International Journal of Scientific Research & Engineering Trends
Volume 12, Issue 3, May-Jun-2026, ISSN (Online): 2395-566X

Generative and Deep Learning In Bioinformatics: A Review of Protein Structure Prediction, Sequence Representation Learning and AI-Assisted Drug Discovery

Tirandaj Fatima Khalid[1] Prof. S. S. Bhor[2] Prof. M. C. Kolse[3] Prof. S. K. Said[4]
Department Of Artificial Intelligence And Data Science Jaihind College Of Engineering Pune, Maharashtra, India;

Abstract—Artificial intelligence, and generative modelling in particular, is creating new opportunities in bioinformatics. Conventional computational biology relied on physics-based simulation, sequence alignment and manual feature engineering, each limited by computational cost and the need for expert interpretation. This paper reviews how deep learning and generative models are used in bioinformatics research. It first examines protein structure prediction, where deep learning achieved a step change in accuracy and extended to complexes of biomolecules, and discusses what such predictions can and cannot tell researchers. It then reviews representation learning on biological sequences, in which transformer models trained on large public sequence databases learn representations that capture structural and evolutionary signal and support downstream tasks such as property and variant-effect prediction. The paper surveys the families of generative models used in the field, including variational autoencoders, graph neural networks, diffusion models and transformers, at the level of their modelling assumptions, and reviews published applications in computational drug discovery such as virtual screening and candidate prioritization. It discusses data resources, evaluation practice, the central role of experimental validation, and the governance of research, including responsible-use and biosecurity frameworks. Research gaps, objectives, a comparison of model families and challenges are presented. The review concludes that these methods are decision-support tools for researchers and do not replace laboratory experimentation or institutional oversight.

Index Terms—Bioinformatics, deep learning, protein structure prediction, representation learning, generative models, computational drug discovery, model evaluation, responsible research

I. INTRODUCTION

1.1 Motivation

Biological data have grown far faster than the capacity to interpret them. Public repositories hold hundreds of millions of sequences and millions of experimentally determined structures and assay results, and no research group can examine these by hand. Bioinformatics therefore depends on computation, and increasingly on machine learning.

The clearest example is protein structure prediction. For decades, determining how a sequence folds required slow experimental methods or approximate simulation. Jumper *et al.* [1] showed with AlphaFold that deep learning can predict structures with accuracy competitive with experiment for many proteins, and Abramson *et al.* [2] extended accurate prediction to interactions among biomolecules. Vamathevan *et al.* [3] reviewed how machine learning is used across drug discovery and development.

This paper reviews these developments at the level of concepts, models and evaluation, and discusses the research practices and governance that responsible use requires.

1.2 Objectives Of The Review

- **O1:** To review deep learning for protein structure prediction and its interpretation.
- **O2:** To examine representation learning on biological sequences.
- **O3:** To compare the generative model families used in bioinformatics.
- **O4:** To review published applications in computational drug discovery.
- **O5:** To discuss evaluation, experimental validation and responsible research practice.

II. OVERVIEW OF KEY CONCEPTS

2.1 Areas Of Ai Use In Bioinformatics

Fig. 1 organizes AI use in bioinformatics into structure prediction, sequence representation learning, property and function prediction, and computational screening in drug discovery.

© 2026 IJSRET

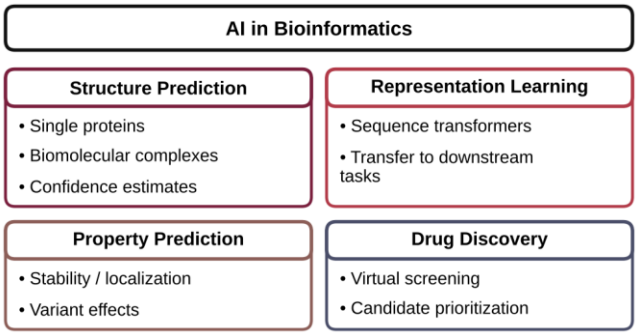


Fig. 1. Areas of AI use in bioinformatics research.

2.2 Predictive And Generative Modelling

A predictive model estimates a property from an input, such as a structure from a sequence. A generative model learns a distribution over data and can propose new samples from it. Both are statistical: their outputs are hypotheses whose value depends on how well the training data represent the question being asked, and on subsequent experimental testing.

III. REVIEW OF EXISTING WORK

Structure prediction. Jumper *et al.* [1] combined attention over evolutionary information with a geometry-aware output to predict protein structures, and reported per-residue confidence estimates that let researchers judge which regions are reliable. Abramson *et al.* [2] generalized the approach to complexes. These results reshaped how structural hypotheses are formed, although predicted structures still require experimental confirmation for mechanistic claims.

Sequence representation learning. Rives *et al.* [4] trained transformer language models on large public protein sequence databases and showed that structural and functional signal emerges in the learned representations without supervision. Lin *et al.* [5] scaled this approach to predict structure from a single sequence at reduced computational cost. Such representations are now standard inputs for downstream predictors of stability, localization and variant effects, mirroring the role of pretrained models in natural language processing [6].

Generative model families. Gómez-Bombarelli *et al.* [7] used a variational autoencoder to embed molecules in a continuous latent space in which optimization of predicted properties becomes feasible. Graph neural networks [8] operate directly on molecular graphs and are widely used for property prediction, as in MoleculeNet benchmarks [9] and chemical property models [10]. Diffusion models [11] and transformer architectures [6] provide the generative machinery most used in current work.

Computational drug discovery. Zhavoronkov *et al.* [12] reported that machine learning shortened the identification of candidate kinase inhibitors, with laboratory testing of the prioritized candidates. Vamathevan *et al.* [3] surveyed applications across target identification, screening and clinical development, and emphasized data quality as the limiting factor. Table I summarizes these works.

TABLE I Summary of Literature Survey

Ref.	Method	Key Findings	Limitations
[8]	Message passing networks	Graph learning on molecules	Property prediction only
[7]	VAE latent chemical space	Continuous optimization	Validity of samples
[9]	MoleculeNet benchmarks	Standard evaluation suite	Benchmark-reality gap
[3]	ML in drug discovery review	Mapped applications	Data quality limits
[12]	ML-assisted candidate finding	Shortened early discovery	Small validated set
[1]	AlphaFold structure prediction	Experiment-level accuracy	Function not predicted
[4]	Protein sequence transformers	Emergent structural signal	Family bias
[5]	Single-sequence structure model	Faster prediction at scale	Lower accuracy on hard cases
[2]	Prediction of complexes	Broader biomolecular scope	Needs validation

3.1 Identified Research Gaps

- Predicting biological function remains much harder than predicting structure.
- Public datasets over-represent well-studied families, biasing models.
- Computational benchmarks correlate imperfectly with experimental outcomes.
- Confidence estimates and uncertainty are often ignored by downstream users.
- Governance and review practices for AI in life-science research are still maturing.

© 2026 IJSRET

IV. SYSTEM ARCHITECTURE

Fig. 2 presents a reference framework for an AI-assisted bioinformatics research workflow, in which computational analysis is explicitly bounded by evaluation and institutional oversight.

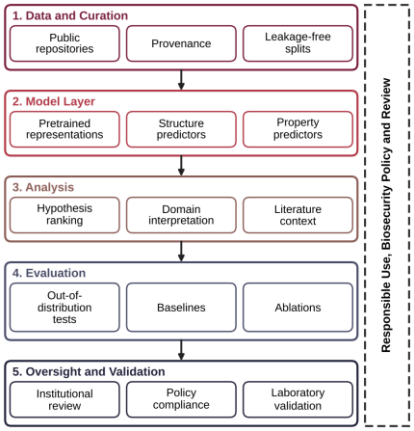


Fig. 2. Reference framework for an AI-assisted bioinformatics workflow with oversight.

4.1 Data And Curation Layer

Public sequence, structure and assay repositories are curated with provenance, de-duplication and splits that avoid leakage between training and evaluation.

4.2 Model Layer

Pretrained sequence representation models, structure predictors and property predictors are selected according to the question, with confidence estimates retained.

4.3 Analysis Layer

Researchers use model outputs to rank and interpret hypotheses, combining them with domain knowledge, literature and statistical analysis.

4.4 Evaluation Layer

Held-out and out-of-distribution benchmarks, ablations and comparison with physics-based or classical baselines test whether model outputs are informative for the specific question.

4.5 Oversight And Validation Layer

Institutional review, responsible-use and biosecurity policies, and experimental validation by qualified laboratories govern what is pursued and how results are reported.

4.6 Implementation Methodology

Fig. 3 describes the workflow at this level. Model quality on a property prediction task is reported with metrics such as

$$RMSE = \sqrt{((1/n) \Sigma (y_i - \hat{y}_i)^2)} \quad (1)$$

and ranking usefulness with correlation or enrichment measures. Because training and deployment distributions differ, evaluation emphasizes out-of-distribution splits, for example by family or time, rather than random splits alone. Confidence estimates are used to decide which outputs are worth following up at all.

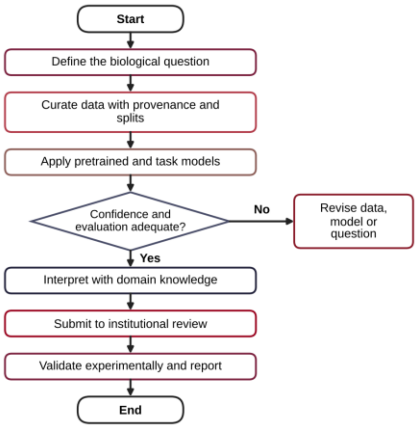


Fig. 3. AI-assisted bioinformatics research workflow.

V. PERFORMANCE COMPARISON

Table II compares generative and predictive model families by the data they consume, their modelling assumptions and their typical research role.

TABLE II Comparison of Model Families Used in Bioinformatics

Family	Input	Main Role	Key Limitation
Sequence transformer	Sequences	Representations	Data-family bias
Graph neural network	Molecular graphs	Property prediction	Needs labels
Variational autoencoder	Molecules	Latent optimization	Sample validity
Diffusion model	Structured data	Distribution modelling	Compute cost
Physics-based simulation	Structures	Mechanistic baseline	Very expensive

5.1 Open Issues And Future Directions

Table III summarizes challenges. Future directions include better function prediction, uncertainty-aware models, standardized out-of-distribution benchmarks, data sharing

© 2026 IJSRET

with provenance, and clearer institutional and policy frameworks for AI-assisted life-science research.

TABLE III Challenges and Mitigation

Challenge	Mitigation
Function prediction	Multimodal data; assay integration
Dataset bias	Curated, diverse repositories
Benchmark-reality gap	Out-of-distribution evaluation
Ignored uncertainty	Confidence-aware reporting
Responsible use	Institutional review; policy compliance

VI. CONCLUSION

Deep learning and generative modelling have changed bioinformatics practice, most visibly in protein structure prediction and in learned representations of biological sequences, and they now support computational stages of drug discovery. The reviewed literature also shows the limits of these methods: they interpolate from biased data, their outputs are hypotheses with uncertainty, and computational benchmarks are not evidence of biological behaviour. Generative AI should therefore be treated as a decision-support tool within a research process whose conclusions rest on experimental validation and institutional oversight.

6.1 Acknowledgment

The authors express their sincere gratitude to the Department of Artificial Intelligence and Data Science, Jaihind College of Engineering, Pune, for providing the facilities, guidance and encouragement needed to carry out this work. The authors also thank the faculty members and colleagues whose valuable suggestions helped improve the quality of this review.

REFERENCES

1. J. Jumper et al., “Highly accurate protein structure prediction with AlphaFold,” Nature, vol. 596, no. 7873, pp. 583–589, 2021.
2. J. Abramson et al., “Accurate structure prediction of biomolecular interactions with AlphaFold 3,” Nature, vol. 630, pp. 493–500, 2024.
3. J. Vamathevan et al., “Applications of machine learning in drug discovery and development,” Nat. Rev. Drug Discov., vol. 18, no. 6, pp. 463–477, 2019.
4. A. Rives et al., “Biological structure and function emerge from scaling unsupervised learning to 250 million protein sequences,” Proc. Natl. Acad. Sci. USA, vol. 118, no. 15, Art. no. e2016239118, 2021.
5. Z. Lin et al., “Evolutionary-scale prediction of atomic-level protein structure with a language model,” Science, vol. 379, no. 6637, pp. 1123–1130, 2023.
6. A. Vaswani et al., “Attention is all you need,” in Proc. Adv. Neural Inf. Process. Syst. (NeurIPS), vol. 30, 2017.
7. R. Gómez-Bombarelli et al., “Automatic chemical design using a data-driven continuous representation of molecules,” ACS Cent. Sci., vol. 4, no. 2, pp. 268–276, 2018.
8. J. Gilmer, S. S. Schoenholz, P. F. Riley, O. Vinyals, and G. E. Dahl, “Neural message passing for quantum chemistry,” in Proc. ICML, 2017, pp. 1263–1272.
9. Z. Wu et al., “MoleculeNet: A benchmark for molecular machine learning,” Chem. Sci., vol. 9, no. 2, pp. 513–530, 2018.
10. K. Yang et al., “Analyzing learned molecular representations for property prediction,” J. Chem. Inf. Model., vol. 59, no. 8, pp. 3370–3388, 2019.
11. J. Ho, A. Jain, and P. Abbeel, “Denoising diffusion probabilistic models,” in Proc. NeurIPS, vol. 33, 2020, pp. 6840–6851.
12. A. Zhavoronkov et al., “Deep learning enables rapid identification of potent DDR1 kinase inhibitors,” Nat. Biotechnol., vol. 37, no. 9, pp. 1038–1040, 2019.

© 2026 IJSRET