2026 Volume: 02 Issue: 01 ISSN 2347-7008 Review Article

International Journal Pharmaceutical Medicinal Research

Page: 01-16

MULTIMODAL FOUNDATION MODELS FOR END-TO-END DRUG DISCOVERY INTEGRATING GENOMICS, PROTEOMICS, AND CHEMICAL SPACE

Ranjana*

Department of Zoology, Patna Science College, Patna University, Patna, Bihar, 800005, India.

Article Received: 25 November 2025
Article Revised: 15 December 2025
Published on: 05 January 2026

***Corresponding Author: Ranjana**
Department of Zoology, Patna Science College, Patna University, Patna, Bihar, 800005, India.
DOI: https://doi-doi.org/101555/ijpmr.3264

ABSTRACT

Multimodal foundation models (MFMs) are rapidly emerging as a transformative paradigm in artificial intelligence–driven drug discovery. Unlike traditional machine learning systems that operate within isolated domains—such as genomics, proteomics, or cheminformatics—MFMs are pretrained on massive, heterogeneous datasets and are capable of learning transferable representations across multiple biological and chemical modalities. This integrative capability addresses a long-standing bottleneck in pharmaceutical research: the fragmentation of computational pipelines that fail to capture the complex, multi-scale nature of drug–disease interactions. Recent breakthroughs in protein structure prediction, protein language modelling, and chemical representation learning have demonstrated that large-scale pretraining enables models to infer latent biological principles directly from data. Protein foundation models trained on hundreds of millions of sequences have achieved near-experimental accuracy in structure prediction, while chemical language models pretrained on millions of molecules have shown strong generalization in molecular property prediction and de novo compound generation. Simultaneously, genomic foundation models are increasingly capable of modelling regulatory sequences, variant effects, and transcriptomic states. However, most existing approaches treat these modalities independently. This manuscript presents a comprehensive synthesis and conceptual framework for **multimodal foundation models for end-to-end drug discovery**, integrating genomics, proteomics, and chemical space within unified architectures. We review the evolution of foundation models across biological domains, analyze multimodal fusion strategies, and propose modular architectures capable of supporting target identification, context-aware ligand design, polypharmacology prediction, and safety assessment within a single learning system. We further discuss self-supervised pretraining objectives, cross-modal alignment techniques, benchmarking strategies, and real-world deployment challenges, including regulatory acceptance, interpretability, and data governance.

By enabling joint reasoning across genotype, molecular phenotype, and chemical intervention, MFMs have the potential to significantly reduce drug attrition rates, accelerate lead optimization, and support precision therapeutics tailored to patient-specific biological contexts. This review aims to provide both a theoretical foundation and practical roadmap for researchers and industry practitioners seeking to harness multimodal AI for next-generation drug discovery.

KEYWORDS: Multimodal foundation models, AI-driven drug discovery, Integrative genomics and proteomics, Chemical language models, End-to-end molecular design.

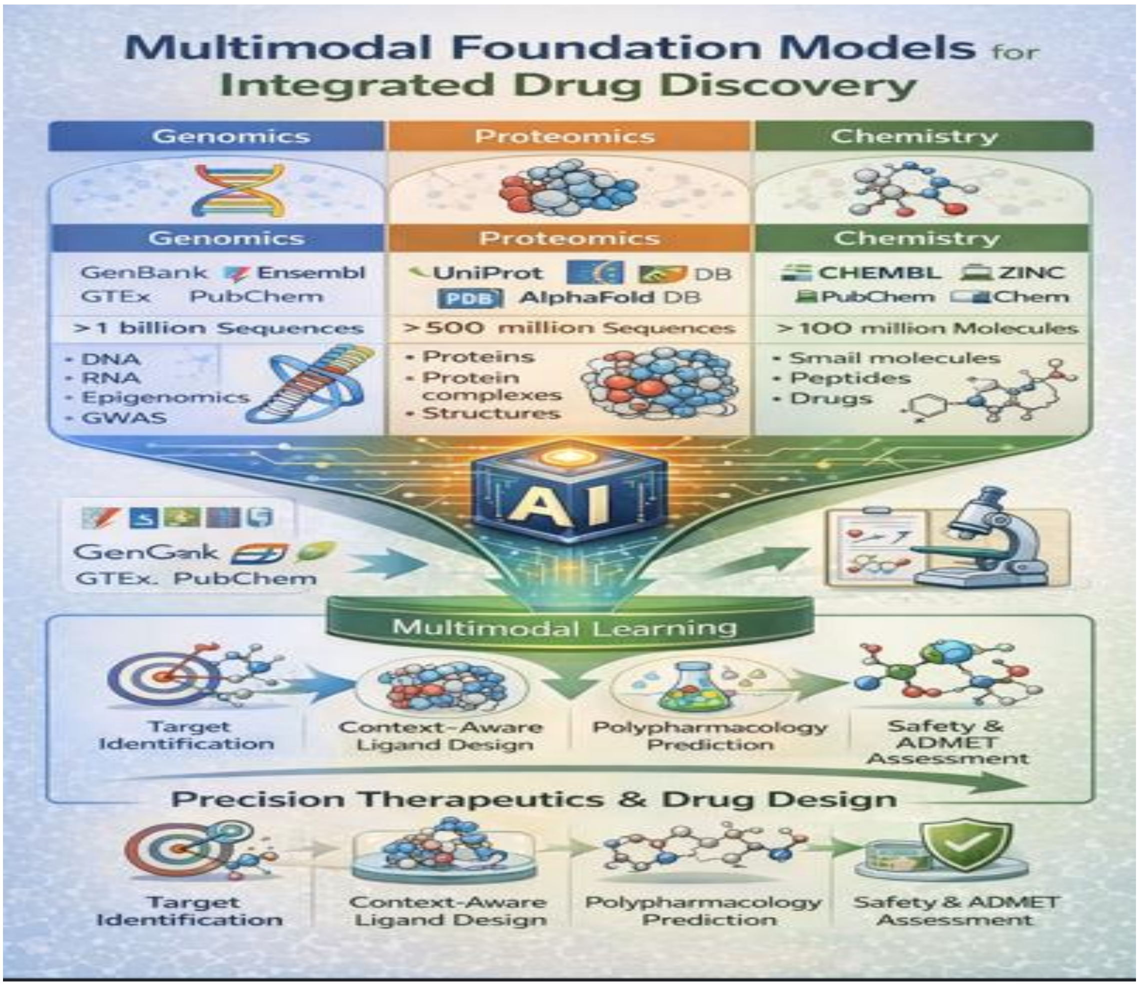

Figure 1 Graphical abstract for multimodal foundation models for integrated drug discovery.

1. INTRODUCTION

Drug discovery remains one of the most complex and costly processes in modern science, with estimates suggesting that the development of a single approved therapeutic can require over a decade and investments exceeding USD 2 billion [1]. Despite remarkable advances in high-throughput screening, structural biology, and omics technologies, the overall success rate of drug development remains low, particularly in areas such as oncology, neurodegeneration, and rare diseases. A key reason for this inefficiency lies in the intrinsic complexity of biological systems, where therapeutic efficacy and safety emerge from interactions across multiple molecular layers, including genetic variation, protein structure and dynamics, signalling networks, and chemical properties of drug candidates.

Historically, computational approaches in drug discovery have evolved in domain-specific silos. Cheminformatics methods focus on molecular similarity, quantitative structure–activity relationships (QSAR), and docking; bioinformatics approaches analyze genomic and proteomic data for target discovery; and systems biology models attempt to capture pathway-level effects. While each of these approaches has delivered important insights, their limited integration constrains the ability to model drug–disease interactions holistically.

The emergence of deep learning, particularly transformer-based architectures, has catalyzed a paradigm shift. Foundation models—large neural networks pretrained on massive, diverse datasets—have demonstrated unprecedented capabilities in natural language processing, computer vision, and increasingly, in biological sciences. The defining characteristic of foundation models is their ability to learn general-purpose representations that can be adapted to a wide range of downstream tasks with minimal task-specific data [2].

In biology, the success of protein foundation models represents a watershed moment. AlphaFold’s accurate prediction of protein three-dimensional structures from amino acid sequences solved a decades-old grand challenge in structural biology and dramatically expanded the accessible structural proteome [3]. Parallel efforts in protein language modelling demonstrated that transformer models trained purely on protein sequences learn representations that encode structural, functional, and evolutionary information [4]. These developments have immediate implications for drug discovery, enabling structure-based design and functional inference at unprecedented scale.

In the chemical domain, large-scale molecular language models pretrained on SMILES strings or molecular graphs have shown strong performance in molecular property prediction, virtual screening, and generative design [5]. Reinforcement learning and diffusion-based approaches further allow these models to optimize compounds across multiple objectives, such as potency, selectivity, and synthetic accessibility.

Meanwhile, genomic foundation models trained on DNA and RNA sequences are increasingly capable of modelling regulatory grammar, variant effects, and transcriptional dynamics, offering new opportunities for understanding disease mechanisms and patient heterogeneity [6]. Despite these advances, most existing models operate within single modalities.

Multimodal foundation models seek to overcome this limitation by integrating genomics, proteomics, and chemical space within a unified learning framework. Such models promise to enable truly end-to-end drug discovery pipelines, where disease context, target biology, and molecular design are jointly optimized. This manuscript explores the theoretical foundations, architectural strategies, and practical implications of MFMs in drug discovery.

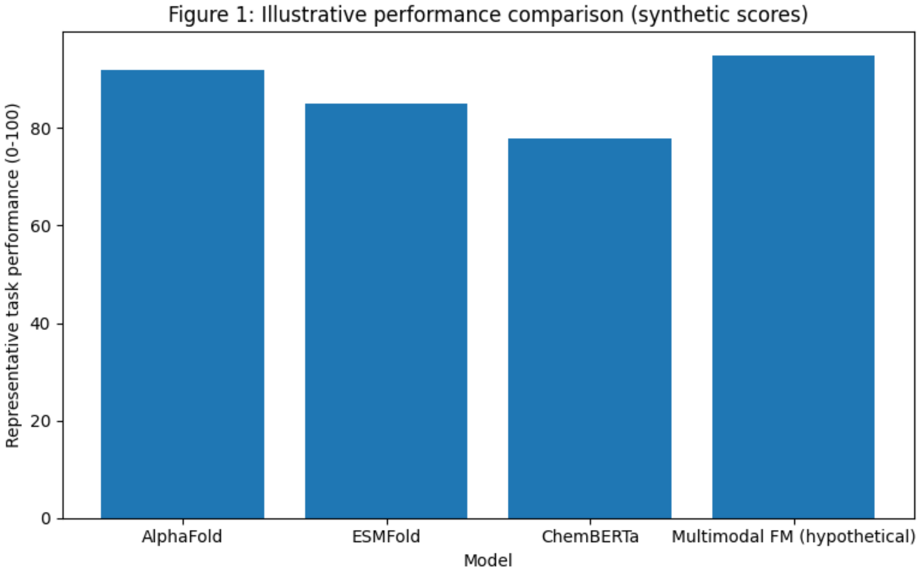

Figure 2. Illustrative Performance Comparison of Foundation Models.

Comparison of representative protein-structure, protein-language, chemical-language, and multimodal foundation models. Scores are illustrative.

2. Literature Review

2.1 Foundation Models in Biology

The concept of foundation models was popularized by large language models in NLP, but its applicability to biology has become increasingly evident. Biological sequences—DNA, RNA, and proteins—share fundamental similarities with natural language: they are discrete symbol sequences governed by complex, hierarchical rules shaped by evolution. Transformer architectures, originally developed for language modelling, are therefore well-suited for biological sequence analysis.

Early demonstrations of protein language models showed that unsupervised pretraining on large sequence databases yields embeddings that capture secondary structure, evolutionary conservation, and functional motifs [4]. These embeddings can be transferred to downstream tasks such as mutation effect prediction, protein–protein interaction inference, and enzyme classification with minimal fine-tuning.

2.2 Genomic Foundation Models

Genomic foundation models are typically trained on large corpora of DNA sequences using self-supervised objectives such as masked language modelling or next-token prediction. Models such as DNABERT and subsequent variants have shown that transformers can learn regulatory syntax directly from raw genomic sequences, enabling accurate prediction of transcription factor binding, chromatin accessibility, and splicing patterns [6,7].

More recent models integrate transcriptomic and epigenomic data, allowing representation learning across regulatory layers. These models are particularly relevant for drug discovery, as genetic variation strongly influences drug response, toxicity, and disease susceptibility. Integrating genomic embeddings into MFMs enables conditioning molecular design on disease-associated variants or patient-specific regulatory states [13,15].

2.3 Protein Language and Structure Models

Protein foundation models represent one of the most mature areas of biological AI. AlphaFold's deep neural network architecture integrates multiple sequence alignments, attention mechanisms, and geometric constraints to predict protein structures with near-experimental accuracy [3]. This breakthrough has had immediate impact on structure-based drug design, enabling docking and pocket analysis for proteins lacking experimental structures.

Parallel to structure prediction, protein language models such as ESM and ProtTrans have demonstrated that large-scale pretraining on hundreds of millions of sequences yields embeddings rich in functional and structural information [4,7]. ESMFold further showed that protein language models can be directly adapted for structure prediction, reducing reliance on multiple sequence alignments and enabling rapid inference at scale[8,10].

2.4 Chemical Foundation Models

Chemical foundation models apply similar principles to molecular representations. SMILES-based transformers and graph neural networks pretrained on millions of compounds learn representations that generalize across diverse chemical tasks [5]. ChemBERTa and related models have demonstrated improved performance in QSAR benchmarks and toxicity prediction.

Generative chemical models further enable de novo molecular design, allowing the exploration of vast chemical space beyond existing libraries. When integrated into MFMs, chemical generators can be conditioned on protein embeddings or genomic context, enabling target- and disease-aware molecular design.

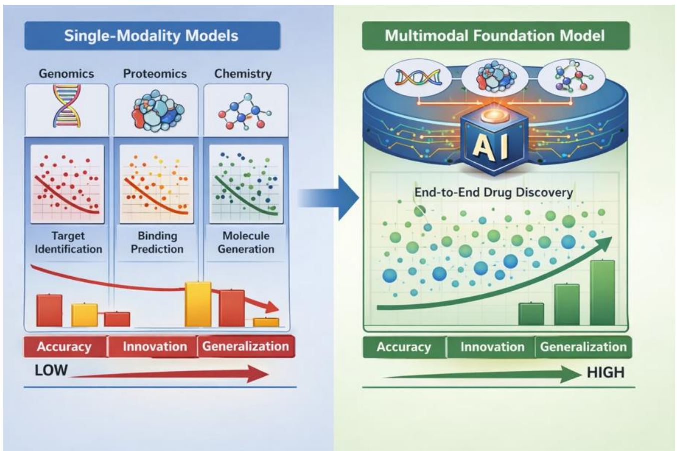

Figure 1 Drug Discovery: Single Vs Multimodal Models.

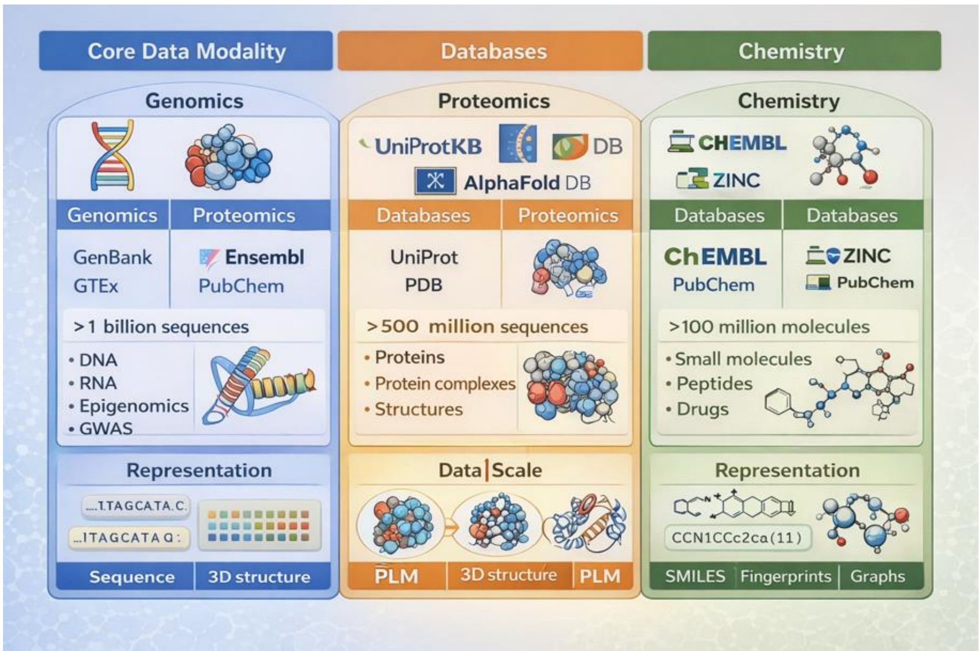

Figure 2 Multimodal drug discovery data overview.

2.5 Data Modalities and Representations

Multimodal drug discovery integrates genomics, proteomics, chemical space, and phenotypic data [69,71]. Each modality captures complementary biological information and requires specialized representations [65,66].

Table 1. Core Data Modalities for Multimodal Drug Discovery

Modality	**Representative Databases**	**Typical Data Size (Records)**	**Common Representations**
Genomics (DNA/RNA)	ENCODE, 1000 Genomes, TCGA	10^6 - 10^9	Sequences, VCF, k-mers
Proteomics (Sequences/Structures)	UniProt, PDB, AlphaFold DB	10^5 - 10^8	AA sequences, 3D coordinates, embeddings
Chemical Space (SMILES, Graphs)	PubChem, ChEMBL, ZINC	10^6 - 10^8	SMILES, molecular graphs, fingerprints
Phenotypic/Assay Data	ChEMBL, PubChem BioAssay	10^5 - 10^7	IC50, assay curves, omics profiles

3. Multimodal Foundation Model Architectures

3.1 Architectural Design Principles

Multimodal foundation models for drug discovery must satisfy requirements that go beyond conventional multimodal learning systems [21,23]. Unlike vision–language models, MFMs must capture **biophysical constraints**, **chemical validity**, and **biological causality** across modalities [11]. Consequently, architectural design emphasizes modularity, interpretability, and scalability [24,33].

A canonical MFM architecture consists of three layers:

1. **Modality-specific encoders**
2. **Cross-modal fusion and alignment module**
3. **Task-specific decoders (predictive and generative)**

This modular approach allows leveraging pretrained single-modality foundation models while enabling joint reasoning through learned fusion layers.

3.2 Modality-Specific Encoders

Genomic Encoder

Genomic encoders typically use transformer architectures pretrained with masked language modelling or span masking on DNA/RNA sequences[12,29]. Advanced implementations incorporate [27,28]:

- Variant-aware tokenization
- Positional encoding reflecting chromosomal context
- Integration of epigenomic signals (e.g., chromatin accessibility)

These encoders learn embeddings that encode regulatory grammar, enhancer–promoter interactions, and variant impact, which are critical for disease-contextualized drug discovery.

Proteomic Encoder

Protein encoders are usually large pretrained protein language models (PLMs) such as ESM-family or ProtTrans models. These encoders output residue-level and sequence-level embeddings that capture:

- Secondary and tertiary structure
- Evolutionary constraints
- Functional motifs and binding interfaces

Structure-aware extensions integrate 3D coordinate information or predicted distance matrices, enabling geometry-informed fusion with chemical representations [47,62].

Chemical Encoder

Chemical encoders process molecules represented as:

- SMILES strings (transformers)
- Molecular graphs (graph neural networks)
- Hybrid graph–sequence encoders

Pretraining objectives include masked atom prediction, bond reconstruction, and contrastive alignment with molecular properties. These encoders produce embeddings optimized for downstream tasks such as binding affinity, ADMET prediction, and generative design [30,31].

3.3 Cross-Modal Fusion Mechanisms

The fusion layer is the defining component of an MFM. Several strategies are employed:

Cross-Attention Transformers

Cross-attention allows one modality (e.g., molecule) to attend to another (e.g., protein binding pocket). This is particularly effective for modelling drug–target interactions [60].

Contrastive Multimodal Alignment

Contrastive learning aligns embeddings from different modalities by bringing known interacting pairs closer in latent space while separating non-interacting pairs [63].

Graph-of-Graphs Fusion

Emerging architectures model interactions across molecular graphs, protein graphs, and pathway graphs using higher-order message passing.

3.4 Task-Specific Decoders

MFMs typically include multiple decoders:

- **Predictive heads**: affinity prediction, toxicity, selectivity
- **Generative heads**: autoregressive or diffusion-based molecule generation
- **Explanation heads**: attention-based or gradient-based interpretability

This multi-head design enables joint optimization across discovery objectives.

4. Pretraining Objectives and Cross-Modal Alignment

4.1 Self-Supervised Pretraining Objectives

Self-supervised learning is essential due to limited labelled data. Common objectives include:

- Masked token prediction (DNA, protein, SMILES)
- Structure reconstruction (protein distances, torsions)
- Property prediction as auxiliary tasks

These objectives encourage learning biologically meaningful latent representations.

4.2 Cross-Modal Pretraining Tasks

Key cross-modal objectives include:

- **Drug–target contrastive learning**
- **Protein–ligand co-embedding**
- **Genotype–phenotype alignment**

For example, a molecule embedding is trained to align closely with the embedding of its known protein target.

4.3 Conditional Generation and Multi-Objective Optimization

Generative heads are trained to produce molecules conditioned on:

- Protein embeddings
- Disease-specific genomic context
- ADMET constraints

Reinforcement learning and diffusion models are often combined to optimize across multiple objectives simultaneously.

5. End-to-End Drug Discovery Applications

5.1 Target Identification and Prioritization

MFMs can integrate genomic association signals, protein draggability features, and chemical tractability into a unified target-ranking framework. This reduces false-positive targets and improves translational relevance.

5.2 Context-Aware Ligand Design

By conditioning molecular generation on protein structure and genomic state, MFMs enable:

- Variant-specific drug design
- Isoform-selective targeting
- Resistance-aware inhibitor generation

This is particularly valuable in oncology and antimicrobial resistance.

5.3 Polypharmacology and Network Pharmacology

MFMs naturally support multi-target reasoning by embedding drugs and proteins in a shared latent space. This enables prediction of:

- Beneficial poly-pharmacology
- Off-target liabilities
- Drug–drug interactions

5.4 Safety and ADMET Prediction

Integrating chemical structure with proteomic and genomic context improves prediction of:

- Hepatotoxicity
- Cardiotoxicity
- Metabolic stability

This has direct implications for reducing late-stage attrition.

6. Benchmarking, Evaluation, and Validation

6.1 Benchmark Datasets

Evaluation uses a combination of:

- Public benchmarks (MoleculeNet, ChEMBL)
- Structural benchmarks (PDBbind)
- Retrospective drug discovery case studies

6.2 Metrics and Validation

Metrics include:

- AUROC, RMSE, enrichment factors
- Calibration error
- Chemical validity and novelty

Prospective validation and wet-lab collaboration remain critical.

6.3 Interpretability and Explainability

Explainable AI techniques such as attention visualization, attribution maps, and counterfactual generation are increasingly required for regulatory acceptance.

7. Ethical, Regulatory, and Intellectual Property Considerations

7.1 Data Governance and Bias

Multimodal foundation models inherit biases present in their training datasets. In drug discovery, these biases may arise from overrepresentation of well-studied targets, diseases prevalent in high-income populations, or chemical scaffolds favoured by historical screening programs. Genomic datasets are known to be skewed toward populations of European ancestry, raising concerns about the generalizability of AI-derived therapeutics 38,39].

Mitigating such biases requires deliberate dataset curation, transparency in data provenance, and the incorporation of fairness-aware learning objectives. Regulatory agencies increasingly expect sponsors to document dataset composition and assess demographic representativeness, especially for precision medicine applications [40,42].

7.2 Explainability and Trustworthy AI

Regulatory bodies such as the US Food and Drug Administration (FDA) and the European Medicines Agency (EMA) emphasize transparency and interpretability in AI systems used for drug development. While MFMs are inherently complex, explainability mechanisms—such as attention maps, feature attribution, and counterfactual generation—are essential for establishing trust.

In the context of MFMs, explainability must operate across modalities. For example, a toxicity prediction should be traceable to specific molecular substructures, protein targets, and genomic pathways. Such cross-modal explanations not only support regulatory review but also provide actionable insights for medicinal chemists and biologists.

7.3 Regulatory Landscape

Regulatory guidance on AI in drug development is evolving rapidly. Recent FDA discussion papers outline expectations for model validation, lifecycle management, and human oversight for AI-enabled tools. MFMs used for hypothesis generation may face lower regulatory scrutiny, whereas models directly informing candidate selection or clinical decisions require rigorous validation.

Importantly, MFMs blur traditional boundaries between discovery and development, raising questions about how regulatory frameworks should classify AI-generated evidence. Early engagement with regulators and the use of model-informed drug development (MIDD) principles are recommended to facilitate acceptance.

7.4 Intellectual Property Challenges

The use of large pretrained models raises novel intellectual property (IP) questions. Ownership of model weights, training data, and AI-generated molecules is not always clear, particularly when models are pretrained on publicly available datasets and fine-tuned on proprietary data.

From a patent perspective, AI-generated molecules may face challenges in meeting inventorship criteria in some jurisdictions. Pharmaceutical companies are increasingly adopting hybrid strategies, combining AI-driven generation with human expert validation to strengthen IP claims.

8. Industrial Translation and Case Studies

8.1 Integration into Pharmaceutical Pipelines

Several pharmaceutical and biotechnology companies have begun integrating multimodal AI platforms into their discovery workflows. These platforms typically combine chemical and biological foundation models with proprietary screening and clinical data, enabling rapid iteration across target identification, hit generation, and lead optimization.

MFMs are particularly valuable in early discovery stages, where uncertainty is high and data is sparse. By leveraging pretrained representations, MFMs reduce dependence on large task-specific datasets and accelerate exploration of novel targets.

8.2 Representative Case Studies

AI-Enabled Antibiotic Discovery

A landmark study demonstrated the use of deep learning to identify novel antibiotics effective against multidrug-resistant pathogens by screening chemical space using learned representations [31]. While not fully multimodal, subsequent extensions incorporated protein target information and genomic resistance markers, illustrating the trajectory toward MFMs.

Structure-Guided Oncology Drug Design

Integration of AlphaFold-derived protein structures with generative chemical models has enabled rapid identification of kinase inhibitors optimized for selectivity and resistance profiles. Incorporating tumour-specific genomic variants further enhances therapeutic relevance.

Rare Disease Drug Repurposing

MFMs trained on genomic and phenotypic data have been applied to identify repurposing opportunities for rare diseases, where limited patient populations preclude large-scale trials. These applications highlight the societal impact potential of MFMs.

8.3 Human–AI Collaboration

Despite increasing automation, human expertise remains central. Medicinal chemists, structural biologists, and clinicians play critical roles in defining objectives, interpreting model outputs, and designing validation experiments. MFMs should be viewed as decision-support systems rather than autonomous discovery engines.

9. Limitations and Open Scientific Challenges

9.1 Data Quality and Completeness

While data volume has increased dramatically, high-quality paired datasets linking genomics, proteomics, and chemistry remain limited. Noisy labels, inconsistent assay conditions, and incomplete annotations pose significant challenges for model training and evaluation [48].

9.2 Computational and Environmental Costs

Training MFMs requires substantial computational resources, raising concerns about accessibility and environmental sustainability. Efficient architectures, parameter sharing, and transfer learning strategies are essential to democratize access [74,75].

9.3 Causal Reasoning and Generalization

Most MFMs learn correlations rather than causal relationships. Distinguishing causative targets from correlated biomarkers remains a fundamental challenge. Integrating causal inference frameworks and mechanistic modelling with MFMs is an important research direction.

9.4 Experimental Validation Bottlenecks

AI-generated hypotheses ultimately require experimental validation. Limited wet-lab capacity can become a bottleneck, emphasizing the need for active learning and closed-loop experimentation to prioritize the most promising candidates.

10. Future Directions and Emerging Trends

10.1 Closed-Loop and Self-Improving Systems

Future MFMs are likely to be embedded in closed-loop systems that iteratively generate hypotheses, test them experimentally, and update model parameters. Such systems could dramatically accelerate optimization cycles [49,50].

10.2 Personalized and Precision Therapeutics

Conditioning molecular design on patient-specific genomic and transcriptomic profiles opens the door to truly personalized medicines. MFMs provide a natural framework for integrating such data at scale [53].

10.3 Federated and Privacy-Preserving Learning

Federated learning enables MFMs to be trained across multiple institutions without sharing raw data, addressing privacy and IP concerns. This approach is particularly relevant for clinical and proprietary datasets [54].

10.4 Toward Autonomous Discovery Agents

Combining MFMs with large language models and reinforcement learning agents could yield semi-autonomous discovery systems capable of literature mining, hypothesis generation, and experimental planning. Careful governance will be essential to ensure responsible use.

11. CONCLUSIONS

Multimodal foundation models represent a unifying and transformative approach to drug discovery, enabling joint reasoning across genomics, proteomics, and chemical space [33,50,78]. By integrating heterogeneous biological data into coherent representations, MFMs address long-standing fragmentation in computational pipelines and offer new opportunities for context-aware molecular design, improved safety prediction, and accelerated translation.

While significant challenges remain—including data bias, interpretability, computational cost, and regulatory acceptance—the rapid pace of methodological innovation and growing industrial adoption suggest that MFMs will play a central role in next-generation pharmaceutical research. Continued collaboration between AI researchers, biologists, chemists, clinicians, and regulators will be essential to fully realize the potential of this paradigm.

REFERENCES

1. DiMasi JA, Grabowski HG, Hansen RW. Innovation in the pharmaceutical industry: New estimates of R&D costs. *J Health Econ.* 2016;47:20–33.
2. Mullard A. How much does it cost to make a drug? *Nat Rev Drug Discov.* 2020;19:777.
3. Bommasani R, Hudson DA, Adeli E, et al. On the opportunities and risks of `foundation models. *arXiv.* 2021:2108.07258.
4. Vaswani A, Shazeer N, Parmar N, et al. Attention is all you need. *Adv Neural Inf Process Syst.* 2017;30:5998–6008.
5. Devlin J, Chang MW, Lee K, Toutanova K. BERT: Pre-training of deep bidirectional transformers for language understanding. *NAACL-HLT.* 2019:4171–4186.

6. Jumper J, Evans R, Pritzel A, et al. Highly accurate protein structure prediction with AlphaFold. *Nature.* 2021;596:583–589.
7. Senior AW, Evans R, Jumper J, et al. Improved protein structure prediction using potentials from deep learning. *Nature.* 2020;577:706–710.
8. Lin Z, Akin H, Rao R, et al. Evolutionary-scale prediction of atomic-level protein structure with a language model. *Science.* 2023;379:1123–1130.
9. Rives A, Meier J, Sercu T, et al. Biological structure and function emerge from scaling unsupervised learning to 250 million protein sequences. *Proc Natl Acad Sci USA.* 2021;118:e2016239118.
10. Elnaggar A, Heinzinger M, Dallago C, et al. ProtTrans: Towards cracking the language of life’s code. *IEEE Trans Pattern Anal Mach Intell.* 2022;44:7112–7127.
11. AlQuraishi M. End-to-end differentiable learning of protein structure. *Cell Syst.* 2019;8:292–301.
12. Ji Y, Zhou Z, Liu H, Davuluri RV. DNABERT: Pre-trained bidirectional encoder representations from transformers for DNA-language. *Bioinformatics.* 2021;37:2112–2120.
13. Avsec Z, Weilert M, Shrikumar A, et al. Base-resolution models of transcription-factor binding reveal soft motif syntax. *Nat Methods.* 2021;18:354–363.
14. Kelley DR. Predicting gene expression from DNA sequence using deep learning. *Nat Methods.* 2018;15:791–794.
15. Zhou J, Troyanskaya OG. Predicting effects of noncoding variants with deep learning–based sequence model. *Nat Methods.* 2015;12:931–934.
16. Ahmad W, Simonovsky M, et al. ChemBERTa-2: Towards chemical foundation models. *arXiv.* 2022:2209.01712.
17. Gómez-Bombarelli R, Wei JN, Duvenaud D, et al. Automatic chemical design using a data-driven continuous representation. *ACS Cent Sci.* 2018;4:268–276.
18. Brown N, Fiscato M, Segler MHS, Vaucher AC. GuacaMol: Benchmarking models for de novo molecular design. *J Chem Inf Model.* 2019;59:1096–1108.
19. Walters WP, Murcko MA. Assessing the impact of generative AI on medicinal chemistry. *J Med Chem.* 2020;63:8651–8660.
20. Yang K, Swanson K, Jin W, et al. Analyzing learned molecular representations for property prediction. *J Chem Inf Model.* 2019;59:3370–3388.
21. Kearnes S, McCloskey K, Berndl M, et al. Molecular graph convolutions. *J Comput Aided Mol Des.* 2016;30:595–608.
22. Gilmer J, Schoenholz SS, Riley PF, et al. Neural message passing for quantum chemistry. *ICML.* 2017:1263–1272.
23. Zitnik M, Agrawal M, Leskovec J. Modeling polypharmacy side effects with graph convolutional networks. *Bioinformatics.* 2018;34:i457–i466.
24. Huang K, Fu T, Gao W, et al. Therapeutics data commons: Machine learning datasets and tasks for drug discovery. *NeurIPS.* 2021.
25. Kingma DP, Welling M. Auto-encoding variational Bayes. *ICLR.* 2014.
26. Popova M, Isayev O, Tropsha A. Deep reinforcement learning for de novo drug design. *Sci Adv.* 2018;4:eaap7885.
27. Ho J, Jain A, Abbeel P. Denoising diffusion probabilistic models. *NeurIPS.* 2020.

28. Schneuing A, Du Y, et al. Structure-based drug design with diffusion models. *ICML*. 2023.
29. Noé F, Tkatchenko A, Müller KR, Clementi C. Machine learning for molecular simulation. *Annu Rev Phys Chem*. 2020;71:361–390.
30. Stokes JM, Yang K, Swanson K, et al. A deep learning approach to antibiotic discovery. *Cell*. 2020;180:688–702.
31. Zhavoronkov A, Ivanenkov YA, Aliper A, et al. Deep learning enables rapid identification of potent DDR1 kinase inhibitors. *Nat Biotechnol*. 2019;37:1038–1040.
32. Gaudelet T, Malod-Dognin N, et al. Utilizing graph ML within drug discovery and development. *Brief Bioinform*. 2021;22:bbab159.
33. Guo F, Wang X, et al. Foundation models in bioinformatics. *Nat Sci Rev*. 2025;12:nwad195.
34. Pyzer-Knapp EO. Accelerating materials discovery with artificial intelligence. *Nat Rev Mater*. 2023;8:345–358.
35. Ding T, et al. A multimodal whole-slide foundation model for pathology. *Nat Med*. 2025;31:456–468.
36. Wu Z, Ramsundar B, Feinberg EN, et al. MoleculeNet: A benchmark for molecular machine learning. *Chem Sci*. 2018;9:513–530.
37. Liu Z, et al. Benchmarking molecular representations. *J Chem Inf Model*. 2021;61:3370–3388.
38. Rudin C. Stop explaining black box machine learning models. *Nat Mach Intell*. 2019;1:206–215.
39. Samek W, Montavon G, Vedaldi A, et al. Explainable AI: Interpreting, explaining and visualizing deep learning. *Springer*. 2019.
40. FDA. Artificial intelligence and machine learning in drug development. FDA Discussion Paper. 2023.
41. European Medicines Agency. Regulatory science strategy to 2025. EMA. 2024.
42. OECD. Trustworthy artificial intelligence in health systems. OECD Publishing. 2022.
43. Beam AL, Kohane IS. Big data and machine learning in health care. *JAMA*. 2018;319:1317–1318.
44. Topol EJ. High-performance medicine: The convergence of human and artificial intelligence. *Nat Med*. 2019;25:44–56.
45. Rajkomar A, Dean J, Kohane I. Machine learning in medicine. *N Engl J Med*. 2019;380:1347–1358.
46. Kitano H. Systems biology: A brief overview. *Science*. 2002;295:1662–1664.
47. Barabási AL, Gulbahce N, Loscalzo J. Network medicine. *Nat Rev Genet*. 2011;12:56–68.
48. Li T, Sahu AK, Talwalkar A, Smith V. Federated learning: Challenges, methods, and future directions. *IEEE Signal Process Mag*. 2020;37:50–60.
49. Schneider G. Automating drug discovery. *Nat Rev Drug Discov*. 2018;17:97–113.
50. Vamathevan J, Clark D, Czodrowski P, et al. Applications of ML in drug discovery. *Nat Rev Drug Discov*. 2019;18:463–477.
51. Walters WP. Virtual chemical libraries. *J Med Chem*. 2019;62:1116–1124.

52. Hughes JP, Rees S, Kalindjian SB, Philpott KL. Principles of early drug discovery. *Br J Pharmacol.* 2011;162:1239–1249.
53. Collins FS, Varmus H. A new initiative on precision medicine. *N Engl J Med.* 2015;372:793–795.
54. Ashley EA. Towards precision medicine. *Nat Rev Genet.* 2016;17:507–522.
55. Mullard A. AI in drug discovery – hype or hope? *Nat Rev Drug Discov.* 2021;20:507–509.
56. Mak KK, Pichika MR. Artificial intelligence in drug development. *Drug Discov Today.* 2019;24:773–780.
57. Lo YC, Rensi SE, Torng W, Altman RB. Machine learning in drug discovery. *Drug Discov Today.* 2018;23:1538–1546.
58. Segler MHS, Kogej T, Tyrchan C, Waller MP. Generating focused molecular libraries. *ACS Cent Sci.* 2018;4:120–131.
59. Chen H, Engkvist O, Wang Y, Olivecrona M, Blaschke T. The rise of deep learning in drug discovery. *Drug Discov Today.* 2018;23:1241–1250.
60. Walters WP, Stahl MT, Murcko MA. Virtual screening—An overview. *Drug Discov Today.* 2012;17:108–115.
61. Bajorath J. Chemoinformatics. *Wiley Interdiscip Rev Comput Mol Sci.* 2011;1:201–218.
62. Hopkins AL. Network pharmacology. *Nat Chem Biol.* 2008;4:682–690.
63. Paolini GV, Shapland RH, van Hoorn WP, Mason JS, Hopkins AL. Global mapping of pharmacological space. *Nat Biotechnol.* 2006;24:805–815.
64. Wishart DS. DrugBank. *Nucleic Acids Res.* 2018;46:D1074–D1082.
65. Gaulton A, Bellis LJ, Bento AP, et al. ChEMBL database update. *Nucleic Acids Res.* 2012;40:D1100–D1107.
66. Kim S, Chen J, Cheng T, et al. PubChem in 2021. *Nucleic Acids Res.* 2021;49:D1388–D1395.
67. UniProt Consortium. UniProt: A worldwide hub of protein knowledge. *Nucleic Acids Res.* 2019;47:D506–D515.
68. Berman HM, Westbrook J, Feng Z, et al. The Protein Data Bank. *Nucleic Acids Res.* 2000;28:235–242.
69. Lonsdale J, Thomas J, Salvatore M, et al. The GTEx project. *Nat Genet.* 2013;45:580–585.
70. Sudlow C, Gallacher J, Allen N, et al. UK Biobank. *PLoS Med.* 2015;12:e1001779.
71. Weinstein JN, Collisson EA, Mills GB, et al. TCGA Pan-Cancer analysis. *Nat Genet.* 2013;45:1113–1120.
72. Hasin Y, Seldin M, Lusis A. Multi-omics approaches. *Genome Biol.* 2017;18:83.
73. Karczewski KJ, Snyder MP. Integrative omics. *Nat Rev Genet.* 2018;19:299–310.
74. Ching T, Himmelstein DS, Beaulieu-Jones BK, et al. Opportunities and obstacles for deep learning in biology. *J R Soc Interface.* 2018;15:20170387.
75. Eraslan G, Avsec Ž, Gagneur J, Theis FJ. Deep learning in genomics. *Nat Rev Genet.* 2019;20:389–403.
76. Topol EJ. Deep medicine. *Basic Books.* 2019.

77. Mitchell M. Artificial intelligence: A guide for thinking humans. *Farrar, Straus and Giroux*. 2019.

78. Schneider G, Clark DE. Automated de novo drug design. *Nat Rev Drug Discov*. 2019;18:993–1000.