Published in final edited form as:

Trends Genet. 2026 May ; 42(5): 423–433. doi:10.1016/j.tig.2026.01.013.

Machine learning for evolutionary genetics and molecular evolution

Nicolas Svetec[1], **UnJin Lee**[1], **Li Zhao**[1,*]

[1]Laboratory of Evolutionary Genetics and Genomics, The Rockefeller University, New York, NY 10065, USA

Abstract

Over the past decade, the rapid expansion of large-scale data and advances in computational power have allowed machine learning, especially deep learning, to reshape many areas of biological research. Evolutionary genetics and molecular evolution are also poised for a similar transformation. In this review, we discuss key advances and ongoing challenges in applying machine learning to the study of genetics and evolution, and we highlight the potential of artificial intelligence to connect genotype, phenotype, and evolutionary history.

Keywords

Artificial intelligence; machine learning; deep learning; population genetics; molecular evolution

Introduction:

Artificial intelligence and evolutionary genetics: from parallel development to new intersections

Unlike more traditionally technology-driven disciplines, the development of evolutionary genetic theories often precedes the experimental capabilities required to validate them. For much of the first half of the twentieth century, evolutionary genetics had already developed complex theoretical models and mathematical frameworks that sought to explain how genetic variation is shaped by mutation, selection, recombination, and drift [1–3]. These frameworks preceded the establishment of DNA-based inheritance [4,5] and

*Correspondence: lzhao@rockefeller.edu.

Author Contribution

N. S. and L.Z. conceived the project. L.Z., N.S., and U.L wrote the manuscript. The authors thank Sasha Mills for critically reading the manuscript.

Declaration of interests

The authors declare no competing interests.

Declaration of Generative AI and AI-assisted technologies in the writing process

During the preparation of this work, the authors used ChatGPT and Claude to assist with grammar editing and glossary preparation. After using these tools, the authors reviewed and edited the content as needed and take full responsibility for the content of the publication.

the emergence of sequence-level or nucleotide-level assessment of homology [6]. The sequencing revolution of the 1970s through 1990s [7,8], followed by explosive genomic growth in the 2000s [9–11], enabled these models to be tested at scale, as biological science started the era of genomic "big data." Investigations of population-level genomes [12], functional genomic datasets [13], and large-scale evolutionary studies [14] have revealed both the power and the limitations of traditional statistical and computational approaches. Despite these advances, simple modeling alone often cannot accurately explain the complex, nonlinear, and high-dimensional processes of evolution, calling for new methods to move the field forward.

In parallel, artificial intelligence (AI, see Glossary), which originated in the 1950s with early symbolic and rule-based systems [15] but experienced limited progress in application for decades [16–19], has undergone a dramatic resurgence since the early 2000s. With the explosion of data, advances in computational hardware such as graphical processing units (GPUs) [20], and breakthroughs in deep learning algorithms [16,21] such as multi-headed attention/transformer models [22,23], AI has rapidly evolved from a theoretical concept into a transformative tool across many domains [24,25], including biology [26,27]. From predicting protein structures to classifying cell types and modeling regulatory networks, AI has uncovered patterns inaccessible to conventional methods. Evolutionary genetics and molecular evolution now stand at the verge of similar breakthroughs. Compared to genomics studies, addressing questions in evolutionary genetics requires tools that integrate across levels of biological organization, from DNA sequences to phenotypes to evolutionary histories, and across scales of time and population size.

The study of evolutionary processes has long relied on statistical models that attempt to extract signals of history, selection, and demographic history from genetic variation. While these methods have provided deep insights, they often struggle with the complexity of genomic and population genomic datasets now available. Alternatively, machine learning (ML) is uniquely capable of integrating diverse data types to detect increasingly subtle patterns (Table 1). ML is broadly defined as computational methods that learn predictive patterns from data irrespective of model architecture, including, but not limited to, more traditional regression, clustering, and/or kernel-based approaches, as well as newer artificial neural network-based models [15,28–30]. After training, the computational time of these ML models is often highly scalable with proper resources [22,29], enabling unprecedented ability to rapidly generate large numbers of accurate predictions. Depending on the accuracy and quality of these predictions, machine learning and deep learning can thus provide new avenues for evolutionary biologists to detect evolutionary patterns [17,31], integrate heterogeneous data types, and move beyond assumptions of simple parametric models. Although these approaches are very promising, there are still technical barriers and, at times, a lack of clear and interpretable advantages of these tools in evolutionary genetics. Therefore, we aim to provide a general overview of the broader application of ML in evolutionary genetics while also highlighting areas for continued improvement in its use across various areas of evolutionary biology.

There are nuances for the terms associated with AI (Table 1). Since deep learning is a subset of machine learning, itself a branch of AI, we will use the term machine learning

(ML) broadly in this paper to include both traditional machine learning and deep learning approaches, although most highlighted advances in this review used deep learning (Figure 1). Our goal is to outline the technical opportunities and the conceptual shifts needed for ML to become a central tool in understanding genome evolution and the emergence of new functions (Figure 2). We aim to discuss how advances in ML may accelerate discovery in evolutionary genetics, highlight key achievements and directions (Figure 2), and critically examine the challenges that remain.

Main Text:

Deep learning for predicting protein structure and sequence function in molecular evolution

A central goal of molecular evolution is to connect genetic variation with molecular function and, ultimately, organismal fitness. One challenge is the difficulty of predicting how sequence changes alter structure, regulation, and phenotype. ML's most immediate impact on the study of molecular evolution is perhaps most evident in the fields of protein structure and regulatory function prediction. The development of AlphaFold marked a turning point in structural biology [32,33], offering scalable and accurate predictions of protein 3D structures directly from amino acid sequences. In evolutionary genetics, this breakthrough enables ancestral sequence reconstructions to be linked to structural models [34,35], allowing researchers to infer how historical substitutions shaped protein folding, stability, and function beyond single-gene studies [36]. In addition to AlphaFold, other models such as RoseTTAFold [37], ESMFold [38,39], and diffusion-based generative approaches [40] are also expanding predictions of individual proteins to multi-protein complexes, membrane proteins, and intrinsically disordered regions, all of which are central to understanding protein evolution. For instance, AlphaFold2 predictions and ancestral reconstruction revealed that sex peptide–receptor binding arose by reusing an ancestral myoinhibitory peptide interface [35], with mutations adding sex peptide–specific contacts and shedding light on the origin of ligand–receptor interactions. While AlphaFold implicitly relies on the evolutionary conservation of protein structure as a guiding biological principle, potentially resulting in insufficient sensitivity for detection of evolutionary changes in protein structures, language models have been shown to perform similarly to AlphaFold for novel protein structural prediction [34,41]. Promisingly, similar language model approaches have successfully revealed the convergent evolution of more distantly related proteins [42], suggesting that such limitations may not be a significant concern. However, parallel development of specialized algorithms, such as those specifically predicting properties of intrinsically disordered proteins [43], along with the rapidly increasing availability of high-quality training datasets are expected to directly address many issues soon [35].

Decoding gene expressional regulation

Deep neural networks trained on functional genomics datasets, such as chromatin accessibility and transcription factor binding, can predict the regulatory effects of sequence variation with remarkable accuracy [27,44]. For example, multiple independent studies have shown that assays for transposase-accessible chromatin with sequencing (ATAC-seq) peaks and DNase-seq peaks [45] are highly predictable [46,47] and that the regulatory

code is robust and conserved in close or distant lineages [46,48]. Since the underlying regulatory mechanisms are conserved, many tools not specifically designed for evolutionary questions may require only minor adjustments to be used in evolutionary studies. For example, tools used for predicting enhancer changes using in-silico mutagenesis may also be used in an evolutionary context [46,48]. There are also a growing number of studies that aim to predict transcription factor binding [49,50]. These models make it possible to study how noncoding mutations influence gene expression, enhancer turnover, and the evolution of gene regulatory networks, and how to use these models to examine their effects from a population or evolutionary genetics perspective. The ongoing and next stage of decoding expression is to quantitatively predict how changes in expression affect cellular and morphological phenotypes.

The prediction of protein fitness landscapes and contemporary evolution

A long-standing challenge in molecular evolution is predicting how genetic changes will shape future evolutionary outcomes. By integrating large-scale mutagenesis data and molecular dynamics simulations, ML models can now map protein fitness landscapes reasonably well [51,52], estimating the effects of single or multiple mutations on stability, binding affinity [35], and enzymatic activity [53,54]. Notably, current investigations into protein fitness landscapes are still mainly limited to a small number of genes that still require experimental validation of accuracy. Moreover, generalizing local sequence rules to the global landscape remains largely unreliable [52]. Given the importance of understanding fitness landscapes for protein engineering and design, we expect continued development of more advanced models in this area.

Evolutionary trajectory prediction is particularly relevant for human health in the context of future viral and pathogen evolution. AI-driven models have been used to forecast likely mutational pathways in influenza [55], HIV [56], and SARS-CoV-2 [57], helping to anticipate antigenic shifts that could undermine vaccines or other medical interventions [55]. Unlike traditional models that rely on fixed assumptions about selective pressures, ML approaches may learn directly from global sequencing data and metadata, such as vaccine efficacy, to identify subtle patterns of convergent evolution or immune escape. As datasets continue to grow, ML has the potential to serve as an early warning system for emerging pathogens, bridging molecular evolution with public health.

Phylogenetics and species tree inference

Reconstructing phylogenetic relationships remains a central challenges in phylogenetics. Traditional likelihood-based or Bayesian methods become computationally challenging or intractable for large datasets or complex models of sequence evolution. Recent work demonstrates that deep learning architectures, such as convolutional and recurrent neural networks (CNNs and RNNs), are now able to infer tree topologies directly from sequence alignments [58,59], which may reduce computational time for real genomic data while maintaining accuracy [60]. ML may also guide the resolution of long-standing issues in phylogenetics [61], such as gene-tree discordance caused by incomplete lineage sorting [62], hybridization[63], or adaptive radiation [63]. The overall goal is that by training models on simulated data with known evolutionary histories, ML systems can eventually learn to

correct discordant signals and improve the accuracy of species tree inference. However, such models are liable to learn incorrect statistical properties and may therefore suffer from similar misspecifications or model violations as likelihood-based methods [64]. It is thus important to evaluate the robustness and accuracy of current or future new ML approaches [59,61,65], as correct phylogenetic reconstruction is essential for downstream interpretation of evolutionary processes [66].

Detecting selection, adaptation, introgression, and other population biology

Separating the footprints of selection from population demographic history, which are generally inferred using maximum-likelihood methods, has traditionally proven difficult, as both processes share the same measurable genomic hallmarks of reduced diversity, extended haplotypes, and skewed allele frequency spectra. Unlike parametric inference models that often rely on simplifying assumptions, ML-based modeling approaches thus have the potential to effectively integrate over multiple summary statistics, genomic windows, or raw genomic data in order to flexibly and accurately capture nonlinear relationships even when such signals prove difficult to extract using inference-based methods. Such ML approaches are shown to be effective at integrating multiple weak signals into robust classifiers when tested on labeled, simulated data, enabling the distinction between selective sweeps and demographic confounders [67,68]. Deep learning models have shown promise in automating genome scans, learning directly from raw haplotype data or summary statistics, and scaling to whole-genome datasets. For example, S/HIC, an early attempt at using an ML approach for population genetics accurately detected selective sweeps under human-relevant demographic histories [69]. Several other ML approaches [68,70–72] have demonstrated accurate distinction between signatures of demography and selection while identifying genomic regions experiencing soft or hard selective sweeps in fruit fly populations. Deep learning models are increasingly applied to detect introgression and admixture [73], thereby learning complex patterns of haplotype sharing that are difficult to capture with traditional summary statistics [73]. Unfortunately, as in phylogenetic tree reconstruction, ML models often perform better on simulated data than on real data in this domain, and ongoing work continues to focus on these challenges to address the “domain adaptation” problem [74,75]. Integration with tools such as ancestral recombination graph (ARG) [76,77] may enable further scaling of deep learning models to petabyte-scale datasets, as is more common in human population genomics [78]. Although there are similar challenges as other applications of ML [61,67], these and future approaches may open the door to automated detection of signals of adaptive evolution across species and populations.

Although a growing number of studies have been published, the application of deep learning in population genetics remains a relatively new frontier [79]. Most current architectures, such as DNNs, CNNs, and RNNs, are often adopted with limited customization to the unique statistical and biological properties of genetic and genomic data. An emerging direction applies machine learning directly to raw SNP data or allele frequency spectra, bypassing human-defined summary statistics. This strategy offers computational convenience and the potential to uncover informative features that capture previously unrecognized patterns of genetic variation. For example, ReLERNN demonstrated that recombination rates can be inferred from pooled sequencing data using allele frequency

patterns [80], and subsequent work showed that genealogical effects shape the allele frequency spectrum, providing a biological explanation for the success of the ML model [81]. Similarly, GANs have been trained directly from real genetic data to generate realistic simulated datasets [82,83]. Follow-up studies demonstrated that GANs trained under neutral evolution can identify genomic regions that deviate from neutrality, consistent with selection, offering a novel and more interpretable framework for detecting non-neutral evolutionary processes [83].

Beyond inference of local genomic features, ML has also been applied to reconstruct population size histories, in some cases achieving performance comparable to well-established approaches such as Approximate Bayesian Computation (ABC) [84]. ABC has been integrated with deep learning to enable efficient comparison of complex demographic models, and for instance, has been used to study historical introgression in humans and provided statistical support for a previously unrecognized third archaic introgression contributing to the ancestry of Asian and Oceanian populations [85].

Decoding epistasis and complex trait evolution

The interaction between genetic variants, or epistasis, is central to understanding adaptation, robustness and constraints in evolution. However, epistasis detection has been hampered by the combinatorial explosion of possible variant interactions in genome-wide datasets. ML models, including random forests [86], gradient boosting [87], and deep neural networks [88], can capture nonlinear and higher-order interactions that are missed in classical regression-based methods. However, many models on epistasis detection only moderately improved the precision when additive heritability was low and dominance effects were high [89]. Thus, understanding epistasis will benefit from biologically informed and causally aware architectures that connect statistical signals to real molecular interaction networks, such as nonlinear or learned SNP embeddings [90].

Modern genomics has expanded beyond DNA sequences to include transcriptomics, proteomics, metabolomics, and chromatin landscapes, often at single-cell resolution. Deep learning models are now helping integrate diverse layers of biological information across scales, from single cells to populations and species. Applied to evolutionary questions, such integration may reveal how regulatory changes shape gene expression patterns, how networks are rewired across species, what is the genetic architecture of a complex trait [91], and how molecular variations contribute to phenotypic plasticity. For instance, deep learning applied to single-cell multi-omics data enables the reconstruction of developmental trajectories [92]. Future cell type atlases at the biodiversity level may reveal clues about how regulatory programs evolve and diverge and contribute to the complexity of cell type evolution [93].

At the population level, genomic and epigenomic data have been uncovering links between sequence evolution and adaptation to environmental stressors [94]. In the context of human population health, a recent autoencoder-based deep learning approach provides a method to impute missing phenotype data and improve the causal inference between genetic variation and disease phenotypes [95], thereby expanding the application of ML in studying complex traits. Importantly, this is an opportune time to investigate sex-biased diseases using

advanced ML tools for two key reasons. First, sex disparities have long existed in disease datasets [96], but recent efforts to generate more sex-balanced data have improved the power to detect genetic signatures associated with disease, enabling more comprehensive trait-level analyses. Second, from an evolutionary perspective, sex-biased constraints, expression variability, and selection signatures in human populations are generally weak [97–99]. ML approaches may therefore provide powerful means to uncover subtle or complex sex-specific genetic architectures that traditional methods might overlook.

Designing models that incorporate evolutionary theory

ML models have traditionally been agnostic to biological principles, learning patterns directly from data. Yet evolutionary biology provides a rich theoretical framework that can guide model design and interpretation. A successful example is AlphaFold [32], which leverages evolutionary information from multiple sequence alignments and co-variation patterns, along with structural templates, to predict protein backbones and folds accurately. Integrating phylogenetic information, population genetic parameters, or evolutionary constraints into ML architectures can make predictions both more accurate and more interpretable. For example, tools such as EVE (evolutionary model of variant effect) [100] use phylogenetic and population genetic information by training deep generative models on multiple sequence alignments across species and learning the evolutionary likelihood of each amino acid at each position. Such evolution-aware machine-learning frameworks improve performance and align computational methods with the central questions of evolutionary theory [66]. However, evolutionary biologists should be cautious when using these tools to address evolution-related questions, as the underlying models may introduce biases that affect evolutionary interpretation. For instance, because AlphaFold implicitly relies on residues conserved under purifying selection to stabilize its structural predictions, it may perform less accurately when analyzing sites evolving rapidly or under positive selection.

Frontiers, challenges, and outlook

ML in evolutionary biology has primarily focused on pattern recognition and prediction. A critical frontier is the development of models that move beyond correlation and capture causal mechanisms informed by evolutionary theory (see Outstanding questions). Emerging approaches in causal ML [101], such as causal graphs and counterfactual modeling, may offer ways to distinguish the effects of selection, drift, possibly demography, and test whether certain regulatory changes cause or merely correlate with phenotypic adaptation.

ML is positioned to detect subtle, high-dimensional patterns in complex evolutionary processes, including epistasis, pleiotropy, and gene–environment interactions, offering new insights into trait architecture and evolutionary dynamics. However, outputs are only as reliable as the underlying data and assumptions. Biases can arise from, but are not limited to, unbalanced training datasets (e.g., over-representation of model organisms and humans), incomplete annotations, or errors in reference genomes. Overfitting is another persistent concern, particularly when deep learning models capture spurious correlations that do not generalize beyond the training context and output misleading biological inferences [102]. These issues are especially pronounced when attempting to detect signatures of selection, adaptation, and introgression, as models are typically trained first on simulated data and

then applied to real genomes. Equally pressing is the issue of explainability. Many deep learning models function as black boxes, providing relatively accurate predictions but limited mechanistic insight into the hidden layers. Improving interpretability is essential for bridging predictive performance with biological understanding (see Outstanding questions).

Profound data imbalances further constrain ML applications. Despite explosive growth in genomic sequencing, knowledge remains uneven across clades, environments, and phenotypes. Model organisms - such as fruit flies, mice, yeast, and nematodes - and humans dominate high-quality datasets, while even their closely related species often lack essential population genomic, multi-omic, and phenotypic data. These disparities risk biasing ML inference and reinforcing blind spots in understanding adaptation and resilience. Addressing this requires global investment in underrepresented taxa, open-access resources, and models that generalize across sparse data or quantify uncertainty for poorly sampled lineages.

As ML becomes increasingly capable of interpreting and forecasting evolutionary trajectories, and of designing novel genetic variants and proteins, ethical guidelines must be established. Predicting pathogen evolution, for example, offers immense public health value but also raises dual-use concerns if misapplied. Therefore, we recommend that existing regulations and protocols regarding the dual use of harmful pathogens be adapted to include evolutionary considerations. Similarly, application of ML models to synthetic biology or *in silico* evolution creates unprecedented potential for the design of proteins [40,103], pathways, and even organisms [104] with unprecedented capabilities, underscoring the need for safety protocols that ensure minimal risk. Careful governance, transparent reporting, and interdisciplinary collaboration are essential to balance innovation with responsibility.

Training and deploying large ML models is computationally expensive, often relying heavily on GPU resources that incur substantial financial and environmental costs. The carbon footprint associated with large-scale model training has prompted increasing concern about whether the scientific gains always justify these expenditures. Moreover, access to high-performance computing infrastructure is unevenly distributed across institutions and countries, raising concerns and potentially exacerbating existing disparities in scientific capacity and participation. These limitations highlight the importance of considering more targeted ML approaches, or even classical statistical methods, when data are limited or questions are narrowly defined. A related and often underappreciated challenge is determining how much training data is sufficient: deep models trained on insufficient or unrepresentative data risk overfitting, poor generalization, and misleading inference, particularly under model misspecification or domain shift.

Despite these challenges, integrating artificial intelligence and evolutionary genetics remains in its early stages. ML is poised to become indispensable for unraveling evolutionary complexity and bridging empirical data with theory. Historically, many important conclusions have been drawn only after simplifying models or reducing data complexity to make analyses tractable. It is now possible to integrate heterogeneous data sources, learn complex nonlinear relationships, test long-standing theoretical predictions, and even uncover unanticipated dynamics. Current success includes protein structure prediction, enhancer activity inference, cell type classification, and selection signature

detection. However, ML has so far been unable to study collective protein-folding dynamics, the biological impact of individual mutations, the rewiring of regulatory networks across entire cell types, or the diversification of lineages and ecosystems over evolutionary time scales. It may also ultimately be impossible for ML to generate accurate predictions for some of these questions, suggesting the existence of as-yet undiscovered laws of nature. Regardless, whether we succeed in answering - or recognize our inability to answer - these deeper questions, the pursuit itself holds the potential to advance our understanding of medicine, agriculture, and conservation, enhance our ability to anticipate future evolutionary changes, and, hopefully, fundamentally reshape our understanding of genetics and evolution.

Acknowledgments

We regret that, due to space limitations, we were unable to cite all relevant studies in this rapidly expanding field. We hope this paper serves as a primer and a bridge to the broader primary literature. This work was supported by the National Institutes of Health (NIH) MIRA grant R35GM133780 to L.Z., the Stavros Niarchos Foundation (SNF) as part of its grant to the SNF Institute for Global Infectious Disease Research at The Rockefeller University to L.Z., and the Marlene Hess Center for Research on Women's Health and Biomedicine through its grant program at The Rockefeller University to N.S. and L.Z.. U.L. was supported by NSF postdoctoral fellowship 2410289.

Glossary

Additive heritability
The proportion of phenotypic variance attributable to the additive effects of genetic variants.

Ancestral recombination graph (ARG)
A genealogical structure that represents both the ancestry and recombination history of a sample of DNA sequences.

Artificial Intelligence (AI)
A broad field of computer science aimed at developing systems capable of perceiving, reasoning, and acting intelligently, often inspired by aspects of human cognition.

Autoencoder
A neural network trained to compress data into a lower-dimensional representation and then reconstruct it, useful for feature extraction and dimensionality reduction.

Convergent Evolution
Independent evolution of similar traits or structures in different lineages.

Convolutional Neural Network (CNN)
A type of deep learning architecture particularly effective at processing grid-like data (such as images or DNA sequences) by using convolutional layers to detect local patterns.

Deep Learning (DL)
A specialized subset of machine learning that uses multi-layered neural networks to automatically learn complex, hierarchical representations from large datasets.

Domain adaptation
An approach in ML that aims to maintain model performance when training data (simulated or empirical) differs from test data.

Dominance effects
Genetic effects that arise from interactions between alleles at the same locus.

Epistasis
The phenomenon where the effect of one genetic variant depends on the presence of other genetic variants, representing gene-gene interactions.

Fitness landscape
A conceptual map representing how different genetic variants affect organismal fitness, with peaks representing optimal genotypes and valleys representing fewer fit variants.

Gene-tree discordance
The phenomenon where evolutionary trees reconstructed from different genes show conflicting topologies.

Generative Adversarial Network (GAN)
A framework consisting of two neural networks (generator and discriminator) that compete with each other to generate realistic simulated data.

Hard sweep
A selective sweep that occurs when a new beneficial mutation arises and rapidly increases in frequency in a population.

Hidden Markov Model (HMM)
A statistical model that represents systems transitioning between hidden states, commonly used in sequence analysis and gene prediction.

Incomplete lineage sorting
A phenomenon where ancestral genetic variation persists through speciation events, leading to gene trees that differ from the species tree.

Introgression
The transfer of genetic material from one species or population to another through hybridization and backcrossing.

Machine Learning (ML)
A subfield of AI focused on algorithms that learn patterns or predictive relationships directly from data without being explicitly programmed.

Phylogenetics
The study of evolutionary relationships among organisms or genes, typically represented as evolutionary trees.

Positive selection
Natural selection that favors beneficial mutations, causing them to increase in frequency.

Purifying selection

Natural selection that removes deleterious mutations from a population, leading to sequence conservation.

Recurrent Neural Network (RNN)
A neural network architecture designed to process sequential data by maintaining memory of previous inputs, commonly used for time-series or sequence analysis.

Selective sweep
A pattern of reduced genetic diversity caused by positive selection rapidly increasing the frequency of an advantageous allele and linked neutral variants.

Soft sweep
A selective sweep that occurs when selection acts on standing genetic variation (pre-existing alleles) or recurrent mutations.

Supervised Learning
ML paradigm in which models are trained on labeled data (input–output pairs).

Transformer
A neural network architecture that uses attention mechanisms to process sequential data, enabling parallel processing and long-range dependencies.

Unsupervised Learning
ML paradigm in which models infer patterns from unlabeled data, often through clustering or dimensionality reduction.

Reference

1. Haldane JBS (1932) The Causes of Evolution, Longmans, Green & Co
2. Wright S (1931) Evolution in Mendelian Populations. Genetics 16, 97 [PubMed: 17246615]
3. Fisher RA (1930) The genetical theory of natural selection, Clarendon Press
4. Avery OT et al. (1944) Studies on the chemical nature of the substance inducing transformation of pneumococcal types: Induction of transformation by a desoxyribonucleic acid fraction isolated from pneumococcus type iii. Journal of Experimental Medicine 79, 137–158 [PubMed: 19871359]
5. Watson JD and Crick FHC (1953) Molecular structure of nucleic acids; a structure for deoxyribose nucleic acid. Nature 171, 737–738 [PubMed: 13054692]
6. Smith TF and Waterman MS (1981) Identification of common molecular subsequences. J. Mol. Biol 147, 195–197 [PubMed: 7265238]
7. Wu R (1970) Nucleotide sequence analysis of DNA. I. Partial sequence of the cohesive ends of bacteriophage λ and 186 DNA. J. Mol. Biol 51, 501–521 [PubMed: 4321727]
8. Wu R and Taylor E (1971) Nucleotide sequence analysis of DNA: II. Complete nucleotide sequence of the cohesive ends of bacteriophage λ DNA. J. Mol. Biol 57, 491–511 [PubMed: 4931680]
9. International Human Genome Sequencing Consortium et al. (2001) Initial sequencing and analysis of the human genome. Nature 409, 860–921 [PubMed: 11237011]
10. Adams MD et al. (2000) The genome sequence of Drosophila melanogaster. Science 287, 2185–95 [PubMed: 10731132]
11. Drosophila 12 Genomes Consortium et al. (2007) Evolution of genes and genomes on the Drosophila phylogeny. Nature 450, 203–18 [PubMed: 17994087]
12. Gibbs RA et al. (2015) A global reference for human genetic variation. Nature 526, 68–74 [PubMed: 26432245]

13. Good PJ et al. (2004) The ENCODE (ENCyclopedia Of DNA Elements) Project. Science 306, 636–40 [PubMed: 15499007]
14. Burke MK et al. (2010) Genome-wide analysis of a long-term evolution experiment with Drosophila. Nature 467, 587–590 [PubMed: 20844486]
15. Turing AM (1950) Computing Machinery and Intelligence. Mind 49, 433–460
16. Rumelhart DE et al. (1986) Learning representations by back-propagating errors. Nature 323, 533–536
17. Hinton GE and Nowlan SJ (1987) How Learning Can Guide Evolution. Complex Systems 1, 495–502
18. Hopfield JJ (1978) Origin of the genetic code: a testable hypothesis based on tRNA structure, sequence, and kinetic proofreading. Proceedings of the National Academy of Sciences 75, 4334–4338
19. Bout DE Van den and Miller TK. (1989) Improving the performance of the Hopfield-Tank neural network through normalization and annealing. Biol. Cybern 62, 129–139
20. Shi S et al. (2017) Benchmarking state-of-the-art deep learning software tools. Proceedings - 2016 7th International Conference on Cloud Computing and Big Data, CCBD 2016 DOI: 10.1109/CCBD.2016.029
21. LeCun Y et al. (1998) Gradient-based learning applied to document recognition. Proceedings of the IEEE 86, 2278–2323
22. Vaswani A et al. (2017) Attention is all you need. in Advances in Neural Information Processing Systems, 2017-Decem, pp. 5999–6009
23. Bahdanau D et al. (2014) Neural Machine Translation by Jointly Learning to Align and Translate. 3rd International Conference on Learning Representations, ICLR 2015 - Conference Track Proceedings at <https://arxiv.org/pdf/1409.0473>
24. He K et al. (2016) Deep residual learning for image recognition. Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition 2016-December, 770–778
25. Zou J et al. (2019) A primer on deep learning in genomics. Nat. Genet 51, 12–18 [PubMed: 30478442]
26. Alipanahi B et al. (2015) Predicting the sequence specificities of DNA- and RNA-binding proteins by deep learning. Nat. Biotechnol 33, 831–838 [PubMed: 26213851]
27. Zhou J and Troyanskaya OG (2015) Predicting effects of noncoding variants with deep learning-based sequence model. Nat. Methods 12, 931–934 [PubMed: 26301843]
28. Fukushima K (1980) Neocognitron: a self organizing neural network model for a mechanism of pattern recognition unaffected by shift in position. Biol. Cybern 36, 193–202 [PubMed: 7370364]
29. Hopfield JJ (1982) Neural networks and physical systems with emergent collective computational abilities. Proc. Natl. Acad. Sci. U. S. A 79, 2554 [PubMed: 6953413]
30. Bengio Y et al. (2003) A Neural Probabilistic Language Model. Journal of Machine Learning Research 3, 1137–1155
31. Wagner A (1996) DOES EVOLUTIONARY PLASTICITY EVOLVE? Evolution 50, 1008–1023 [PubMed: 28565284]
32. Jumper J et al. (2021) Highly accurate protein structure prediction with AlphaFold. Nature 596, 583–589 [PubMed: 34265844]
33. Abramson J et al. (2024) Accurate structure prediction of biomolecular interactions with AlphaFold 3. Nature 2024 630:8016 630, 493–500 [PubMed: 38718835]
34. Peng J and Zhao L (2024) The origin and structural evolution of de novo genes in Drosophila. Nat. Commun 15, 810 [PubMed: 38280868]
35. Peng J et al. (2024) The Origin and Evolution of Sex Peptide and Sex Peptide Receptor Interactions. Mol. Biol. Evol 41, msae065 [PubMed: 38518286]
36. Pillai AS et al. (2020) Origin of complexity in haemoglobin evolution. Nature 581, 480–485 [PubMed: 32461643]
37. Baek M et al. (2021) Accurate prediction of protein structures and interactions using a three-track neural network. Science (1979). 373, 871–876

38. Rives A et al. (2021) Biological structure and function emerge from scaling unsupervised learning to 250 million protein sequences. Proceedings of the National Academy of Sciences 118, e2016239118
39. Lin Z et al. (2023) Evolutionary-scale prediction of atomic-level protein structure with a language model. Science (1979). 379, 1123–1130
40. Watson JL et al. (2023) De novo design of protein structure and function with RFdiffusion. Nature 620, 1089–1100 [PubMed: 37433327]
41. Peng J et al. (2025) Gene regulatory networks and essential transcription factors for de novo-originated genes. Nat. Ecol. Evol 9, 1487–1498 [PubMed: 40659874]
42. Cao Z et al. (2025) Language models reveal a complex sequence basis for adaptive convergent evolution of protein functions. Proceedings of the National Academy of Sciences 122, e2418254122
43. Peng J and Zhao L (2026) A predicted structural interactome reveals binding interference from intrinsically disordered regions. PLoS Comput. Biol 22, e1013899 [PubMed: 41570036]
44. Eraslan G et al. (2019) Deep learning: new computational modelling techniques for genomics. Nat. Rev. Genet 20, 389–403 [PubMed: 30971806]
45. Kelley DR et al. (2016) Basset: Learning the regulatory code of the accessible genome with deep convolutional neural networks. Genome Res. 26, 990–999 [PubMed: 27197224]
46. Khodursky S et al. (2023) The evolution and mutational robustness of chromatin accessibility in Drosophila. Genome Biol. 24, 232 [PubMed: 37845780]
47. de Almeida BP et al. (2024) Targeted design of synthetic enhancers for selected tissues in the Drosophila embryo. Nature 626, 207–211 [PubMed: 38086418]
48. Gao VR et al. (2024) ChromaFold predicts the 3D contact map from single-cell chromatin accessibility. Nat. Commun 15, 9432 [PubMed: 39487131]
49. Pampari A et al. (2025) ChromBPNet: bias factorized, base-resolution deep learning models of chromatin accessibility reveal cis-regulatory sequence syntax, transcription factor footprints and regulatory variants. bioRxiv DOI: 10.1101/2024.12.25.630221
50. Avsec Ž et al. (2021) Base-resolution models of transcription-factor binding reveal soft motif syntax. Nat. Genet 53, 354–366 [PubMed: 33603233]
51. Ding X et al. (2019) Deciphering protein evolution and fitness landscapes with latent space models. Nat. Commun 10, 5644 [PubMed: 31822668]
52. Freschlin CR et al. (2024) Neural network extrapolation to distant regions of the protein fitness landscape. Nat. Commun 15, 6405 [PubMed: 39080282]
53. Hsu C et al. (2022) Learning protein fitness models from evolutionary and assay-labeled data. Nat. Biotechnol 40, 1114–1122 [PubMed: 35039677]
54. Hopf TA et al. (2017) Mutation effects predicted from sequence co-variation. Nat. Biotechnol 35, 128–135 [PubMed: 28092658]
55. Shi W et al. (2025) Influenza vaccine strain selection with an AI-based evolutionary and antigenicity model. Nat. Med DOI: 10.1038/S41591-025-03917-Y
56. Marcus JL et al. (2020) Artificial Intelligence and Machine Learning for HIV Prevention: Emerging Approaches to Ending the Epidemic. Curr. HIV/AIDS Rep 17, 171 [PubMed: 32347446]
57. Thadani NN et al. (2023) Learning from prepandemic data to forecast viral escape. Nature 622, 818–825 [PubMed: 37821700]
58. Suvorov A et al. (2019) Accurate Inference of Tree Topologies from Multiple Sequence Alignments Using Deep Learning. Syst. Biol 69, 221
59. Smith ML and Hahn MW (2023) Phylogenetic inference using generative adversarial networks. Bioinformatics 39, btad543 [PubMed: 37669126]
60. Zou Z et al. (2020) Deep Residual Neural Networks Resolve Quartet Molecular Phylogenies. Mol. Biol. Evol 37, 1495–1507 [PubMed: 31868908]
61. Mo YK et al. (2024) Applications of machine learning in phylogenetics. Mol. Phylogenet. Evol 196, 108066 [PubMed: 38565358]

62. Feng S et al. (2022) Incomplete lineage sorting and phenotypic evolution in marsupials. Cell 185, 1646–1660.e18 [PubMed: 35447073]
63. Morales-Briones DF et al. (2018) Phylogenomic analyses reveal a deep history of hybridization and polyploidy in the Neotropical genus Lachemilla (Rosaceae). New Phytologist 218, 1668–1684 [PubMed: 29604235]
64. Thompson A et al. (2024) Deep Learning and Likelihood Approaches for Viral Phylogeography Converge on the Same Answers Whether the Inference Model Is Right or Wrong. Syst. Biol 73, 183–206 [PubMed: 38189575]
65. Smith ML and Hahn MW (2023) Phylogenetic inference using generative adversarial networks. Bioinformatics 39, btad543 [PubMed: 37669126]
66. Ye C et al. (2025) Predicting functional constraints across evolutionary timescales with phylogeny-informed genomic language models. bioRxiv DOI: 10.1101/2025.09.21.677619
67. Schrider DR and Kern AD (2018) Supervised Machine Learning for Population Genetics: A New Paradigm. Trends in Genetics 34, 301–312 [PubMed: 29331490]
68. Sheehan S and Song YS (2016) Deep Learning for Population Genetic Inference. PLoS Comput. Biol 12, e1004845 [PubMed: 27018908]
69. Schrider DR and Kern AD (2016) S/HIC: Robust Identification of Soft and Hard Sweeps Using Machine Learning. PLoS Genet. 12, e1005928 [PubMed: 26977894]
70. Pavlidis P et al. (2010) Searching for Footprints of Positive Selection in Whole-Genome SNP Data From Nonequilibrium Populations. Genetics 185, 907 [PubMed: 20407129]
71. Schrider DR and Kern AD (2017) Soft Sweeps Are the Dominant Mode of Adaptation in the Human Genome. Mol. Biol. Evol 34, 1863–1877 [PubMed: 28482049]
72. Ronen R et al. (2013) Learning natural selection from the site frequency spectrum. Genetics 195, 181–193 [PubMed: 23770700]
73. Flagel L et al. (2019) The Unreasonable Effectiveness of Convolutional Neural Networks in Population Genetic Inference. Mol. Biol. Evol 36, 220–238 [PubMed: 30517664]
74. Mo Z and Siepel A (2023) Domain-adaptive neural networks improve supervised machine learning based on simulated population genetic data. PLoS Genet. 19, e1011032 [PubMed: 37934781]
75. Cobb KA and Smith ML (2025) The reasonable effectiveness of domain adaptation for inference of introgression. bioRxiv DOI: 10.1101/2025.01.17.633659
76. Korfmann K et al. (2024) Simultaneous Inference of Past Demography and Selection from the Ancestral Recombination Graph under the Beta Coalescent. Peer Community Journal 4, e33
77. Whitehouse LS et al. (2024) Tree Sequences as a General-Purpose Tool for Population Genetic Inference. Mol. Biol. Evol 41, msae223 [PubMed: 39460991]
78. Hejase HA et al. (2022) A Deep-Learning Approach for Inference of Selective Sweeps from the Ancestral Recombination Graph. Mol. Biol. Evol 39, msab332 [PubMed: 34888675]
79. Korfmann K et al. (2023) Deep Learning in Population Genetics. Genome Biol. Evol 15, evad008 [PubMed: 36683406]
80. Adrion JR et al. (2020) Predicting the Landscape of Recombination Using Deep Learning. Mol. Biol. Evol 37, 1790–1808 [PubMed: 32077950]
81. Hahn MW and Mishra SR (2026) Estimating recombination using only the allele frequency spectrum. Genetics 232, iyaf108 [PubMed: 40472340]
82. Booker WW et al. (2023) This population does not exist: learning the distribution of evolutionary histories with generative adversarial networks. Genetics 224, iyad063 [PubMed: 37067864]
83. Riley R et al. (2024) Interpreting generative adversarial networks to infer natural selection from genetic data. Genetics 226, iyae024 [PubMed: 38386895]
84. Sanchez T et al. (2021) Deep learning for population size history inference: Design, comparison and combination with approximate Bayesian computation. Mol. Ecol. Resour 21, 2645–2660 [PubMed: 32644216]
85. Mondal M et al. (2019) Approximate Bayesian computation with deep learning supports a third archaic introgression in Asia and Oceania. Nature Communications 2019 10:1 10, 246-
86. Jiang R et al. (2009) A random forest approach to the detection of epistatic interactions in case-control studies. BMC Bioinformatics 10, S65 [PubMed: 19208169]

87. Verma SS et al. (2018) Collective feature selection to identify crucial epistatic variants. BioData Min. 11, 5 [PubMed: 29713383]

88. Abdollahi-Arpanahi R et al. (2020) Deep learning versus parametric and ensemble methods for genomic prediction of complex phenotypes. Genet. Sel. Evol 52, 12 [PubMed: 32093611]

89. Bellot P et al. (2018) Can Deep Learning Improve Genomic Prediction of Complex Human Traits? Genetics 210, 809–819 [PubMed: 30171033]

90. van Hilten A et al. (2025) Detecting genetic interactions with visible neural networks. Commun. Biol 8, 874 [PubMed: 40473911]

91. van der Bijl W et al. (2025) Deep learning reveals the complex genetic architecture of male guppy colouration. Nat. Ecol. Evol 9, 1614–1625 [PubMed: 40596731]

92. Erfanian N et al. (2023) Deep learning applications in single-cell genomics and transcriptomics data analysis. Biomedicine & Pharmacotherapy 165, 115077 [PubMed: 37393865]

93. Sebé-Pedrós A et al. (2025) The Biodiversity Cell Atlas: mapping the tree of life at cellular resolution. Nature 645, 877–885 [PubMed: 40993253]

94. Argentieri MA et al. (2025) Integrating the environmental and genetic architectures of aging and mortality. Nat. Med 31, 1016–1025 [PubMed: 39972219]

95. An U et al. (2023) Deep learning-based phenotype imputation on population-scale biobank data increases genetic discoveries. Nat. Genet 55, 2269–2276 [PubMed: 37985819]

96. Straw I and Wu H (2022) Investigating for bias in healthcare algorithms: a sex-stratified analysis of supervised machine learning models in liver disease prediction. BMJ Health Care Inform. 29, e100457

97. Khodursky S et al. (2022) Sex differences in inter-individual gene expression variability across human tissues. PNAS Nexus 1, pgac243 [PubMed: 36712323]

98. Fair BJ et al. (2020) Gene expression variability in human and chimpanzee populations share common determinants. Elife 9, e59929 [PubMed: 33084571]

99. Glassberg EC et al. (2019) Evidence for Weak Selective Constraint on Human Gene Expression. Genetics 211, 757–772 [PubMed: 30554168]

100. Frazer J et al. (2021) Disease variant prediction with deep generative models of evolutionary data. Nature 599, 91–95 [PubMed: 34707284]

101. Prosperi M et al. (2020) Causal inference and counterfactual prediction in machine learning for actionable healthcare. Nat. Mach. Intell 2, 369–375

102. Bernett J et al. (2024) Guiding questions to avoid data leakage in biological machine learning applications. Nature Methods 2024 21:8 21, 1444–1453 [PubMed: 39122953]

103. Alley EC et al. (2019) Unified rational protein engineering with sequence-based deep representation learning. Nat. Methods 16, 1315–1322 [PubMed: 31636460]

104. King SH et al. (2025) Generative design of novel bacteriophages with genome language models. bioRxiv DOI: 10.1101/2025.09.12.675911

Outstanding Questions

- How can ML models be made more interpretable, ensuring that predictions contribute to mechanistic understanding rather than remaining black boxes?
- How can structural and functional prediction models be improved to capture evolutionary changes, particularly rapid or adaptive evolution and polymorphisms within species?
- Can causal ML frameworks disentangle the relative contributions of selection, drift, and demography in shaping genomic patterns?
- What strategies can mitigate biases introduced by uneven genomic sampling across taxa?
- How can ML models reliably predict evolutionary signatures in species under-sampled or with limited high-quality training data?
- Is it possible to predict the evolution of a trait or a species using large-scale models, and if so, what trait datasets do these models need?
- What ethical guidelines are needed to guide the use of ML in studying evolutionary trajectories, particularly for predicting pathogen evolution and the use of synthetic biology?

Highlights

- Recent advances in ML especially deep learning have enabled breakthroughs in biology and are poised to play an important role in evolutionary biology.
- ML models are increasingly capable of connecting genetic variation to molecular function, integrating structural, regulatory, and phenotypic data at multiple scales.
- ML models can now be used directly on raw SNPs, haplotypes, or allele frequency spectra, reducing reliance on summary statistics and enabling the discovery of previously unrecognized evolutionary patterns.
- Emerging approaches move beyond correlation, using causal inference to generate mechanistic insights into evolutionary processes.
- Integrating multi-omics data through ML is revealing hidden layers of complexity, such as epistasis, regulatory evolution, and phenotypic plasticity, that shape adaptation and diversification.

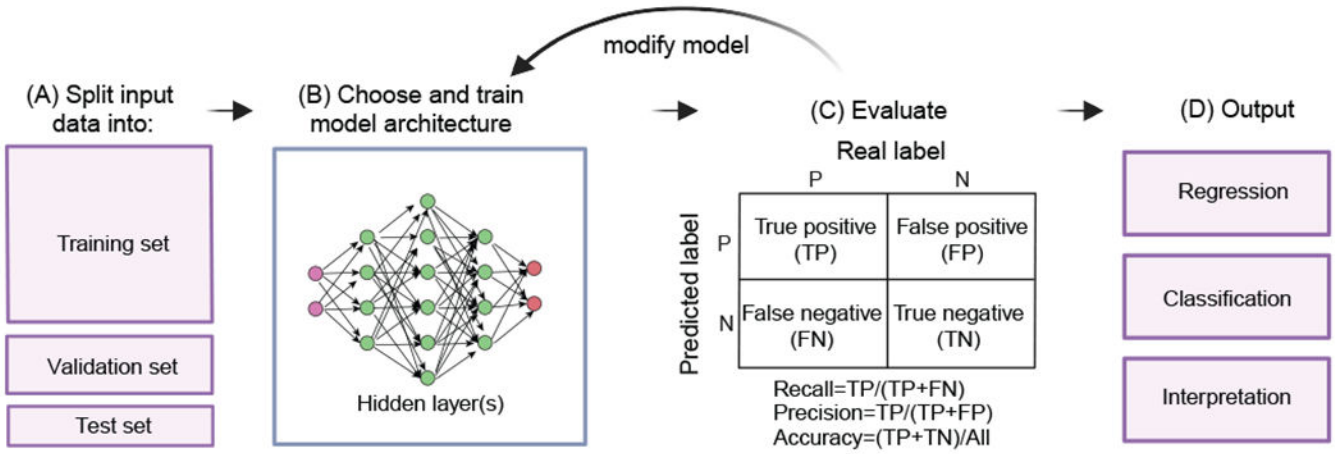

Figure 1.

Overview of a deep learning workflow for genomic sequence prediction. (A) The dataset is divided into training, validation, and test sets. Positive and negative examples need to be balanced for potential confounders to ensure that the model learns biologically relevant features rather than artifacts. Most data are usually used for training rather than for validation or testing. (B) The model architecture is selected based on knowledge. (C) Model performance is evaluated using true positive (TP), false positive (FP), false negative (FN), and true negative (TN) rates. Precision and recall are often used to evaluate model performance. Based on the results, the model architecture and hyperparameters often need to be tweaked. (D) Model interpretation is performed, and the outputs can be used for classification or quantitative prediction tasks.

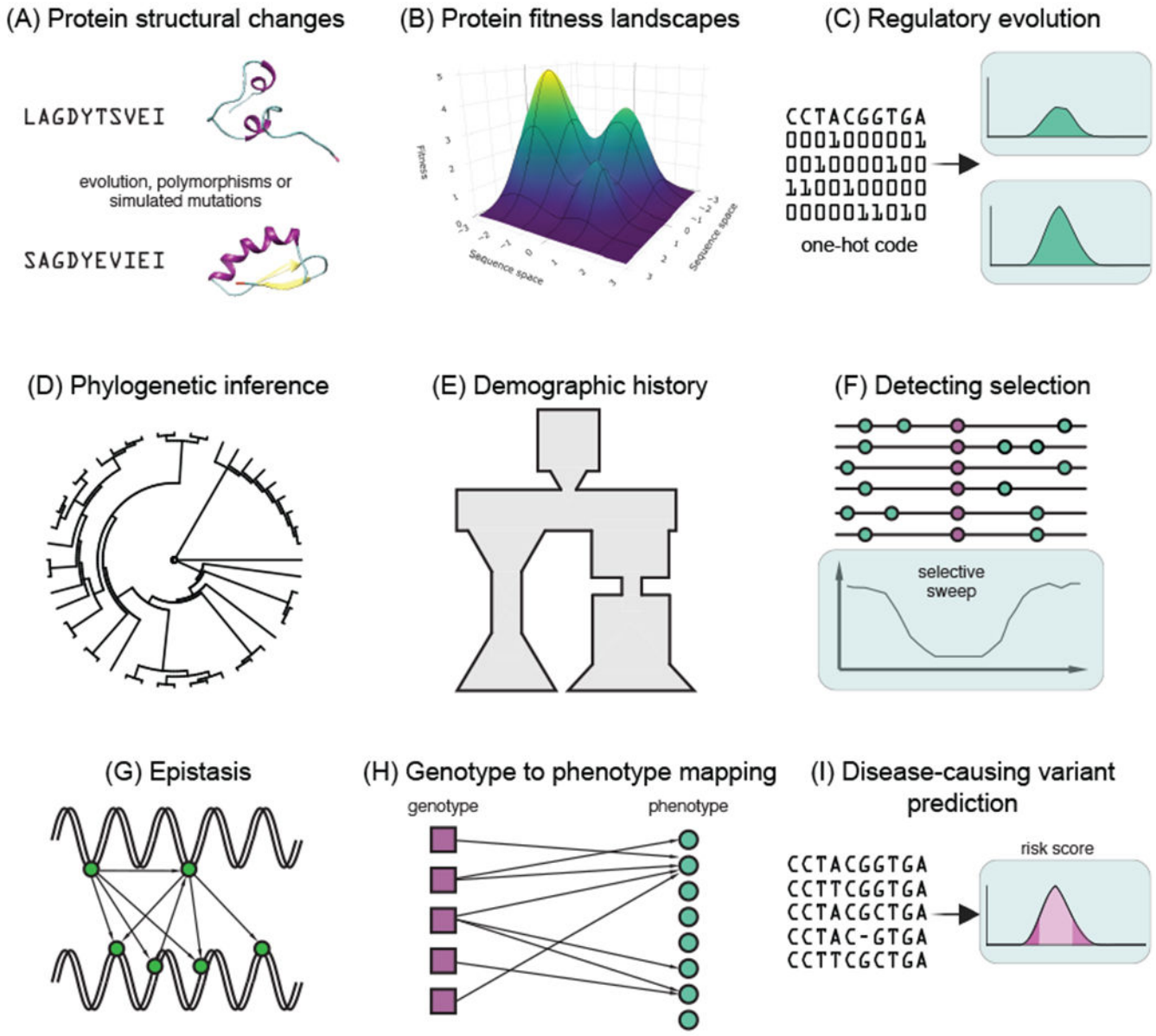

Figure 2.

Applications of machine learning in evolutionary biology. (A) Protein structure prediction and analysis of structural changes in the context of evolution, population polymorphism, or simulated mutations. (B) Protein fitness landscapes that reveal evolutionary principles and provide practical insights for virology and immunology. (C) Regulatory architecture and its evolution. (D) Phylogenetic inference. (E) Demographic history reconstruction. (F) Detection of selection and classification of selection types. (G) Understanding epistasis and the genetic basis of complex traits. (H) Mapping from genotypes to phenotypes, and (I) predicting disease-causing variants. All panels are illustrative examples and do not refer to specific studies.

Table 1

Relationships among AI, machine learning, and deep learning in biology.

Term	Definition and scope	Representative methods / models	Examples of applications in biology
Artificial Intelligence (AI)	Broad field aiming to create systems that can perceive, reason, and act intelligently.	Knowledge-based systems, logical reasoning, search algorithms.	clinical decision support systems, laboratory robotics for experiment planning
Machine Learning (ML)	Subfield of AI focused on algorithms that learn from data to improve performance on a task without explicit programming.	**Supervised learning:** Linear Regression, Logistic Regression, Support Vector Machine (SVM), Decision Trees.	Predicting gene expression from genomic features, classifying cell types from single-cell RNA-seq, Genome-wide association studies (GWAS) trait prediction, protein–protein interaction prediction, phenotype classification from imaging data.
		Unsupervised learning: Clustering, Principal Component Analysis (PCA).	
		Probabilistic models: Bayesian networks, Hidden Markov Models (HMMs).	
Deep Learning (DL)	Subfield of ML using multi-layered neural networks to automatically learn large-scale hierarchical data representations.	**Neural architectures:** Convolutional Neural Network (CNN), Recurrent Neural Network (RNN) / Long Short-Term Memory (LSTM), Transformer, Autoencoder, Generative Adversarial Network (GAN).	Protein structure prediction (such as AlphaFold), enhancer and promoter prediction from sequence data, tissue and pathology image recognition, gene regulatory network inference, integration of multi-omics data