Saudi Journal of Medicine and Public Health

https://saudijmph.com/index.php/pub

https://doi.org/10.64483/jmph-191

The Biology and Laboratory Paradigm Shift: A Review of Machine Learning and Artificial Intelligence in Biomolecular Data Interpretation and Predictive Modelling

Saeed Ali Alasmari [(1)] , Abdulmajeed Saad Bin Baz [(2)] , Mohammed Awadh Alshehri [(3)] , Saad Salem Aldawsari [(4)] , Awadh Jarallah Alkaabi [(5)] , Saad Mahdi Saleh Alamri [(2)] , Turki Saeed Alwadaie [(6)] , ALhanouf Mohammed Moredh [(2)] , Turkia Mohammed Alharthi [(7)] , Abdullah Ahmed Alamer [(8)] , Abdullah Salman Al Salman [(9)]

(1) Dirayiah Hospital, Ministry of Health, Saudi Arabia,
(2) Regional Laboratory,Ministry of Health, Saudi Arabia,
(3) King Saud Medical City,Ministry of Health, Saudi Arabia,
(4) Riyadh Regional Lab,Ministry of Health, Saudi Arabia,
(5) MCH Najran,Ministry of Health, Saudi Arabia,
(6) KSMC, Ministry of Health, Saudi Arabia,
(7) Ministry Of Health Branch In Riyadh, Saudi Arabia,
(8) Security Forces Hospital, Ministry of Health, Saudi Arabia,
(9) Jalajil PHC,Ministry of Health, Saudi Arabia

Abstract

Background: The life sciences are experiencing an explosion of data from high-throughput genomics, proteomics, and metabolomics. It is a challenging problem to interpret the complex data sets in parallel with developments in artificial intelligence (AI) and machine learning (ML).

Aim: This review categorizes the groundbreaking contribution of AI/ML to biomolecular data science during the period 2015-2024, elucidating its use in multi-omics analysis, protein structure prediction, and experimental automation.

Methods: We performed a systematic literature review highlighting the application of sophisticated computational models such as deep neural networks, graph neural networks, and transformer architectures in diverse biomolecular data.

Results: Our results establish that AI/ML has changed the discipline at its core. These technologies facilitate the discovery of new biomarkers and drug targets from multi-omics data and have made breakthrough achievements in protein structure prediction using AlphaFold2. In addition, AI is now automating experimental design, making closed-loop systems that accelerate discovery.

Conclusion: AI and ML are no longer ancillary tools but intrinsic drivers of a new paradigm in molecular biology. Although data quality and interpretability challenges persist, the incorporation of AI is imperative for decoding the patterns of complex biological systems and developing personalized medicine.

Keywords: Artificial Intelligence, Machine Learning, Deep Learning, Genomics, Proteomics, Metabolomics, Protein Structure Prediction, AlphaFold, Predictive Modeling, Multi-omics, Experimental Automation.

1. Introduction

The 21st century has witnessed biology shift away from a qualitative discipline per se to a quantitative data-rich discipline. Technologies such as next-generation sequencing (NGS), mass spectrometry-based proteomics, and high-resolution metabolomics produce terabytes of data per experiment daily (Hasin et al., 2017). This data deluge, although rich, overwhelmed the capacity of traditional statistical and computational means to analyze and interpret. Its high dimensionality, intrinsic noise, and complexity require more sophisticated, adaptive, and non-linear methods.

Enter Artificial Intelligence (AI) and Machine Learning (ML). AI, the broad term for the capacity of machines to perform tasks that would otherwise require human intelligence, has discovered a fertile ground in biology. ML, an AI, is fed algorithms that can learn patterns and associations from data without being specifically programmed for every case (LeCun et al., 2015). The period after 2015 has been particularly explosive, with deep learning—deep learning being an ML technique using artificial neural networks with numerous layers—sweeping the scene across domains. In biomedicine, this intersection has accelerated a paradigm shift such that scientists can move from descriptive analysis to predictive and generative modeling (Eraslan et al., 2019).

*Corresponding author e-mail: s515s515@hotmail.com (Saeed Ali Alasmari).

Receive Date: 25 November 2024, Revise Date: 25 December 2024, Accept Date: 31 December 2024

This review aims to provide an in-depth description of the application of AI and ML for the prediction of complex biomolecular data as well as for developing predictive models. We shall start by looking at the area of AI in integrating multi-omics data, such as genomics, proteomics, and metabolomics. We will then consider the landmark achievement of AI in structural biology, such as AlphaFold2, and its consequences. We will next examine the new field of AI-supported experimental design and verification. We will then consider the challenges and opportunities of this rapidly evolving discipline. Along the way, we will highlight exemplary work between 2015 and 2024 to illustrate state-of-the-art and trends.

Machine Learning and AI for Multi-Omics Data Interpretation

The "omics" revolution has provided us with a systems-level view of biology. Each of the omics layers—genomics (DNA), transcriptomics (RNA), proteomics (proteins), and metabolomics (metabolites)—gives only a partial picture. AI and ML are uniquely positioned compared to any other technology to integrate these various types of data and create a more comprehensive model of cellular function and disease dysfunction (Figure 1).

Genomics and Transcriptomics: From Variant Calling to Functional Prediction

Genomics is at the vanguard biology field to adopt large-scale data analysis. ML was first adopted in solving such problems as identifying locations where proteins interact with DNA or classifying genomic sequences. Later, deep learning models have vastly improved performance. Early variant callers employed hand-crafted statistical models. Google Health's DeepVariant model, utilizing deep learning, re-framed variant calling as an image class problem and overlayed aligned sequencing reads on an image to use a convolutional neural network (CNN) to identify insertions, deletions, and single-nucleotide polymorphisms (SNPs) far more accurately than methods before (Poplin et al., 2018). This work showed how domain shift—using biological information in the form of visual patterns—could produce breakthroughs.

Besides identification, comprehension of the functional impact of non-coding variants is a critical challenge. ExPecto and Sei are programs that use deep learning models that have been trained on a vast corpus of genomic and epigenomic data to predict the transcriptional and epigenetic effect of any sequence variant, including those in regulatory regions, to prioritize pathogenic mutations (Zhou et al., 2018; Chen et al., 2022). scRNA-seq and bulk data are high-dimensional and sparse. Autoencoders—a type of neural network used for dimensionality reduction—are used by ML to compress this data into interpretable latent representations. The representations can then be used for cell type labeling, trajectory inference (pseudotime analysis), and denoising (Lopez et al., 2018). For instance, approaches like scVI (single-cell Variational Inference) provide a probabilistic framework for normalization, visualization, and differential expression analysis of scRNA-seq data, effectively addressing technical noise and batch effects (Lopez et al., 2018). More recently, generative models like Generative Adversarial Networks (GANs) and diffusion models have been used to create synthetic, high-quality single-cell data for dataset augmentation and for in-silico perturbation studies (Marouf et al., 2020).

Proteomics: Unpacking the Proteome's Complexity

Information proteomics, which is predominantly generated by mass spectrometry, is further complicated by the dynamic nature of protein expression, post-translational modifications (PTMs), and protein-protein interactions. Peptide sequence-to-mass spectrum matching is one of the core activities in proteomics. Traditional database search engines can be ambiguous. Deep learning methods such as MS²PIP and Prosit directly predict the fragmentation spectrum of a peptide sequence, leading to more confident identifications and revealing newly unassigned spectra (Gessulat et al., 2019). The depth and accuracy of proteome coverage have improved significantly due to this.

Together with Phosphorylation, PTMs have important roles in signaling. Prediction of PTM sites from sequence alone is a classic bioinformatics problem. Deep learning architectures with protein language model embeddings reached new benchmarks for phosphorylation, acetylation, and glycosylation site prediction (Ofer et al., 2021). Besides, ML has been used to integrate proteomic data with other omics layers to identify signaling networks that are perturbed in cancer and other diseases and determine new biomarker and drug target discovery (Zhang et al., 2021).

Metabolomics: Omics Cascade End

Metabolomics provides a direct readout of cell phenotype and is very dynamic. However, identification of metabolites from mass spectra is still an enormous bottleneck. A minimal percentage of spectral features in the typical untargeted metabolomics experiment can be properly identified. ML techniques are being used to predict a candidate metabolite's mass spectrum from its structure, and vice versa. FingerID utilizes support vector machines (SVMs) and deep learning to map fragmentation spectra onto molecular structures by searching in silico fragmentation libraries (Dührkop et al., 2019). This has made it possible to annotate unknown metabolites.

By integrating metabolomic data with clinical variables, ML classifiers (Random Forests, XGBoost) have succeeded in identifying metabolite diagnostic signatures for diseases like cancer, diabetes, and neurological disorders (Anwardeen et al., 2023). Not only do the models yield diagnostic

capacity, but they may also shed light on the pathogenic metabolic pathways implicated.

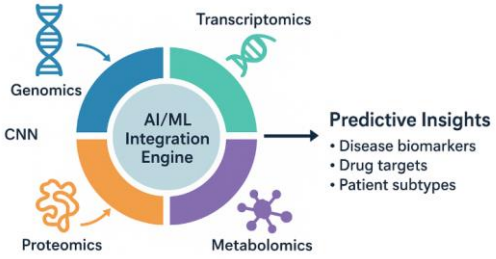

Figure 1: AI-Driven Multi-Omics Data Integration Framework

Multi-Omics Data Integration

The true power of AI lies in the integration of multiple omics layers. Heterogeneity of data is the challenge. Earlier methods used concatenation or kernel-based operations, but deep learning can offer more universal and powerful solutions. Multi-modal autoencoders and deep neural network models can learn a shared representation of different types of omics data in a shared latent space. This unified representation can then be utilized for improved disease subtyping, patient stratification, and survival prediction (Picard et al., 2021). Genomic, transcriptomic, and histopathological image data, for example, have been integrated through deep learning and employed to predict cancer prognosis superior to any single data type alone (Mobadersany et al., 2018).

Graph Neural Networks (GNNs)

Living organisms are networked by their biology. GNNs directly process graph structures, e.g., protein-protein interaction networks or gene regulatory networks, with nodes representing biomolecules and edges representing their interactions. With the incorporation of multi-omics data as node features, GNNs predicted novel interactions, identified disease modules, and ranked candidate genes (Zitnik et al., 2019). This is a move towards modeling biology in its natural network context (Table 1).

Table 1: Overview of Key AI/ML Applications to Multi-Omics Data Analysis

Omics Field	Key Task	Traditional/Baseline Method	AI/ML Approach	Example Tool/Model (Citation)	Impact/Advancement
Genomics	Variant Calling	GATK, Samtools	CNN (image classification)	DeepVariant (Poplin et al., 2018)	Higher accuracy, especially in difficult genomic regions.
Genomics	Non-coding Variant Effect	PWM, GWAS	Deep learning on regulatory code	ExPecto, Sei (Zhou et al., 2018; Chen et al., 2022)	Functional interpretation of variants in non-coding regions.
Transcriptomics	scRNA-seq Analysis	PCA, t-SNE	Variational Autoencoders	scVI, scANVI (Lopez et al., 2018)	Effective denoising, batch correction, and latent space representation.
Transcriptomics	Synthetic Data Generation	N/A	Generative Adversarial Networks	scGAN (Marouf et al., 2020)	Data augmentation, in-silico experimentation.
Proteomics	Peptide Identification	Database Search (e.g., MaxQuant)	Spectrum Prediction with DNNs	Prosit, MS²PIP (Gessulat et al., 2019)	Increased proteome coverage and identification confidence.
Proteomics	PTM Prediction	Sequence Motif Analysis	Embeddings from Protein Language Models	(Ofer et al., 2021)	State-of-the-art accuracy in predicting modification sites.
Metabolomics	Metabolite Annotation	Spectral Library Search	In-silico Fragmentation & ML	CSI: FingerID (Dührkop et al., 2019)	Dramatically increased annotation rates for unknown metabolites.
Metabolomics	Disease Biomarker Discovery	Univariate Statistics	Multivariate Classifiers (XGBoost, RF)	(Anwardeen et al., 2023)	Identification of robust, multi-metabolite diagnostic signatures.
Multi-Omics	Data Integration	MOFA, iCluster	Multi-modal Autoencoders	(Picard et al., 2021)	Learning joint representations for

					superior patient stratification.
Multi-Omics	Network Biology	Gene Set Enrichment	Graph Neural Networks (GNNs)	(Zitnik et al., 2019)	Modeling biology as interactive networks for prediction.

The AI Advances in Protein Structure Prediction

For over 50 years, the "protein folding problem" or protein three-dimensional structure prediction from the amino acid sequence has been a grand challenge in biology. AlphaFold2 was the advance by DeepMind, an AI program that achieved accuracy comparable to experimental methods (Jumper et al., 2021).

The Pre-AlphaFold2 Landscape

Before AlphaFold2, computational methods like homology modeling and fragment assembly were at best roughly correct, often not working for proteins with no close structurally characterized homologs. The CASP experiments consistently found the gap between computational prediction and experimental structure. Early ML approaches combined predicted input features like contact maps, but progress was incremental (Senior et al., 2020).

The AlphaFold2 Architecture: A Technical Leap

AlphaFold2 was not an incremental achievement but a revolutionary idea. The most basic innovation in AlphaFold2 is its end-to-end deep learning approach, which avoids intermediate steps like prediction of the contact map.

1. **Evolutionary Sequence Analysis:**

The target sequence input is not only the sequence but a multiple sequence alignment (MSA) of homologs, and this MSA has evolutionary constraints. There is an independent deep learning module, the Evoformer, that receives the MSA and a corresponding representation of residues and predicts evolutionary and co-evolutionary relationships (Jumper et al., 2021).

2. **The Structure Module:**

It is the most innovative component. It accepts the representations from the Evoformer and outputs directly the 3D coordinates of all the atoms. It uses an attention-based mechanism (a transformer model) to reason about spatial relationships between residues, effectively "folding" the protein in silico in a single, end-to-end pass (Jumper et al., 2021).

3. **Iterative Refinement:**

The system iterates, using its own output to refine the predicted structure, optimizing local geometry and steric clash. The result was a system capable of predicting protein structures at sub-atomic accuracy for the majority of CASP14 targets, solving the fundamental single-chain protein folding problem.

Ramifications and Subsequent Developments

The publication of AlphaFold2 and the subsequent AlphaFold Protein Structure Database, with predicted structures for virtually all but a minuscule number of the proteins in the cataloged human proteome, and over 20 other model organisms, was an earthquake (Varadi et al., 2022). Structure-based drug design relies on structural information about the target protein. AlphaFold2 has generated high-quality models for the vast majority of proteins without an experimental structure, providing novel opportunities for virtual screening and lead optimization (Thornton et al., 2021). It has been used, for instance, to model recalcitrant targets like G-protein-coupled receptors (GPCRs) and membrane proteins. Design is the reverse of folding. With AlphaFold2 and after, there have been models like RoseTTAFold and RFdiffusion, using similar architectural ideas to design new proteins that do not occur naturally (Baek et al., 2021; Watson et al., 2023). It has massive potential for designing novel enzymes, drugs, and biomaterials.

The field is racing to more difficult problems. AlphaFold-Multimer and subsequent versions are specifically aimed at predicting protein complex structures (Evans et al., 2021). Although there remain difficulties, in particular for very flexible complexes, the rate of progress is rapid. Parallel and complementary to this has been the development of Protein Language Models (pLMs), such as ESM (Evolutionary Scale Modeling) and ProtTrans. These are transformer models of a gigantic size, which are pre-trained on millions of protein sequences from databases. They acquire fundamental laws of protein syntax and semantics and produce high-strength numerical embeddings for each sequence (Rives et al., 2021). These embeddings are now a default option for a wide range of downstream tasks, from protein function and stability prediction to the effect of missense mutations, and will usually outperform MSAA-derived features, especially for orphan sequences with few homologs (Brandes et al., 2022).

AI in Automated Experimental Design and Validation

The final frontier of AI in biomolecular science is to complete the loop from hypothesis, prediction, experiment, and analysis. AI is beginning to shift from the role of a passive analytical tool to an active participant in the scientific process.

Experimental Parameters Optimization

Biological assays typically come with a vast parameter space (e.g., levels of reagents, temperatures, time points). Active learning and AI-powered Bayesian optimization can explore this space comprehensively to find good conditions using many fewer experiments than traditional grid searches (Malkomes& Garnett, 2018). This is being applied to

optimizing CRISPR guide RNA design, PCR, and protein crystallization.

Self-Driving Laboratories

The "self-driving lab" idea combines robotic automation with AI planning. The AI proposes an experiment based on a predefined objective (e.g., to synthesize a molecule with specific properties), a robotic platform experiment, and the results are returned to the AI to update its model and propose the next experiment. This has been demonstrated in fields such as materials science and is being applied in biology to automate strain engineering in synthetic biology and the discovery of new genetic circuits (Seifrid et al., 2022).

AI for Data Validation and Reproducibility

The reproducibility crisis of science is also, in part, a data quality problem. AI models can be instructed to find outliers, detect technical artifacts, and even flag potentially falsified images on scientific articles (Gendron et al., 2022). ML algorithms, for example, can filter Western blot images or flow cytometry data for signs of improper manipulations or of poor quality as an initial line of defense in data analysis and peer review.

Hypothesis Formation with Generative AI

Large language models (LLMs) like GPT-4, if trained on the vast corpus of scientific literature (e.g., PubMed), can also act as superhuman assistants in literature generation. They can abstract existing knowledge, identify unmapped connections between unrelated fields, and generate novel, testable hypotheses (Wang et al., 2023). While not replacing scientists, they can significantly accelerate the initial phase of research development (Table 2).

Table 2: Advances through AI in Protein Science and Experimental Automation

Domain	Specific Challenge	Pre-AI Paradigm	AI/ML Solution	Key Model/System (Citation)	Impact/Advancement
Protein Structure	Single-chain 3D Prediction	Homology Modeling, Physics-based	End-to-End Deep Learning (Transformers)	AlphaFold2 (Jumper et al., 2021)	Solved the core folding problem; atomic-level accuracy.
Protein Structure	Rapid, Accessible Prediction	N/A	Simplified, Open-Source AF2 Architecture	RoseTTAFold (Baek et al., 2021)	Democratized high-accuracy structure prediction.
Protein Science	Functional & Stability Prediction	Evolutionary Analysis (MSA-dependent)	Protein Language Models (pLMs)	ESM-2, ProtTrans (Rives et al., 2021)	Powerful sequence-only embeddings for diverse prediction tasks.
Protein Design	De Novo Protein Creation	Rational Design, Phage Display	Inverse Folding & Generative Models	RFdiffusion, ProteinMPNN (Watson et al., 2023; Dauparas et al., 2022)	Creation of novel functional proteins and enzymes from scratch.
Protein Complexes	Protein-Protein Interaction Structures	Docking Simulations	Specialized Multimer Prediction	AlphaFold-Multimer (Evans et al., 2021)	Improved accuracy for quaternary structure prediction.
Experiment Design	Parameter Optimization	One-factor-at-a-time, Grid Search	Bayesian Optimization	(Malkomes& Garnett, 2018)	Finds optimal experimental conditions with minimal trials.
Experiment Design	Synthetic Biology & Chemistry	Manual Design-Build-Test Cycles	Self-Driving Laboratories	(Seifrid et al., 2022)	Fully automated, closed-loop discovery systems.
Validation	Image Fraud Detection	Manual Peer Review	Image Analysis with CNNs	(Gendron et al., 2022)	Automated screening for image duplication and manipulation.
Hypothesis Generation	Literature Mining & Connection	Manual Literature Review	Fine-tuned Large Language Models	GPT-4, Galactica (Wang et al., 2023)	Accelerated knowledge synthesis and novel hypothesis generation.

Challenges, Limitations, and Future Directions

Although the advances are really inspiring, the integration of AI in biomolecular science is marred by several major hurdles.

1. **Data Quality and Quantity:**

The quality and quantity of training data are a fundamental part of the performance of AI models. Noisy, biased, or badly annotated data will produce biased and unreliable models. The "garbage in, garbage out" maxim is the most significant principle. Further, for the majority of rare diseases or biological settings, large datasets do not exist, and therefore, few-shot or zero-shot learning strategies must be created (Feuerriegel et al., 2024).

2. **Model Explainability and Interpretability (XAI):**

Deep models are typically referred to as "black boxes." Understanding why a model is making a particular prediction is critical to knowing what biological knowledge is being derived and building trust, especially in clinical settings. Techniques like SHAP (Shapley Additive exPlanations) and LIME (Local Interpretable Model-agnostic Explanations) are being adapted to work on biological models in an attempt to highlight which features (like specific genomic positions or metabolites) contributed most to a prediction (Lundberg & Lee, 2017). The design of inherently interpretable models represents an active area for future research.

3. **Generalization and Robustness:**

Models trained with data from a single cell type, species, or technology fail to generalize to others (domain shift). Achieving robustness and generalizability across biological contexts is a primary challenge that can be tackled by curated training data and algorithmic advancements (Yang et al., 2022).

4. **Computational Resources:**

Training such high-quality models as AlphaFold2 or large pLMs requires significant computational resources and energy, constituting an entry barrier for small laboratories and also causing concerns regarding the carbon footprint of AI research (Strubell et al., 2019). Efficient models and algorithms are the solution.

5. **Ethical and Societal Implications:**

The ability to predict disease risk from genomic data, design new pathogens, or generate synthetic biological data raises profound ethical issues. Data privacy, consent, algorithmic bias (e.g., models being less accurate on underrepresented groups), and the potential for dual-use call for anticipatory governance and input from bioethicists, policymakers, and the public (Raikar et al., 2023).

Future Directions

In the coming times, the trajectory of AI in biomolecular science is towards some groundbreaking frontiers. One of the primary directions is the development of foundation models for biology-large-scale, multi-modal pre-trained models over a range of data types from DNA and protein sequences to cellular images and scientific texts. These models, adaptable to a huge range of downstream tasks with minimal fine-tuning, promise to be the universal platform for biological discovery (Moor et al., 2023). Concurrently, the advent of spatial omics technologies demands sophisticated AI for spatial omics to disentangle the complex spatial patterns of gene and protein expression in tissues, thereby unveiling a profound tissue architecture comprehension in health and disease (Moses & Pachter, 2022). On a higher integration level, the ambitious vision of digital twins would create comprehensive, dynamic AI models of biological systems, from individual cells to entire patients. By combining multi-omics, clinical, and lifestyle data, digital twins may be able to simulate disease development and personalize reactions to treatment, revolutionizing predictive medicine (Bruynseels et al., 2018). Lastly, in order to go beyond correlation and into true mechanistic understanding, the field must embrace causal AI. Building models that learn causal relationships from high-dimensional observational data is the next crucial step, enabling predictions of the outcome of intervention and solidifying AI's role not only in the discovery of patterns, but in informing actionable biological knowledge (Schölkopf et al., 2021).

Conclusion

This decade has been revolutionary in the life sciences, driven by the immense synergy between biomolecular data and AI/ML. We have progressed from using ML to assistive tasks to implementing deep learning machines to solve problems of existential importance, most aptly exemplified by the solution to the problem of predicting protein structure. AI is no longer just an analytical tool; it is becoming a discovery engine, capable of interpreting the richness of multi-omics spaces, anticipating accurate structural models at scale, and even designing and executing experiments independently. While issues around data, interpretability, and ethics remain, the trend is set. AI and ML are now fundamental tools in the biomolecular scientist's toolkit, ushering in the era of predictive, personalized, and programmable biology. The future will be defined by our ability to apply these technologies judiciously to crack the remaining frontiers of life and translate these predictions into actionable therapies and understanding of health and disease.

References

1. Anwardeen, N. R., Diboun, I., Mokrab, Y., Althani, A. A., & Elrayess, M. A. (2023). Statistical methods and resources for biomarker discovery using metabolomics. *BMC bioinformatics*, *24*(1), 250. https://doi.org/10.1186/s12859-023-05383-0
2. Baek, M., DiMaio, F., Anishchenko, I., Dauparas, J., Ovchinnikov, S., Lee, G. R., ... & Baker, D. (2021). Accurate prediction of protein structures

and interactions using a three-track neural network. *Science*, *373*(6557), 871-876. https://doi.org/10.1126/science.abj8754
3. Brandes, N., Ofer, D., Peleg, Y., Rappoport, N., & Linial, M. (2022). ProteinBERT: a universal deep-learning model of protein sequence and function. *Bioinformatics*, *38*(8), 2102-2110. https://doi.org/10.1093/bioinformatics/btac020
4. Bruynseels, K., Santoni de Sio, F., & Van den Hoven, J. (2018). Digital twins in health care: ethical implications of an emerging engineering paradigm. *Frontiers in genetics*, *9*, 31. https://doi.org/10.3389/fgene.2018.00031
5. Chen, K. M., Wong, A. K., Troyanskaya, O. G., & Zhou, J. (2022). A sequence-based global map of regulatory activity for deciphering human genetics. *Nature genetics*, *54*(7), 940-949. https://doi.org/10.1038/s41588-022-01102-2
6. Dauparas, J., Anishchenko, I., Bennett, N., Bai, H., Ragotte, R. J., Milles, L. F., ... & Baker, D. (2022). Robust deep learning–based protein sequence design using ProteinMPNN. *Science*, *378*(6615), 49-56. https://doi.org/10.1126/science.add2187
7. Dührkop, K., Fleischauer, M., Ludwig, M., Aksenov, A. A., Melnik, A. V., Meusel, M., ... & Böcker, S. (2019). SIRIUS 4: a rapid tool for turning tandem mass spectra into metabolite structure information. *Nature methods*, *16*(4), 299-302. https://doi.org/10.1038/s41592-019-0344-8
8. Eraslan, G., Avsec, Ž., Gagneur, J., & Theis, F. J. (2019). Deep learning: new computational modelling techniques for genomics. *Nature reviews genetics*, *20*(7), 389-403. https://doi.org/10.1038/s41576-019-0122-6
9. Evans, R., O'Neill, M., Pritzel, A., Antropova, N., Senior, A., Green, T., ... & Hassabis, D. (2021). Protein complex prediction with AlphaFold-Multimer. *biorxiv*, 2021-10. https://doi.org/10.1101/2021.10.04.463034
10. Feuerriegel, S., Hartmann, J., Janiesch, C., & Zschech, P. (2024). Generative ai. *Business & Information Systems Engineering*, *66*(1), 111-126. https://doi.org/10.1007/s12599-023-00834-7
11. Gendron, Y., Andrew, J., & Cooper, C. (2022). The perils of artificial intelligence in academic publishing. *Critical Perspectives on Accounting*, *87*, 102411. https://doi.org/10.1016/j.cpa.2021.102411
12. Gessulat, S., Schmidt, T., Zolg, D. P., Samaras, P., Schnatbaum, K., Zerweck, J., ... & Wilhelm, M. (2019). Prosit: proteome-wide prediction of peptide tandem mass spectra by deep learning. *Nature methods*, *16*(6), 509-518. https://doi.org/10.1038/s41592-019-0426-7
13. Hasin, Y., Seldin, M., & Lusis, A. (2017). Multi-omics approaches to disease. *Genome biology*, *18*(1), 83. https://doi.org/10.1186/s13059-017-1215-1
14. Jumper, J., Evans, R., Pritzel, A., Green, T., Figurnov, M., Ronneberger, O., ... & Hassabis, D. (2021). Highly accurate protein structure prediction with AlphaFold. *nature*, *596*(7873), 583-589. https://doi.org/10.1038/s41586-021-03819-2
15. LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. *nature*, *521*(7553), 436-444. https://doi.org/10.1038/nature14539
16. Lopez, R., Regier, J., Cole, M. B., Jordan, M. I., & Yosef, N. (2018). Deep generative modeling for single-cell transcriptomics. *Nature methods*, *15*(12), 1053-1058. https://doi.org/10.1038/s41592-018-0229-2
17. Lundberg, S. M., & Lee, S. I. (2017). A unified approach to interpreting model predictions. *Advances in neural information processing systems*, *30*.
18. Malkomes, G., & Garnett, R. (2018). Automating Bayesian optimization with Bayesian optimization. *Advances in Neural Information Processing Systems*, *31*.
19. Marouf, M., Machart, P., Bansal, V., Kilian, C., Magruder, D. S., Krebs, C. F., & Bonn, S. (2020). Realistic in silico generation and augmentation of single-cell RNA-seq data using generative adversarial networks. *Nature communications*, *11*(1), 166. https://doi.org/10.1038/s41467-019-14018-z
20. Mobadersany, P., Yousefi, S., Amgad, M., Gutman, D. A., Barnholtz-Sloan, J. S., Velázquez Vega, J. E., ... & Cooper, L. A. (2018). Predicting cancer outcomes from histology and genomics using convolutional networks. *Proceedings of the National Academy of Sciences*, *115*(13), E2970-E2979. https://doi.org/10.1073/pnas.1717139115
21. Moor, M., Banerjee, O., Abad, Z. S. H., Krumholz, H. M., Leskovec, J., Topol, E. J., & Rajpurkar, P. (2023). Foundation models for generalist medical artificial intelligence. *Nature*, *616*(7956), 259-265. https://doi.org/10.1038/s41586-023-05881-4
22. Moses, L., & Pachter, L. (2022). Museum of spatial transcriptomics. *Nature methods*, *19*(5), 534-546. https://doi.org/10.1038/s41592-022-01409-2
23. Ofer, D., Brandes, N., & Linial, M. (2021). The language of proteins: NLP, machine learning & protein sequences. *Computational and Structural Biotechnology Journal*, *19*, 1750-1758. https://doi.org/10.1016/j.csbj.2021.03.022
24. Picard, M., Scott-Boyer, M. P., Bodein, A., Périn, O., & Droit, A. (2021). Integration strategies of multi-omics data for machine learning analysis. *Computational and Structural Biotechnology Journal*, *19*, 3735-3746. https://doi.org/10.1016/j.csbj.2021.06.030
25. Poplin, R., Chang, P. C., Alexander, D., Schwartz, S., Colthurst, T., Ku, A., ... & DePristo, M. A. (2018). A universal SNP and small-indel variant

caller using deep neural networks. *Nature biotechnology*, *36*(10), 983-987. https://doi.org/10.1038/nbt.4235

26. Raikar, G. V. S., Raikar, A. S., & Somnache, S. N. (2023). Advancements in artificial intelligence and machine learning in revolutionising biomarker discovery. *Brazilian Journal of Pharmaceutical Sciences*, *59*, e23146. https://doi.org/10.1590/s2175-97902023e23146
27. Rives, A., Meier, J., Sercu, T., Goyal, S., Lin, Z., Liu, J., ... & Fergus, R. (2021). Biological structure and function emerge from scaling unsupervised learning to 250 million protein sequences. *Proceedings of the National Academy of Sciences*, *118*(15), e2016239118. https://doi.org/10.1073/pnas.2016239118
28. Seifrid, M., Pollice, R., Aguilar-Granda, A., Morgan Chan, Z., Hotta, K., Ser, C. T., ... & Aspuru-Guzik, A. (2022). Autonomous chemical experiments: Challenges and perspectives on establishing a self-driving lab. *Accounts of Chemical Research*, *55*(17), 2454-2466. https://doi.org/10.1021/acs.accounts.2c00220
29. Schölkopf, B., Locatello, F., Bauer, S., Ke, N. R., Kalchbrenner, N., Goyal, A., & Bengio, Y. (2021). Toward causal representation learning. *Proceedings of the IEEE*, *109*(5), 612-634. https://doi.org/10.1109/JPROC.2021.3058954
30. Senior, A. W., Evans, R., Jumper, J., Kirkpatrick, J., Sifre, L., Green, T., ... & Hassabis, D. (2020). Improved protein structure prediction using potentials from deep learning. *Nature*, *577*(7792), 706-710. https://doi.org/10.1038/s41586-019-1923-7
31. Strubell, E., Ganesh, A., & McCallum, A. (2020, April). Energy and policy considerations for modern deep learning research. In *Proceedings of the AAAI conference on artificial intelligence* (Vol. 34, No. 09, pp. 13693-13696). https://doi.org/10.1609/aaai.v34i09.7123
32. Thornton, J. M., Laskowski, R. A., & Borkakoti, N. (2021). AlphaFold heralds a data-driven revolution in biology and medicine. *Nature Medicine*, *27*(10), 1666-1669. https://doi.org/10.1038/s41591-021-01533-0
33. Varadi, M., Anyango, S., Deshpande, M., Nair, S., Natassia, C., Yordanova, G., ... & Velankar, S. (2022). AlphaFold Protein Structure Database: massively expanding the structural coverage of protein-sequence space with high-accuracy models. *Nucleic acids research*, *50*(D1), D439-D444. https://doi.org/10.1093/nar/gkab1061
34. Wang, H., Fu, T., Du, Y., Gao, W., Huang, K., Liu, Z., ... & Zitnik, M. (2023). Scientific discovery in the age of artificial intelligence. *Nature*, *620*(7972), 47-60. https://doi.org/10.1038/s41586-023-06221-2
35. Watson, J. L., Juergens, D., Bennett, N. R., Trippe, B. L., Yim, J., Eisenach, H. E., ... & Baker, D. (2023). De novo design of protein structure and function with RFdiffusion. *Nature*, *620*(7976), 1089-1100. https://doi.org/10.1038/s41586-023-06415-8
36. Yang, K. D., Belyaeva, A., Venkatachalapathy, S., Damodaran, K., Katcoff, A., Radhakrishnan, A., ... & Uhler, C. (2021). Multi-domain translation between single-cell imaging and sequencing data using autoencoders. *Nature communications*, *12*(1), 31. https://doi.org/10.1038/s41467-020-20249-2
37. Zhang, B., Whiteaker, J. R., Hoofnagle, A. N., Baird, G. S., Rodland, K. D., & Paulovich, A. G. (2019). Clinical potential of mass spectrometry-based proteogenomics. *Nature Reviews Clinical Oncology*, *16*(4), 256-268. https://doi.org/10.1038/s41571-018-0135-7
38. Zhou, J., Theesfeld, C. L., Yao, K., Chen, K. M., Wong, A. K., & Troyanskaya, O. G. (2018). Deep learning sequence-based ab initio prediction of variant effects on expression and disease risk. *Nature genetics*, *50*(8), 1171-1179. https://doi.org/10.1038/s41588-018-0160-6
39. Zitnik, M., Nguyen, F., Wang, B., Leskovec, J., Goldenberg, A., & Hoffman, M. M. (2019). Machine learning for integrating data in biology and medicine: Principles, practice, and opportunities. *Information Fusion*, *50*, 71-91. https://doi.org/10.1016/j.inffus.2018.09.012
40. Zrimec, J., Börlin, C. S., Buric, F., Muhammad, A. S., Chen, R., Siewers, V., ... & Zelezniak, A. (2020). Deep learning suggests that gene expression is encoded in all parts of a co-evolving interacting gene regulatory structure. *Nature communications*, *11*(1), 6141. https://doi.org/10.1038/s41467-020-19921-4