pharmacokinetics (PK). K-BERT [64] differs from BERT by adopting three distinct pre-trained tasks as part of its pre-training phase, which goes beyond mere discovery of the SMILES paradigm to understand its essence. Masked Atoms Modeling and Triplet Masked Contrastive Learning tasks are introduced in Mole-BERT [65], a graph-based pre-training neural network based on BERT. A network can acquire a comprehensive understanding of molecular graph 'language' through these tasks. Using self-supervised learning, KPGT [66] is pretrained for the Line Graph Transformer. The molecular graph is processed into a molecular line graph, and molecular fingerprints are used as additional knowledge, resulting in better prediction ability in downstream tasks such as molecular properties prediction. To train molecular property prediction models, researchers have gradually adopted the large model + contrast learning paradigm due to the increasing prominence of contrastive learning. In MolCLR [67], a contrastive learning pre-training architecture, the data from one molecular graph before and after enhancement is treated as a positive sample, while data from different molecular graphs is considered a negative sample. MoleculesSTM [68], constructs a multi-modal molecular text pre-training model with two branches for molecular prediction, which reduces the representation distance between chemical structure and text description.

Drug-like molecules generation

Virtual screening libraries usually contain a few compounds, not entire drug-like chemicals. As part of MolGPT [69], an additional training task is included to facilitate conditional prediction. As well as being capable of generating innovative and effective molecules, the model also captures specific statistical characteristics within the dataset. Recently, researchers have introduced target protein information into molecular generation to identify potential target molecules. In Pocket2Mol [70], chemical constraints are captured through an E(3)-equivariant generative model. Using a neural network architecture with E(3)-equivariant, protein pockets and molecular fragments can be extracted more accurately. Using a conditional deep generative model, PMDM [71] can efficiently generate 3D molecules that are highly affinable to specific proteins. In order to preserve the geometric properties of molecules, the system uses a dual diffusion strategy that captures both local and global interactions between atoms as well as a dynamic kernel that is equivariant. Researchers are also gradually studying multi-target molecules in addition to single-target molecules. A deep generative model, POLYGON [72], can design new polypharmacology compounds that inhibit multiple targets simultaneously using encoder-decoder architectures and reinforcement learning strategies.

Drug-target interaction identification

Drug-Target Interaction (DTI) provides valuable guidance for optimizing pharmaceutical agents. DrugBAN [73] uses Frequent Contiguous Subsequence (FCS) mining to extract high-quality substructures of targets and drugs. To explicitly learn drug-target interactions, it then constructs a bilinear attention network framework. To enhance the generalization to novel drug-target pairs, a Conditional Domain Adversarial Network (CDAN) is used to harmonize the interaction representations across various domains. In EIHGN [74], four independent GNN are used to learn node representations from four distinct atomic interactions to model complexes as heterogeneous graphs. As a result, the risk of overshadowing non-covalent interaction information during message passing is minimized. EIHGN

ORIGINAL UNEDITED MANUSCRIPT

Downloaded from https://academic.oup.com/nsr/advance-article/doi/10.1093/nsr/nwaf028/7979309 by guest on 08 February 2025

Table 4. Bioinformatics FMs in drug discovery.

Model	Architecture	Description
- LANGUAGE FMS -		
SMILES-BERT *ACM BCB, 2019*	• BERT • Pretrain + Finetune	- Molecular representation - Molecular property prediction
MolGPT *J Chem Inf Model, 2021*	• Transformer (Decoder-only) • Pretrain + Finetune • 6 million parameters	- Molecule generation via properties - Molecule generation via scaffolds
X-MOL *Sci Bull, 2022*	• Transformer • Pretrain + Finetune	- Molecular property prediction - Chemical reaction analysis - Molecule optimization
K-BERT *Brief Bioinform, 2022*	• BERT • Pretrain + Finetune • 110 million parameters	- Atom feature prediction - Molecular feature prediction
DrugBAN *Nat Mach Intell, 2023*	• FCS • No Pretrain	- Drug-target pairs prediction
- VISION FMS -		
PMDM *Nat Commun, 2024*	• EGNNs + SchNet • No pretrain	- Molecule generation of specific target
POLYGON *Nat Commun, 2024*	• VAE • No pretrain	- Molecule generation of multi-target
- GRAPH FMS -		
Mole-BERT *ICLR, 2022*	• GINs • Pretrain + Finetune	- Molecular property prediction - Drug-target affinity prediction
MolCLR *Nat Mach Intell, 2022*	• GNN + Contrastive Learning • Pretrain + Finetune	- Molecular representation - Molecular property prediction
Pocket2Mol *ICML, 2022*	• MPNN • No pretrain	- Molecule generation via 3D pockets
KPGT *Nat Commun, 2023*	• Line Graph Transformer • Pretrain + Finetune • 100 million parameters	- Molecular representation - Molecular property prediction
EIHGN *IEEE TPAMI, 2024*	• 3D GNN • No Pretrain	- Protein-ligand binding prediction
- MULTIMODAL FMS -		
MoleculesSTM *Nat Mach Intell, 2023*	• MolBART + GIN + SciBERT • Pretrain + Finetune • 120 million parameters	- Structure-text retrieval - Molecule editing

ORIGINAL UNEDITED MANUSCRIPT

Downloaded from https://academic.oup.com/nsr/advance-article/doi/10.1093/nsr/nwaf028/7979309 by guest on 08 February 2025

also decomposes affinity prediction values into the sum of non-covalent interaction forces predicted between target and drug atoms.

BIOINFORMATICS FM IN SINGLE CELLS

Single-cell RNA sequencing (scRNA-seq) technology has paved the way for numerous breakthroughs. Single-cell language models can be used to identify cell states, discover novel cell types, infer regulation networks, and integrate multi-omics data. scGPT [75] provides a unified pre-training pipeline tailored to non-sequential datasets. Through the use of stacked transformer layers and multiple heads, scGPT is capable of general-purpose pre-training and fine-tuning for specific applications, enabling learning to be transferred to downstream tasks. To infer the missing single-cell proteome from the transcriptome, scTranslator [76] proposes a large, pre-trained, generative model derived from both NLP and genetic central dogma. In scTranslator, the protein abundance is inferred from paired bulk data, then paired single-cell data, and finally from scRNA-seq datasets as a transformer-based model. scButterfly [77] learns latent factors within individual modalities to perform cross-modal translation using a dual aligned variational autoencoder and data augmentation scheme. A masked VAE is trained by scButterfly, then the latent representations are cross-modally aligned. scFoundation [78] algorithm presents a novel pre-trained method called Read-Depth-Aware (RDA) modeling. Nicheformer [79] is a transformer-based approach to learning cellular representations from dissociated single cells and transcriptomics data for many downstream applications. CELLama [80] creates cellular data embedding sentences encapsulating gene expressions and metadata. Bioinformatics FMs in single-cell multi-omics analysis are shown in Table 5. In order to evaluate single-cell foundation models, scEval [81] evaluates how hyperparameters and LLM training, which presents a summary of single-cell LLMs and their limitations, as well as possible future developments. Several single-cell LLMs were evaluated on eight tasks with 22 datasets.

Cell clustering

In order to understand the complex landscape of cellular heterogeneity within biological samples, the process of cell clustering is crucial. To learn cell embedding for clustering, the encoder and decoder structures of scFoundation [78] are transformer-based models, and only genes that are not masked are fed into the encoder. MarsGT [82] infers and identifies rare cell clusters from multi-omics data generated by single cells. MarsGT builds a multi-head attention mechanism on heterogeneous graphs of genes and cells. In scPROTEIN [83], peptide quantification uncertainty and other data problems are addressed in a unified framework through deep graph contrastive learning. A variety of downstream tasks can be performed using scPROTEIN's versatile cell embeddings.

Cell type annotation

When annotating a single cell, biological labels are assigned to each cell or cluster, typically cell type or cell state. With the remarkable success of LLMs in NLP and CV, single-cell RNA sequencing data can now be analyzed for cell type annotation. Several computational tools have emerged for annotation of scRNA-seq data using language models, including TOSICA [84] and scBERT [85]. TOSICA incorporates knowledge-based masks from GSEA into a fully connected weight matrix to create an interpretable cell type annotation method. The pre-training phase of scBERT is designed to eliminate batch effects and enhance generalizability through a comprehensive understanding of gene-gene interactions. When fine-tuning, reference datasets influence model parameters since a classifier is added to PTMs. Thus, scBERT enables the discovery of unbiased long-range interactions, data-driven annotation of cell types.

Multi-omics integration

Integration of various omics technologies offers several advantages over analyses based on individual omics data. Large models are valuable tools for finding solutions to scMulti-omics data's features-variance, sparsity, and cell heterogeneity because of their adaptability, generalization capabilities, and feature extraction abilities. As part of scMulti-omics integration tasks, scGPT [75] uses supplementary token sets to signify distinct sequencing modalities. Transformators incorporate modality tokens, either at the feature or cell level, into their output. Incorporating this intentionally prevents the transformer from highlighting features associated with the same modalities, while simultaneously undermining those associated with different modalities. In DeepMAPS [86], scMulti-omics data is integrated and mapped into biological networks by using graph transformer. Due to the fact that DeepMAPS builds a graph with nodes

ORIGINAL UNEDITED MANUSCRIPT

Downloaded from https://academic.oup.com/nsr/advance-article/doi/10.1093/nsr/nwaf028/7979309 by guest on 08 February 2025

Table 5. Bioinformatics FMs in single-cell analysis.

Model	Architecture	Description
- LANGUAGE FMS -		
scBERT *Nat Mach Intell, 2022*	• BERT • Pretrain + Fine-tuning	- Gene-gene interactions prediction - Cell type annotation - Novel cell type discovery
scMVP *Genome Biol, 2022*	• Transformer • No Pretrain	- Data imputation - Cell groups identification
scTranslator *Genome Res, 2023*	• Transformer • Pretrain + Fine-tuning	- Gene-gene interactions prediction - Gene pseudo-knockout - Cell clustering
TOSICA *Nat Commun, 2023*	• Transformer • No Pretrain	- Cell type annotation
scFoundation *Nat Methods, 2024*	• Transformer • Pretrain + Fine-tuning • 100 million parameters	- Gene expression enhancement - Single-cell drug response prediction - Tissue drug response identification
scGPT *Nat Methods, 2024*	• Transformer • Pretrain + Fine-tuning	- Cell clustering - Batch correction - Gene regulatory networks inference
mvTCR *Nat Commun, 2024*	• Transformer • No Pretrain	- Cell-level embedding - Atlas-level analysis
- VISION FMS -		
scButterfly *Nat Commun, 2024*	• VAE • Pretrain + Fine-tuning	- Cell type annotation - Poor-quality data enhancement - Integrative multi-omics analysis
MIDAS *Nat Biotechnol, 2024*	• VAE + Transfer Learning • No Pretrain	- Modality alignment - Data imputation - Batch correction
- GRAPH FMS -		
DeepMAPS *Nat Commun, 2023*	• Graph Transformer • No Pretrain	- Cell clustering - Biological network construction
SiGra *Nat Commun, 2023*	• Graph Transformer • No Pretrain	- Spatial profiles augmentation
MarsGT *Nat Commun, 2024*	• Graph Transformer • No Pretrain	- Cell clustering - Gene regulatory networks inference
scPROTEIN *Nat Methods, 2024*	• GCN + Contrastive Learning • No Pretrain	- Cell type annotation - Batch correction - Cell type annotation - Proteomic data exploration
- MULTIMODAL FMS -		
GLUE *Nat Biotechnol, 2022*	• Graph VAE • No Pretrain	- Triple-omics data integration - Integrative regulation inference

ORIGINAL UNEDITED MANUSCRIPT

Downloaded from https://academic.oup.com/nsr/advance-article/doi/10.1093/nsr/nwaf028/7979309 by guest on 08 February 2025

for genes and cells, the features of all other modes are mapped to genes. The transformer in DeepMAPS builds relations between cells and genes as well as gene-gene relations using local and global features. A cell-level embedding is created using mvTCR [87], which is easily scalable to atlase-level analysis and fits well into a standard analysis pipeline. Using separate encoders, mvTCR combines different modalities to produce a joint representation. Using an image-augmented Graph transformer, SiGra [88] reveals single-cell spatial information. Through the use of multimodalities and transcriptomics, SiGra enhances data quality and recognizes spatial domains simultaneously. GLUE [22] integrates unpaired multi-omics data and infers regulatory interactions. GLUE models cross-layer regulatory interactions explicitly by leveraging prior biological knowledge. As well as integrating triple-omics, GLUE can also handle regulating inference and annotation correction. MIDAS [89] uses a modular encoder network and a decoder network to integrate and transfer multimodal data from single cells.

DATA IN BIOINFORMATICS FM

In bioinformatics, FMs are reliant on biological data quality, which constitutes a massive amount of multi-omics data. Biological databases are commonly used in bioinformatics FMs, as shown in Table 6. In genomics, The Cancer Genome Atlas (TCGA) [90] analyzes more than 20,000 cancer samples matched to normal samples, covering 33 different cancer types; TargetFinder [91] provides a pipeline for identifying or characterizing gene targets of distal enhancers; ArrayExpress [92] contains data from high-throughput functional genomics experiments. In transcriptomics, Gene Expression Omnibus (GEO) [93] archives functional genomics data derived from microarrays and other high-throughput methodologies; Encyclopedia of DNA Elements (ENCODE) [94] is a comprehensive database of essential elements in the human genome, including proteins and RNA, as well as regulatory elements that control cell activity and gene expression. In proteomics, Universal Protein Resource (UniProt) [95] contains comprehensive protein sequences and annotations; Protein Data Bank (PDB) [96] contains sequences and 2D/3D structures of large biological molecules. In drug molecular, ChEMBL [97] is a meticulously curated database of bioactive molecules exhibiting drug-like properties; ZINC [98] is a publicly accessible database of commercially available compounds designed for virtual screening; PubChem [99] is the collection of chemical information that is freely accessible. In single-cell data, Single-cell Expression Atlas [100] serves as a comprehensive database for single-cell gene expression across various species; Human Cell Atlas [101] aims to map each cell type within the human body, thereby creating a 3D atlas of human cells.

FUTURE DIRECTIONS

Our study focuses on various applications of Bioinformatics FMs, which accurately model the intricate complexities of molecular biology. Pre-training architectures capture patterns related to source data; Fine-tuning strategies analyze task data to solve biological problems accurately. The landscape of FMs in bioinformatics is shown in Figure 3. New insights can be gained into the dynamic interaction between molecules by exploring these cutting-edge technologies. Our final objective is to discuss challenges and opportunities related to explainability of foundation models, and architectures of large-scale models.

Pre-training paradigm

The emerging methodologies for training foundation models in artificial intelligence enable the execution of specific downstream tasks, thereby allowing AI to be fine-tuned for highly specialized applications even when only a limited number of training examples are available. Currently, several studies are exploring the application of prompt learning and contrastive learning to pre-trained models within the field of bioinformatics, which require further development. Prompt learning and contrastive learning have become pivotal techniques in bioinformatics, particularly when utilized with pre-trained models to enhance both model performance and interpretability across various bioinformatics tasks. KANO [102] exploits external fundamental domain knowledge to enhance molecular contrastive learning and fine-tune learning. Microscopic atomic associations are investigated while maintaining the molecular semantics of an element-oriented knowledge graph. In fine-tuning, functional prompts are designed to elicit task-specific knowledge. PromptProtein [103] offers an innovative pre-training and fine-tuning framework based on prompt-guided training. With prompt-guided multi-task pre-training, it learns to focus on different structure levels based on multiple prompt signals. By providing downstream tasks with on-demand flexibil-

ORIGINAL UNEDITED MANUSCRIPT

Downloaded from https://academic.oup.com/nsr/advance-article/doi/10.1093/nsr/nwaf028/7979309 by guest on 08 February 2025

Table 6. Biological databases commonly used in bioinformatics FMs.

Database	Description	Web Link
- GENOMICS -		
TCGA *2.5 petabytes*	Database of cancer genome atlas varieties.	`https://www.cancer.gov/ccg/research/genome-sequencing/tcga`
TargetFinder *100 thousand*	Database of promoter-enhancer interaction pairs.	`https://github.com/shwhalen/targetfinder`
ArrayExpress *78 thousand*	Database of the high-throughput functional genomics data.	`https://www.ebi.ac.uk/biostudies/arrayexpres`
- TRANSCRIPTOMICS -		
GEO *7.3 billion*	Database of public functional genomics data.	`https://www.ncbi.nlm.nih.gov/geo/`
ENCODE *100 million*	Database of functional elements in the human genome.	`https://www.encodeproject.org/`
- PROTEOMICS -		
UniProt *245 million*	Database of protein sequences and functions.	`https://www.uniprot.org/`
PDB *222 thousand*	Database of 3D structural data for large biological molecules.	`https://www.rcsb.org/`
- DRUG -		
ChEMBL *2.4 million*	Database of bioactive compounds with drug-like properties.	`https://www.ebi.ac.uk/chembl/`
ZINC *750 million*	Database of commercially available compounds.	`https://zinc15.docking.org`
PubChem *119 million*	Database of freely available chemical information.	`https://pubchem.ncbi.nlm.nih.gov/`
- SINGLE CELL -		
Single-cell Expression Atlas *10 million*	Database of single-cell gene expression across species.	`https://www.ebi.ac.uk/gxa/sc/home`
Human Cell Atlas *61.8 million*	Database of the multi-omic human cell atlas.	`https://www.humancellatlas.org/`

ORIGINAL UNEDITED MANUSCRIPT

Downloaded from https://academic.oup.com/nsr/advance-article/doi/10.1093/nsr/nwaf028/7979309 by guest on 08 February 2025

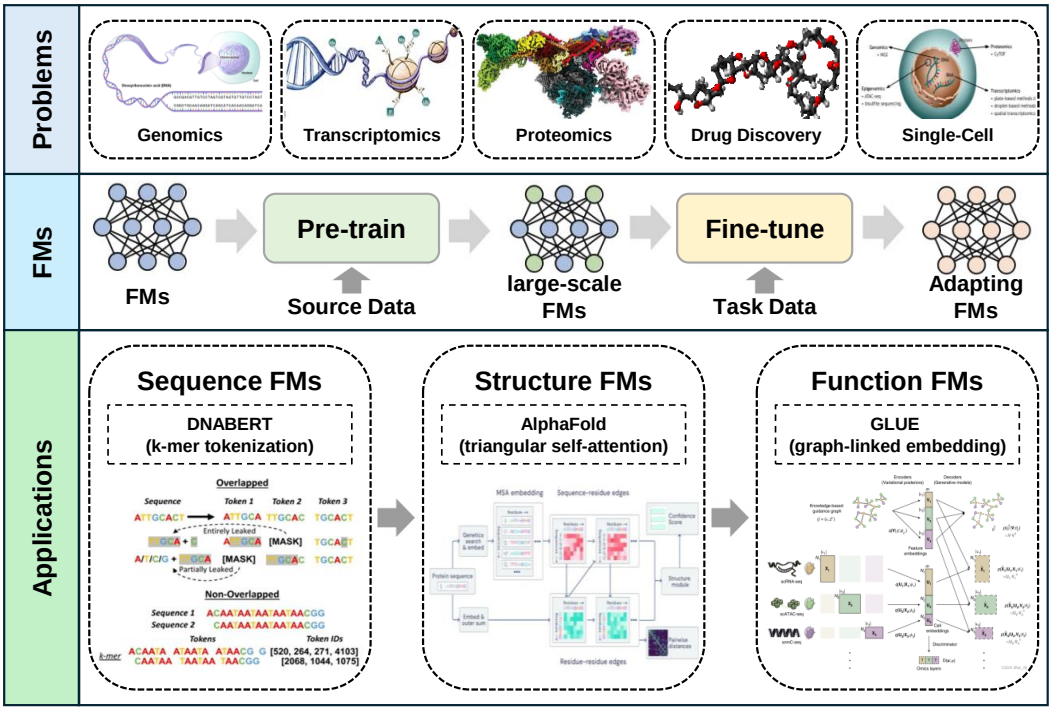

Figure 3. Landscape of bioinformatics FMs.

ity, prompt fine-tuning modules allow them to utilize respective levels of structure information.

Evaluation Framework

Several AI models have been trained on large datasets and applied to downstream applications. Foundation models offer opportunities and risks from their capabilities and technical principles to their applications and societal impacts. The scale and effectiveness of models across so many tasks encourage homogenization. Using three key components for model evaluation (models, data, and metrics), UltraEval [104] showcases itself as a user-friendly evaluation model that is lightweight, comprehensive, modular, and efficient. Some studies have also evaluated the performance of some bioinformatics research fields, including protein engineering, drug design, and analysis of single-cell multi-omics data. scBackdoor [105] has been introduced for single-cell pre-training models to assess the attack success rate. This poses a significant potential threat to single-cell research, particularly in relation to AI pre-training models that rely on open data.

Model explainability

Bioinformatics also faces challenges in providing interpretable FMs and acquiring logical evidence. For instance, computer-aided drug discovery includes docking, scoring, and screening. To generate a drug molecule, properties such as effectiveness, novelty, and similarity to existing drugs must be considered. However, existing methods lack extensive studies on actual chemical or biological experimental validation to prove their efficiency. FMs may be able to solve complex biological problems more efficiently through knowledge graphs with interpretability and explainability. The use of causal inference has been shown to enhance predictive accuracy, fairness, robustness, and explainability of NLP models by tracing causal relationships among variables. With its improved sampling efficiency, CIMI [106] provides more faithful and generalizable explanations, making it particularly suitable for large pretrained models.

Hallucination Detection

Foundation models are employed to construct comprehensive biological maps in such a diverse environment. However, there are difficulties transitioning from derivable approaches to multi-focus frameworks. For example, large language models are universal tools for a broad range of biological datasets and applications. In order to improve understanding of the cellular landscape, various model architectures are synergized and scaled up to extract meaningful features from raw data. Furthermore, LLMs can reason and an-

ORIGINAL UNEDITED MANUSCRIPT

Downloaded from https://academic.oup.com/nsr/advance-article/doi/10.1093/nsr/nwaf028/7979309 by guest on 08 February 2025

swer questions impressively, but they tend to hallucinate false results and answer questions unsubstantiated [107]. Hallucination in FMs refers to the generation of content that strays from factual reality or includes fabricated information. Current hallucination detection techniques lack accuracy, low latency, and low cost all at once. Luna [108] has been fine-tuned to detect hallucinations in RAG, which allows language models to incorporate external knowledge retrieval mechanisms to enhance their capabilities.

CONCLUSIONS

Foundation models of future artificial intelligence are able to scale up as the training data gets more sophisticated. Modifying these models further can result in outstanding performance across a variety of application domains. Furthermore, corporations have achieved remarkable progress by homogenizing models with LLMs within NLP. Due to their ability to comprehend and manipulate human language, these models are revolutionary and transformative. High throughput data are integral to these bioinformatics problems: DNA sequence in genomics, RNA sequence in transcriptomics, protein sequence and structure in proteomics, molecular SMILES in drug discovery, and multi-omics data in single cells. Deep learning mechanisms are integrated to acquire biological insights, such as CNN for protein 3D structure features, RNN for time-series single-cell RNA sequencing features, Transformer for biological sequence features, and GNN for molecular topology features. By leveraging massive biological datasets, large-scale models are pre-trained and can be applied to a variety of tasks (few-shot, zero-shot, or fine-tuned). Foundation models solve several core biological problems and various downstream tasks rapidly and efficiently.

AUTHOR CONTRIBUTIONS

Jianxin Wang contributed to conception and design; Fei Guo, Renchu Guan, Qi Liu, Xiaowo Wang and Jianxin Wang contributed to investigation, acquisition, analysis and interpretation; Fei Guo drafted the manuscript; Renchu Guan, Yaohang Li, Qi Liu, Xiaowo Wang, Can Yang and Jianxin Wang critically revised the manuscript; Jianxin Wang led supervision and project administration; All authors agree to maintain integrity and accuracy in all aspects of their work.

FUNDING

This work was supported by the National Key Research and Development Program of China (2021YFF1201200), the National Natural Science Foundation of China under Grants (62350004, 62332020, 62322215), and the Project of Xiangjiang Laboratory (23XJ01011).

Conflict of interest statement. None declared.

REFERENCES

1. Bommasani R, Hudson DA, Adeli E *et al.* On the opportunities and risks of foundation models. ArXiv: 2108.07258.
2. Mikolov T, Sutskever I, Chen K *et al.* Distributed representations of words and phrases and their compositionality. *Proceedings of the Advances in Neural Information Processing Systems*, volume 26 (2013) 3111–9.
3. Vaswani A, Shazeer N, Parmar N *et al.* Attention is all you need. *Proceedings of the Advances in Neural Information Processing Systems*, volume 30 (2017) .
4. Devlin J, Chang MW, Lee K *et al.* Bert: Pre-training of deep bidirectional transformers for language understanding. *Proceedings of the Conference of the North American chapter of the Association for Computational Linguistics: Human Language Technologies* (2019) 4171–86.
5. Brown T, Mann B, Ryder N *et al.* Language models are few-shot learners. *Proceedings of the Advances in Neural Information Processing Systems*, volume 33 (2020) 1877–901.
6. Krizhevsky A, Sutskever I and Hinton GE. Imagenet classification with deep convolutional neural networks. *Proceedings of the Advances in Neural Information Processing Systems*, volume 25 (2012) .
7. He K, Zhang X, Ren S *et al.* Deep residual learning for image recognition. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (2016) 770–8.
8. Kirillov A, Mintun E, Ravi N *et al.* Segment anything. *Proceedings of the IEEE/CVF International Conference on Computer Vision* (2023) 4015–26.
9. Gilmer J, Schoenholz SS, Riley PF *et al.* Neural message passing for quantum chemistry. *Proceedings of the International Conference on Machine Learning* (2017) 1263–72.
10. Xu K, Hu W, Leskovec J *et al.* How powerful are graph neural networks? *Proceedings of the International Conference on Learning Representations* (2019) .
11. Ying C, Cai T, Luo S *et al.* Do transformers really perform badly for graph representation? *Proceedings of the Advances in Neural Information Processing Systems*, volume 34 (2021) 28877–88.
12. Edge D, Trinh H, Cheng N *et al.* From local to global: A graph rag approach to query-focused summarization. ArXiv: 2404.16130.
13. Dosovitskiy A, Beyer L, Kolesnikov A *et al.* An image is worth 16x16 words: Transformers for image recognition at scale. *Proceedings of the International Conference on Learning Representations* (2021) .

ORIGINAL UNEDITED MANUSCRIPT

Downloaded from https://academic.oup.com/nsr/advance-article/doi/10.1093/nsr/nwaf028/7979309 by guest on 08 February 2025

14. Radford A, Kim JW, Hallacy C *et al.* Learning transferable visual models from natural language supervision. *Proceedings of the International Conference on Machine Learning* (2021) 8748–63.
15. Chen Z, Wei L and Gao G. Foundation models for bioinformatics. *Quant. Biol.* 2024; **12**: 339–44.
16. Sarumi OA and Heider D. Large language models and their applications in bioinformatics. *Computat. Struct. Biotech.* 2024; **23**: 3498–505.
17. Guo Z, Liu J, Wang Y *et al.* Diffusion models in bioinformatics and computational biology. *Nature Rev. Bioeng.* 2024; **2**: 136–54.
18. Li Q, Hu Z, Wang Y *et al.* Progress and opportunities of foundation models in bioinformatics. *Brief. Bioinform.* 2024; **25**: bbae548.
19. Moor M, Banerjee O, Abad ZSH *et al.* Foundation models for generalist medical artificial intelligence. *Nature* 2023; **616**: 259–65.
20. Ji Y, Zhou Z, Liu H *et al.* Dnabert: Pre-trained bidirectional encoder representations from transformers model for dna-language in genome. *Bioinformatics* 2021; **37**: 2112–20.
21. Senior A, Evans R, Jumper J *et al.* Improved protein structure prediction using potentials from deep learning. *Nature* 2020; **577**: 706–10.
22. Cao ZJ and Gao G. Multi-omics single-cell data integration and regulatory inference with graph-linked embedding. *Nat. Biotechnol.* 2022; **40**: 1458–66.
23. Jumper J, Evans R, Pritzel A *et al.* Highly accurate protein structure prediction with alphafold. *Nature* 2021; **596**: 583–9.
24. Abramson J, Adler J, Dunger J *et al.* Accurate structure prediction of biomolecular interactions with alphafold 3. *Nature* 2024; 1–3.
25. Dalla-Torre H, Gonzalez L, Mendoza-Revilla J *et al.* The nucleotide transformer: Building and evaluating robust foundation models for human genomics. BioRxiv: 2023–01.
26. Zhou Z, Ji Y, Li W *et al.* Dnabert-2: Efficient and effective foundation model for multi-species genome. *Proceedings of the International Conference on Learning Representations* (2024) .
27. Nguyen E, Poli M, Faizi M *et al.* Hyenadna: Long-range genomic sequence modeling at single nucleotide resolution. *Proceedings of the Advances in Neural Information Processing Systems*, volume 36 (2024) .
28. Nguyen E, Poli M, Durrant MG *et al.* Sequence modeling and design from molecular to genome scale with evo. *Science* 2024; **386**: eado9336.
29. Li S, Wang Z, Liu Z *et al.* Vqdna: Unleashing the power of vector quantization for multi-species genomic sequence modeling. *Proceedings of the International Conference on Machine Learning* (2024) 28717–33.
30. Schiff Y, Kao CH, Gokaslan A *et al.* Caduceus: Bi-directional equivariant long-range dna sequence modeling. *Proceedings of the International Conference on Machine Learning* (2024) 43632–48.
31. Zhou J and Troyanskaya OG. Predicting effects of noncoding variants with deep learning–based sequence model. *Nat. Methods* 2015; **12**: 931–934.
32. Benegas G, Batra SS and Song YS. Dna language models are powerful predictors of genome-wide variant effects. *Proceedings of the National Academy of Sciences* 2023; **120**: e2311219120.
33. Avsec Ž, Agarwal V, Visentin D *et al.* Effective gene expression prediction from sequence by integrating long-range interactions. *Nat. Methods* 2021; **18**: 1196–203.
34. Luo H, Chen C, Shan W *et al.* ienhancer-bert: A novel transfer learning architecture based on dna-language model for identifying enhancers and their strength. *Proceedings of the International Conference on Intelligent Computing* (2022) 153–65.
35. Zhang P, Wang H, Xu H *et al.* Deep flanking sequence engineering for efficient promoter design using deepseed. *Nat. Commun.* 2023; **14**: 6309.
36. Yu Y, He W, Jin J *et al.* idna-abt: Advanced deep learning model for detecting dna methylation with adaptive features and transductive information maximization. *Bioinformatics* 2021; **37**: 4603–10.
37. Jin J, Yu Y, Wang R *et al.* idna-abf: Multi-scale deep biological language learning model for the interpretable prediction of dna methylations. *Genome Biol.* 2022; **23**: 219.
38. Ni P, Nie F, Zhong Z *et al.* Dna 5-methylcytosine detection and methylation phasing using pacbio circular consensus sequencing. *Nat. Commun.* 2023; **14**: 4054.
39. Chen J, Hu Z, Sun S *et al.* Interpretable rna foundation model from unannotated data for highly accurate rna structure and function predictions. ArXiv: 2204.00300.
40. Zhang Y, Lang M, Jiang J *et al.* Multiple sequence alignment-based rna language model and its application to structural inference. *Nucleic Acids Res.* 2024; **52**: e3.
41. Sumi S, Hamada M and Saito H. Deep generative design of rna family sequences. *Nat. Methods* 2024; **21**: 435–43.
42. Zhao Y, Oono K, Takizawa H *et al.* Generrna: A generative pre-trained language model for de novo rna design. BioRxiv: 2024–02.
43. Akiyama M and Sakakibara Y. Informative rna base embedding for rna structural alignment and clustering by deep representation learning. *NAR Genomics and Bioinformatics* 2022; **4**: lqac012.
44. Chen K, Zhou Y, Ding M *et al.* Self-supervised learning on millions of pre-mrna sequences improves sequence-based rna splicing prediction. BioRxiv: 2023–01.
45. Zhang L, Qin X, Liu M *et al.* Bert-m7g: A transformer architecture based on bert and stacking ensemble to identify rna n7-methylguanosine sites from sequence information. *Comput. Math. Method. M.* 2021; **2021**: 7764764.
46. Soylu NN and Sefer E. Bert2ome: Prediction of 2-o-methylation modifications from rna sequence by transformer architecture based on bert. *ACM T. Comput. Bi.* 2023; **20**: 2177–89.
47. Brandes N, Ofer D, Peleg Y *et al.* Proteinbert: A universal deep-learning model of protein sequence and function. *Bioinformatics* 2022; **38**: 2102–10.
48. Zhang N, Bi Z, Liang X *et al.* Ontoprotein: Protein pretraining with gene ontology embedding. *Proceedings of the International Conference on Learning Representations* (2022) .
49. Hu B, Tan C, Wu L *et al.* Advances of deep learning in protein science: A comprehensive survey. ArXiv: 2403.05314.
50. Rao RM, Liu J, Verkuil R *et al.* Msa transformer. *Proceedings of the International Conference on Machine Learning* (2021) 8844–56.

ORIGINAL UNEDITED MANUSCRIPT

Downloaded from https://academic.oup.com/nsr/advance-article/doi/10.1093/nsr/nwaf028/7979309 by guest on 08 February 2025

51. Elnaggar A, Heinzinger M, Dallago C *et al.* Prottrans: Toward understanding the language of life through self-supervised learning. *IEEE T. Pattern Anal.* 2021; **44**: 7112–27.
52. Rao R, Bhattacharya N, Thomas N *et al.* Evaluating protein transfer learning with tape. *Proceedings of the Advances in Neural Information Processing Systems*, volume 32 (2019) .
53. Lin Z, Akin H, Rao R *et al.* Evolutionary-scale prediction of atomic-level protein structure with a language model. *Science* 2023; **379**: 1123–30.
54. Krapp LF, Abriata LA, Cortés Rodriguez F *et al.* Pesto: Parameter-free geometric deep learning for accurate prediction of protein binding interfaces. *Nat. Commun.* 2023; **14**: 2175.
55. Ferruz N, Schmidt S and Höcker B. Protgpt2 is a deep unsupervised language model for protein design. *Nat. Commun.* 2022; **13**: 4348.
56. Munsamy G, Lindner S, Lorenz P *et al.* Zymctrl: A conditional language model for the controllable generation of artificial enzymes. *Proceedings of the NeurIPS Machine Learning in Structural Biology Workshop* (2022) .
57. Madani A, Krause B, Greene ER *et al.* Large language models generate functional protein sequences across diverse families. *Nat. Biotechnol.* 2023; **41**: 1099–106.
58. Riesselman AJ, Ingraham JB and Marks DS. Deep generative models of genetic variation capture the effects of mutations. *Nat. Methods* 2018; **15**: 816–22.
59. Alley EC, Khimulya G, Biswas S *et al.* Unified rational protein engineering with sequence-based deep representation learning. *Nat. Methods* 2019; **16**: 1315–22.
60. Jiang K, Yan Z, Di Bernardo M *et al.* Rapid in silico directed evolution by a protein language model with evolvepro. *Science* 2024; eadr6006.
61. Wang S, Guo Y, Wang Y *et al.* Smiles-bert: Large scale unsupervised pre-training for molecular property prediction. *Proceedings of the ACM Conference on Bioinformatics, Computational Biology, and Health Informatics* (2019) 429–36.
62. Xue D, Zhang H, Xiao D *et al.* X-mol: Large-scale pre-training for molecular understanding and diverse molecular analysis. *Science Bulletin (Beijing)* 2022; **67**: 899–902.
63. Xiong G, Wu Z, Yi J *et al.* Admetlab 2.0: An integrated online platform for accurate and comprehensive predictions of admet properties. *Nucleic Acids Res.* 2021; **49**: W5–14.
64. Wu Z, Jiang D, Wang J *et al.* Knowledge-based bert: A method to extract molecular features like computational chemists. *Brief. Bioinform.* 2022; **23**: bbac131.
65. Xia J, Zhao C, Hu B *et al.* Mole-bert: Rethinking pre-training graph neural networks for molecules. *Proceedings of the International Conference on Learning Representations* (2023) .
66. Li H, Zhang R, Min Y *et al.* A knowledge-guided pre-training framework for improving molecular representation learning. *Nat. Commun.* 2023; **14**: 7568.
67. Wang Y, Wang J, Cao Z *et al.* Molecular contrastive learning of representations via graph neural networks. *Nat. Mach. Intell.* 2022; **4**: 279–87.
68. Liu S, Nie W, Wang C *et al.* Multi-modal molecule structure–text model for text-based retrieval and editing. *Nat. Mach. Intell.* 2023; **5**: 1447–57.
69. Bagal V, Aggarwal R, Vinod P *et al.* Molgpt: Molecular generation using a transformer-decoder model. *J. Chem. Inf. Model.* 2021; **62**: 2064–76.
70. Peng X, Luo S, Guan J *et al.* Pocket2mol: Efficient molecular sampling based on 3d protein pockets. *Proceedings of the International Conference on Machine Learning* (2022) 17644–55.
71. Huang L, Xu T, Yu Y *et al.* A dual diffusion model enables 3d molecule generation and lead optimization based on target pockets. *Nat. Commun.* 2024; **15**: 2657.
72. Munson BP, Chen M, Bogosian A *et al.* De novo generation of multi-target compounds using deep generative chemistry. *Nat. Commun.* 2024; **15**: 3636.
73. Bai P, Miljković F, John B *et al.* Interpretable bilinear attention network with domain adaptation improves drug–target prediction. *Nat. Mach. Intell.* 2023; **5**: 126–36.
74. Yang Z, Zhong W, Lv Q *et al.* Interaction-based inductive bias in graph neural networks: Enhancing protein–ligand binding affinity predictions from 3d structures. *IEEE T. Pattern Anal.* 2024; **46**: 8191–208.
75. Cui H, Wang C, Maan H *et al.* scgpt: Toward building a foundation model for single-cell multi-omics using generative ai. *Nat. Methods* 2024; 1–11.
76. Liu X, Shen Q and Zhang S. Cross-species cell-type assignment from single-cell rna-seq data by a heterogeneous graph neural network. *Genome Res.* 2023; **33**: 96–111.
77. Cao Y, Zhao X, Tang S *et al.* Chen, shengquan. *Nat. Commun.* 2024; **15**: 2973.
78. Hao M, Gong J, Zeng X *et al.* Large-scale foundation model on single-cell transcriptomics. *Nat. Methods* 2024; 1–11.
79. Schaar AC, Tejada-Lapuerta A, Palla G *et al.* Nicheformer: A foundation model for single-cell and spatial omics. BioRxiv: 2024–04.
80. Choi H, Park J, Kim S *et al.* Cellama: Foundation model for single cell and spatial transcriptomics by cell embedding leveraging language model abilities. BioRxiv: 2024–05.
81. Zhao H, Liu T, Li K *et al.* Evaluating the utilities of large language models in single-cell data analysis. BioRxiv: 2023.
82. Wang X, Duan M, Li J *et al.* Marsgt: Multi-omics analysis for rare population inference using single-cell graph transformer. *Nat. Commun.* 2024; **15**: 338.
83. Li W, Yang F, Wang F *et al.* scprotein: A versatile deep graph contrastive learning framework for single-cell proteomics embedding. *Nat. Methods* 2024; **21**: 623–34.
84. Chen J, Xu H, Tao W *et al.* Transformer for one stop interpretable cell type annotation. *Nat. Commun.* 2023; **14**: 223.
85. Yang F, Wang W, Wang F *et al.* scbert as a large-scale pretrained deep language model for cell type annotation of single-cell rna-seq data. *Nat. Mach. Intell.* 2022; **4**: 852–66.
86. Ma A, Wang X, Li J *et al.* Single-cell biological network inference using a heterogeneous graph transformer. *Nat. Commun.* 2023; **14**: 964.
87. Drost F, An Y, Bonafonte-Pardàs I *et al.* Multi-modal generative modeling for joint analysis of single-cell t cell receptor and gene expression data. *Nat. Commun.* 2024; **15**: 5577.

ORIGINAL UNEDITED MANUSCRIPT

Downloaded from https://academic.oup.com/nsr/advance-article/doi/10.1093/nsr/nwaf028/7979309 by guest on 08 February 2025

88. Tang Z, Li Z, Hou T *et al.* Sigra: Single-cell spatial elucidation through an image-augmented graph transformer. *Nat. Commun.* 2023; **14**: 5618.
89. He Z, Hu S, Chen Y *et al.* Mosaic integration and knowledge transfer of single-cell multimodal data with midas. *Nat. Biotechnol.* 2024; 1–12.
90. Weinstein JN, Collisson EA, Mills GB *et al.* The cancer genome atlas pan-cancer analysis project. *Nat. Genet.* 2013; **45**: 1113–20.
91. Whalen S, Truty RM and Pollard KS. Enhancer–promoter interactions are encoded by complex genomic signatures on looping chromatin. *Nat. Genet.* 2016; **48**: 488–96.
92. Sarkans U, Gostev M, Athar A *et al.* The biostudies database—one stop shop for all data supporting a life sciences study. *Nucleic Acids Res.* 2018; **46**: D1266–70.
93. Edgar R, Domrachev M and Lash AE. Gene expression omnibus: Ncbi gene expression and hybridization array data repository. *Nucleic Acids Res.* 2002; **30**: 207–10.
94. de Souza N. The encode project. *Nat. Methods* 2012; **9**: 1046.
95. Uniprot: The universal protein knowledgebase in 2023. *Nucleic Acids Res.* 2023; **51**: D523–31.
96. Burley SK, Bhikadiya C, Bi C *et al.* Rcsb protein data bank (rcsb. org): Delivery of experimentally-determined pdb structures alongside one million computed structure models of proteins from artificial intelligence/machine learning. *Nucleic Acids Res.* 2023; **51**: D488–508.
97. Zdrazil B, Felix E, Hunter F *et al.* The chembl database in 2023: A drug discovery platform spanning multiple bioactivity data types and time periods. *Nucleic Acids Res.* 2024; **52**: D1180–92.
98. Tingle BI, Tang KG, Castanon M *et al.* Zinc-22– a free multi-billion-scale database of tangible compounds for ligand discovery. *J. Chem. Inf. Model.* 2023; **63**: 1166–76.
99. Kim S, Chen J, Cheng T *et al.* Pubchem 2023 update. *Nucleic Acids Res.* 2023; **51**: D1373–80.
100. Moreno P, Fexova S, George N *et al.* Expression atlas update: Gene and protein expression in multiple species. *Nucleic Acids Res.* 2022; **50**: D129–40.
101. Regev A, Teichmann SA, Lander ES *et al.* The human cell atlas. *eLife* 2017; **6**: e27041.
102. Fang Y, Zhang Q, Zhang N *et al.* Knowledge graph-enhanced molecular contrastive learning with functional prompt. *Nat. Mach. Intell.* 2023; **5**: 542–53.
103. Wang Z, Zhang Q, HU SW *et al.* Multi-level protein structure pre-training via prompt learning. *Proceedings of the International Conference on Learning Representations* (2023) .
104. He C, Luo R, Hu S *et al.* Ultraeval: A lightweight platform for flexible and comprehensive evaluation for llms. ArXiv: 2404.07584.
105. Feng S, Li S, Chen L *et al.* Unveiling potential threats: Backdoor attacks in single-cell pre-trained models. *Cell Discov.* 2024; **10**: 122.
106. Wu C, Wang X, Lian D *et al.* A causality inspired framework for model interpretation. *Proceedings of the ACM SIGKDD Conference on Knowledge Discovery and Data Mining* (2023) 2731–41.
107. Farquhar S, Kossen J, Kuhn L *et al.* Detecting hallucinations in large language models using semantic entropy. *Nature* 2024; **630**: 625–30.
108. Belyi M, Friel R, Shao S *et al.* Luna: An evaluation foundation model to catch language model hallucinations with high accuracy and low cost. ArXiv: 2406.00975.

ORIGINAL UNEDITED MANUSCRIPT

Downloaded from https://academic.oup.com/nsr/advance-article/doi/10.1093/nsr/nwaf028/7979309 by guest on 08 February 2025