Tech & AI Systems | Jan 13, 2026 | 11 min read | By Bryan White

The Code of Life: How Large Language Models are Designing New Proteins

Biology * Public Health * AI

Introduction: From Sequence Prediction to Generative Biology

The integration of Large Language Models (LLMs) and Generative Artificial Intelligence (GenAI) into the natural sciences represents a foundational shift from predictive classification to de novo molecular generation.[1][5] Whereas earlier machine learning architectures primarily annotated existing biological datasets, the 2024-2025 generation of foundation models learns the underlying syntax and semantics of biological sequences -- whether composed of nucleotides, amino acids, or chemical bond graphs -- to write novel proteins, plan multi-step chemical syntheses, and formulate testable mechanistic hypotheses.[3][6]

Bibliometric analyses from early 2025 confirm that publications utilizing generative foundation models in the life sciences have surpassed traditional single-task deep learning studies, paralleling a rapid expansion in FDA-authorized AI medical devices.[1][2] Across bioinformatics, chemoinformatics, and pharmacoinformatics, generative models now simulate hundreds of millions of years of evolution in silico, compress solid-state materials discovery from years to weeks, and act as autonomous "AI Co-Scientists" in elucidating antimicrobial resistance pathways.[13][28][36]

Tokenization and Multimodal Transformer Architectures

Applying the Transformer architecture to molecular science relies on self-attention mechanisms that compute pairwise interactions across every position in a sequence.[7] In protein folding, residues separated by hundreds of positions along the linear polypeptide chain (primary structure) converge in three-dimensional space (tertiary structure) to form a catalytic active site.[8][9] Self-attention matrices directly capture these distal physical couplings across three core tokenization paradigms:

* Genomic Byte Pair Encoding (BPE): To process a 3-billion-base-pair human genome, early models relied on fixed overlapping k-mers, creating bloated vocabularies. Modern genomics models such as DNABERT-2 replace k-mers with Byte Pair Encoding, merging frequent nucleotide motifs (such as promoter TATA boxes or transcription factor binding sites) into variable-length tokens and matching state-of-the-art accuracy with 21-fold fewer parameters.[11][18]

* Chemical Atom Pair Encoding (APE): Flattening 3D molecular graphs into 1D SMILES strings often produces invalid valences under generic text tokenizers. While SELFIES enforces a 100% valid formal grammar, coupling SMILES with Atom Pair Encoding (APE) -- which groups bonded chemical environments -- significantly outperforms BPE across biophysics benchmarks.[12][24]

* Multimodal Latent Spaces: Rather than treating biological translation and folding solely as 1D text, frontier models such as ESM3 and AlphaFold 3 unify 1D sequence

strings, 3D atomic coordinates, and functional annotations within a shared generative architecture.[13][16]

Bioinformatics: Protein and Genomic Language Models

Unlike classical physics-based simulations that estimate conformational Gibbs free energy landscapes through brute-force molecular dynamics, Protein Language Models (PLMs) extract evolutionary constraints directly from billions of natural sequences.[9][15] In 2024, EvolutionaryScale released ESM3, a frontier multimodal PLM trained to reason jointly across sequence, structure, and function.[13] Prompted via chain-of-thought constraints to coordinate chromophore active-site residues within a beta-barrel scaffold, ESM3 generated a novel functional fluorescent protein, esmGFP, sharing only 58% sequence identity with its closest natural homolog -- equivalent to roughly 500 million years of natural selection and random mutation.[14]

Whereas ESM3 functions as a generative writer of novel enzymes and biologics, Google DeepMind's AlphaFold 3 excels as an analytical structural engine, jointly predicting high-accuracy biomolecular complexes across proteins, DNA, RNA, and small-molecule ligands.[16][17] In modern computational workflows, researchers pair generative and predictive models in tandem: a candidate sequence is synthesized in silico by ESM3 to satisfy desired catalytic or geometric constraints and subsequently evaluated by AlphaFold 3 to verify ligand-binding pocket stability prior to physical expression. In evolutionary biology, conversational frameworks such as NJGPT automate sequence alignment and neighbor-joining phylogenetic tree construction from natural-language prompts, while researchers have reciprocally applied phylogenetic distance algorithms to trace architectural speciation across encoder-decoder and decoder-only LLM families.[19][20][21] At the same time, because public training repositories like UniProt heavily overrepresent model organisms (such as Escherichia coli, mice, and humans), parameter-efficient fine-tuning is increasingly deployed to overcome sequence-similarity bias across underrepresented viral and extremophile proteomes.[22][23]

Chemoinformatics: Molecular Grammars and Autonomous Synthesis

To prevent structural hallucinations when generating small molecules, researchers at MIT and IBM introduced a hybrid architecture in 2025 pairing an LLM planner with Graph Neural Networks (GNNs) operating directly on molecular topology, elevating valid synthesis plan rates from 5% to 35% and outperforming text-only models ten times larger.[4] Beyond static design, tool-augmented chemical agents such as ChemCrow, CheMatAgent, and ChemActor integrate reaction databases, retrosynthesis tree-search algorithms, and hazard verification tools to plan multi-step organic syntheses while screening out explosive or toxic intermediates before laboratory execution.[25][26][27]

In inorganic materials science, Microsoft and the Pacific Northwest National Laboratory (PNNL) validated this generative-screening pipeline using Azure Quantum Elements.[28] Starting from 32 million candidate inorganic crystal structures, AI foundation models filtered the space down to 500,000 stable compounds and 18 viable solid-state battery electrolytes in just 80 hours -- a screening campaign that would traditionally require years of density functional theory calculations -- enabling

rapid wet-lab synthesis of a reduced-lithium prototype battery.[10][28][29]

Pharmacoinformatics, Agentic Co-Scientists, and Validation Risks

In therapeutic discovery, platforms such as Insilico Medicine's PandaOmics combine multi-omics knowledge graphs with LLM literature mining to detect "attention spikes" around emerging gene-disease targets.[30][31] Downstream, benchmarks including DDI-JUDGE and Ab-VS-Bench show that fine-tuned LLMs (AUC ~0.788) and zero-shot models (AUC ~0.642) rapidly pre-screen drug interactions and antibody binding affinities while optimizing multi-objective ADMET profiles.[32][33][34]

In early 2025, Google's multi-agent AI Co-Scientist demonstrated autonomous hypothesis generation when tasked with explaining how chromosomal phage-inducible chromosomal islands (cf-PICIs) drive horizontal gene transfer across bacteria.[36] Using an internal tournament of agents, the system hypothesized a "tail piracy" mechanism -- wherein genetic elements hijack capsid tails from diverse bacteriophages via adaptor and connector proteins -- independently matching unpublished experimental findings that had taken human microbiologists nearly a decade to confirm.[35][36]

Nevertheless, critical challenges persist. Chemical hallucinations require mitigation via Retrieval-Augmented Generation (RAG) and multi-hop prompt optimization (MIPRO).[37] Crucially, evaluations by the National Institute of Standards and Technology (NIST) using safe biological proxies revealed a pronounced "validation gap": many AI-generated protein homologs predicted to fold cleanly proved catalytically inert in wet-lab assays, underscoring the necessity of physical experimental verification.[38][39]

| Scientific Domain | | Foundation Model / System | | Core Architecture | | Validated Breakthrough (2024-2025) |

| Generative Proteomics | | ESM3 (EvolutionaryScale) | | Multimodal sequence-structure-function PLM | | Designed de novo fluorescent protein (esmGFP) across 500M years of evolutionary distance.[13][14] |

| Structural Biology | | AlphaFold 3 (DeepMind) | | Diffusion biomolecular complex predictor | | High-accuracy joint modeling of protein, DNA, RNA, and ligand interactions.[16] |

| Genomics | | DNABERT-2 | | Byte Pair Encoding (BPE) Transformer | | Captures long-range regulatory dependencies with 21x fewer parameters.[11][18] |

| Synthetic Chemistry | | MIT-IBM Graph-LLM & ChemCrow | | Hybrid LLM-GNN & tool-augmented agents | | Boosted valid synthesis route generation from 5% to 35%; automated hazard planning.[4][25] |

| Materials Discovery | | Azure Quantum Elements | | HPC-accelerated GNN screening | | Screened 32M inorganic candidates in 80 hours to synthesize a solid-state electrolyte.[28][29] |

| Autonomous Discovery | | AI Co-Scientist (Google) | | Multi-agent tournament reasoning system | | Independently derived bacteriophage "tail piracy" horizontal gene transfer mechanism.[35][36] |

Conclusion: The Era of Self-Driving Science

Generative AI in the molecular and life sciences has matured from isolated predictive models into integrated, multimodal infrastructure.[40] As multi-agent reasoning engines like the AI Co-Scientist converge with automated cloud laboratories and high-throughput wet-lab validation loops, computational biology and chemistry are entering an era of "self-driving science." Sustaining this acceleration will depend on closing the experimental validation gap, expanding sequence diversity across neglected branches of the tree of life, and enforcing rigorous biosecurity benchmarks.

Updates... (Nov 22, 2026): Replaced bot-blocked OpenReview citations (#20 and #27) with verified arXiv preprints (Yax et al., Apr 6, 2024, arXiv:2404.04671; Wang et al., May 11, 2025, arXiv:2505.07027), added exact publication dates and primary authorship across all 40 references, and standardized reference section markup. (Prior Nov 14, 2026 update standardized bibliographic records to canonical peer-reviewed and preprint DOIs alongside official FDA medical-device authorization listings.)

References & Primary Sources

* [1] Awasthi, R., Ramachandran, S., Mishra, S., Mahapatra, D., et al. (Feb 27, 2025). Artificial Intelligence in Healthcare: 2024 Year in Review. medRxiv. -- https://doi.org/10.1101/2025.02.26.25322978

* [2] U.S. Food and Drug Administration (FDA) (Sep 22, 2026). List of Artificial Intelligence-Enabled Medical Devices. -- https://www.fda.gov/medical-devices/artificial-intelligence-enabled-medical-devices/list-artificial-intelligence-enabled-medical-devi

* [3] Chakraborty, C., Bhattacharya, M., Pal, S., Chatterjee, S., et al. (Dec 2025). AI-enabled language models (LMs) to large language models (LLMs) and multimodal large language models (MLLMs) in drug discovery and development. Journal of Advanced Research. -- https://doi.org/10.1016/j.jare.2025.02.011

* [4] MIT News (Apr 9, 2025). Could LLMs help design our next medicines and materials? -- https://news.mit.edu/2025/could-llms-help-design-our-next-medicines-and-materials-0409

* [5] Reddy, C. K., & Shojaee, P. (Dec 16, 2024). Towards Scientific Discovery with Generative AI: Progress, Opportunities, and Challenges. arXiv:2412.11427. -- https://arxiv.org/html/2412.11427v1

* [6] MIT News (Dec 19, 2024). Need a research hypothesis? Ask AI. -- https://news.mit.edu/2024/need-research-hypothesis-ask-ai-1219

* [7] Chen, J. Y., et al. (Jan 21, 2025). Evaluating the advancements in protein language models for encoding strategies in protein function prediction: a comprehensive review. Frontiers in Bioengineering and Biotechnology. -- https://doi.org/10.3389/fbioe.2025.1506508

* [8] Wang, L., Li, X., Zhang, H., Wang, J., et al. (Feb 8, 2025). A Comprehensive Review of Protein Language Models. arXiv:2502.06881. -- https://arxiv.org/html/2502.06881v1

* [9] Nguyen, V. T. D., & Hy, T.-S. (Mar 20, 2025). Advances in Protein Representation Learning: Methods, Applications, and Future Directions. arXiv:2503.16659. -- https://arxiv.org/html/2503.16659v1

* [10] Chen, C., Nguyen, D., Lee, S., Baker, N., et al. (Jul 9, 2024). Accelerating Computational Materials Discovery with Machine Learning and Cloud High-Performance Computing: from Large-Scale Screening to Experimental Validation. Journal of the American Chemical Society. -- https://doi.org/10.1021/jacs.4c03849

* [11] Zhou, Z., Ji, Y., Li, W., Dutta, P., et al. (Jun 26, 2023; revised Mar 18, 2024). DNABERT-2: Efficient Foundation Model and Benchmark For Multi-Species Genome. arXiv:2306.15006. -- https://arxiv.org/html/2306.15006v2

* [12] Leon, M., et al. (Oct 23, 2024). Comparing SMILES and SELFIES tokenization for enhanced chemical language modeling. Scientific Reports. -- https://doi.org/10.1038/s41598-024-76440-8

* [13] EvolutionaryScale (Jun 25, 2024). ESM3: Simulating 500 million years of evolution with a language model. -- https://www.evolutionaryscale.ai/blog/esm3-release

* [14] Hayes, T., Rao, R., Akin, H., Sofroniew, N., et al. (Feb 21, 2025). Simulating 500 million years of evolution with a language model. Science. -- https://doi.org/10.1126/science.ads0018

* [15] Bepler, T., & Berger, B. (Jun 2021). Learning the protein language: Evolution, structure, and function. Cell Systems. -- https://doi.org/10.1016/j.cels.2021.05.017

* [16] Abramson, J., Adler, J., Dunger, J., Evans, R., et al. (May 8, 2024). Accurate structure prediction of biomolecular interactions with AlphaFold 3. Nature. -- https://doi.org/10.1038/s41586-024-07487-w

* [17] Future Medicine AI (Jun 27, 2024). ESM3 | A New Frontier for Advanced Protein Design. -- https://www.fmai-hub.com/esm3-a-new-frontier-for-advanced-protein-design/

* [18] Feng, H., et al. (Aug 2024). Benchmarking DNA Foundation Models for Genomic and Genetic Tasks. bioRxiv. -- https://doi.org/10.1101/2024.08.16.608288

* [19] Wang, Z., Huang, H., Li, T., Rodrigo, A., et al. (Dec 4, 2024). NJGPT: A Large Language Model-Driven, User-Friendly Solution for Phylogenetic Tree Construction. bioRxiv. -- https://doi.org/10.1101/2024.12.02.626464

* [20] Yax, N., Oudeyer, P.-Y., & Palminteri, S. (Apr 6, 2024). PhyloLM: Inferring the Phylogeny of Large Language Models and Predicting their Performances in Benchmarks. arXiv:2404.04671. -- https://arxiv.org/abs/2404.04671

* [21] Wu, Z., Zhao, H., Wang, Z., Guo, J., Wang, Q., & He, B. (Sep 29, 2025). LLM DNA: Tracing Model Evolution via Functional Representations. arXiv:2509.24496. -- https://arxiv.org/html/2509.24496v1

* [22] Ding, F., & Steinhardt, J. (Mar 12, 2024). Protein language models are biased by unequal sequence sampling across the tree of life. bioRxiv. -- https://doi.org/10.1101/2024.03.07.584001

* [23] Sawhney, R., Ferrell, B., Dejean, T., Schreiber, Z., et al. (Sep 18, 2025).

Fine-tuning protein language models unlocks the potential of underrepresented viral proteomes. PeerJ. -- https://doi.org/10.7717/peerj.19919

* [24] Jang, Y., Kim, J., & Ahn, S. (2025). Improving Chemical Understanding of LLMs via SMILES Parsing. Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing (EMNLP). -- https://doi.org/10.18653/v1/2025.emnlp-main.791

* [25] M. Bran, A., et al. (May 8, 2024). Augmenting large language models with chemistry tools. Nature Machine Intelligence. -- https://doi.org/10.1038/s42256-024-00832-8

* [26] Yu, B., Baker, F. N., Chen, Z., Herb, G., et al. (Nov 11, 2024). ChemToolAgent: The Impact of Tools on Language Agents for Chemistry Problem Solving. arXiv:2411.07228. -- https://arxiv.org/html/2411.07228v2

* [27] Wang, H., Guo, J., Kong, L., Ramprasad, R., & Schwaller, P. (May 11, 2025). LLM-Augmented Chemical Synthesis and Design Decision Programs. arXiv:2505.07027. -- https://arxiv.org/abs/2505.07027

* [28] Microsoft Source (Jan 9, 2024). Discoveries in weeks, not years: How AI and high-performance computing are speeding up scientific discovery. -- https://news.microsoft.com/source/features/innovation/how-ai-and-hpc-are-speeding-up-scientific-discovery/

* [29] Utility Dive (Jan 10, 2024). DOE lab, Microsoft find new battery material in AI-based energy storage research initiative. -- https://www.utilitydive.com/news/pnnl-microsoft-artificial-intelligence-research-battery/704176/

* [30] Insilico Medicine (Sep 30, 2025). Showcasing gen-AI breakthroughs, Insilico Medicine presents fall edition of Pharma.AI quarterly launch. -- https://insilico.com/news/rd037pu8d1-showcasing-gen-ai-breakthroughs-insilico

* [31] Insilico Medicine (2025). Prioritization Scores | PandaOmics Help Documentation. -- https://insilico.com/pandaomics/help/scores

* [32] Li, T., Fang, Z., Zhang, X., Tang, K., et al. (Jul 11, 2025). DrugLM: A Unified Framework to Enhance Drug-Target Interaction Predictions by Incorporating Textual Embeddings via Language Models. bioRxiv. -- https://doi.org/10.1101/2025.07.09.657250

* [33] Qi, H., et al. (Jun 2, 2025). Improving drug-drug interaction prediction via in-context learning and judging with large language models. Frontiers in Pharmacology. -- https://doi.org/10.3389/fphar.2025.1589788

* [34] Zhang, Y., & Tsuda, K. (Dec 15, 2025). Ab-VS: Evaluating Large Language Models for Virtual Antibody Screening via Antibody-Antigen Interaction Prediction. bioRxiv. -- https://doi.org/10.1101/2025.07.26.666985

* [35] Penad s, J. R., et al. (Sep 9, 2025). AI mirrors experimental science to uncover a mechanism of gene transfer crucial to bacterial evolution. Cell. -- https://doi.org/10.1016/j.cell.2025.08.018

* [36] Google Research (Feb 19, 2025). Accelerating scientific breakthroughs with an AI co-scientist. -- https://research.google/blog/accelerating-scientific-breakthroughs-with-an-ai-co-scientist/

* [37] Reed, S. M. (Apr 22, 2025). Augmented and Programmatically Optimized LLM Prompts Reduce Chemical Hallucinations. Journal of Chemical Information and Modeling. -- https://doi.org/10.1021/acs.jcim.4c02322

* [38] National Institute of Standards and Technology (NIST) (Jun 20, 2025). Experimental Evaluation of AI-Driven Protein Design Risks Using Safe Biological Proxies. -- https://www.nist.gov/publications/experimental-evaluation-ai-driven-protein-design-risks-using-safe-biological-proxies

* [39] Drug Target Review (Sep 11, 2025). Making sense of AI: bias, trust and transparency in pharma R&D. -- https://www.drugtargetreview.com/article/183358/making-sense-of-ai-bias-trust-and-transparency-in-pharma-rd/

* [40] Ishida, S., Sato, T., Honma, T., & Terayama, K. (Mar 24, 2025). Large language models open new way of AI-assisted molecule design for chemists. Journal of Cheminformatics. -- https://doi.org/10.1186/s13321-025-00984-8