

# Transformer-Based Deep Learning for Protein Structure and Function Prediction: From Sequence Understanding to Biological Insight: A Review



Simon J. Moore <sup>1\*</sup>

## Abstract

There is, perhaps, something quietly transformative happening in how we understand proteins. For decades, the field relied on a combination of experimental precision and evolutionary inference—methods that were undeniably powerful, yet often limited by scale, cost, and the boundaries of known biology. What has changed, more recently, is not just the volume of data, but the way we interpret it. This review explores the emergence of Transformer-based deep learning models as a turning point in protein science, where sequences are no longer treated merely as biochemical strings, but as a form of language—structured, contextual, and, to some extent, interpretable. At the center of this shift lies the idea that long-range dependencies—once difficult to capture—can now be modeled directly through attention mechanisms. These models appear capable of extracting structural and functional signals from raw sequences alone, sometimes without explicit evolutionary guidance. And yet, their success raises questions that feel as important as the answers they provide: what exactly are these systems learning, and how reliably can we trust their predictions? By tracing the evolution from alignment-based methods to large-scale representation learning, this review attempts to situate Transformer models within a broader computational narrative. It suggests that we are moving—perhaps cautiously—toward a framework where

biological complexity can be read, predicted, and even designed with increasing fluency. **Keywords:** Transformer models; Protein structure prediction; Protein language models; Bioinformatics; Deep learning

## Keywords

Transformer models; Protein structure prediction; Protein language models; Bioinformatics; Deep learning

## 1. Introduction

Proteins, in many respects, sit at the very center of biological organization. They do not merely participate in life processes—they enable them, orchestrating everything from enzymatic catalysis and intracellular transport to immune surveillance and signal transduction. Yet, despite this centrality, understanding how proteins function has never been straightforward. It hinges on a deceptively simple premise: that a protein's function emerges from its three-dimensional structure, which in turn is encoded—somehow—within its linear amino acid sequence (Anfinsen, 1973). This sequence–structure–function paradigm has guided decades of biochemical inquiry, but it has also, perhaps quietly, exposed the limits of our predictive capacity.

There is, for instance, the well-known paradox articulated

**Significance |** Reveals how Transformer models redefine protein science by enabling sequence-driven prediction, bridging biological complexity, and accelerating computational discovery.

## \*Correspondence.

Department of Life Sciences, Imperial College London, London, SW7 2AZ, UK. E-mail: [simon.moore@imperial.ac.uk](mailto:simon.moore@imperial.ac.uk),

Editor Vishwanadham Mandala, Ph.D. And accepted by the Editorial Board Jan 21, 2021 (received for review Nov 20, 2020)

## Author Affiliation:

<sup>1</sup> Department of Life Sciences, Imperial College London, London, SW7 2AZ, UK.

## Please cite this article as:

Moore, S. J. (2021). "Transformer-Based Deep Learning for Protein Structure and Function Prediction: From Sequence Understanding to Biological Insight: A Review", *Bioinfo Chem*, 3(1), 1-12, 10734

© 2021 Bioinfo Chem, a publication of Eman Research, USA.  
This is an open access article under the CC BY-NC-ND license.  
(<http://creativecommons.org/licenses/by-nc-nd/4.0/>)  
(<https://publishing.emanresearch.org>)

by Levinthal (1968). If a protein were to sample all possible conformations randomly, the time required to reach its native fold would exceed the age of the universe. And yet, in living systems, proteins fold reliably and rapidly—often within milliseconds. This apparent contradiction suggests that folding is not random but constrained, guided by intrinsic physicochemical principles embedded within the sequence itself. Still, translating that insight into predictive models has proven far more difficult than the theory might initially suggest.

Historically, structural biology has relied on experimental techniques to resolve this challenge. Methods such as X-ray crystallography, nuclear magnetic resonance (NMR) spectroscopy, and, more recently, cryo-electron microscopy have provided remarkably detailed views of protein structures (Berman et al., 2000). These approaches are, without question, powerful. However, they are also resource-intensive, technically demanding, and often slow. As biological data generation accelerated—particularly with the rise of high-throughput sequencing—the imbalance between known protein sequences and experimentally determined structures became increasingly pronounced. Databases such as UniProt expanded at an extraordinary pace, while structural repositories like the Protein Data Bank (PDB) grew more modestly (UniProt Consortium, 2015; Berman et al., 2000). The resulting “sequence–structure gap” is not merely a quantitative issue; it represents a fundamental bottleneck in translating genomic information into biological understanding.

To address this gap, computational methods emerged as an essential complement to experimental approaches. Early bioinformatics tools relied heavily on sequence alignment and homology-based inference. Algorithms such as BLAST (Altschul et al., 1997) and profile hidden Markov models implemented in HMMER (Eddy, 1998; Finn et al., 2011) leveraged evolutionary conservation to infer structural and functional properties. These methods were, and remain, highly effective when homologous sequences are available. However, their limitations become apparent in less well-characterized regions of sequence space—particularly for so-called “orphan proteins” that lack detectable homologs, or for structurally flexible regions where alignment signals are weak.

At the same time, the scale of biological data introduced new computational challenges. Constructing multiple sequence alignments (MSAs) for large datasets became

increasingly expensive, both in time and computational resources. More subtly, reliance on evolutionary information may obscure cases where novel or engineered proteins deviate from established patterns. In such contexts, traditional methods begin to feel not inadequate, exactly, but incomplete.

The emergence of deep learning introduced a new conceptual framework for tackling these problems. Neural network architectures, particularly convolutional neural networks (CNNs) and recurrent neural networks (RNNs), were adapted to model biological sequences. Among these, long short-term memory (LSTM) networks offered a mechanism for capturing sequential dependencies (Hochreiter & Schmidhuber, 1997). Yet, despite their promise, these models encountered persistent limitations. Sequential processing constrained parallelization, and long-range dependencies—so critical in protein folding, where distant residues interact spatially—remained difficult to capture effectively. The vanishing gradient problem further complicated the learning of long-distance relationships, often diminishing the influence of earlier residues as sequences grew longer.

A more decisive shift began with the introduction of the Transformer architecture (Vaswani et al., 2017). Originally developed for natural language processing, Transformers departed from recurrence entirely, instead relying on self-attention mechanisms. This seemingly technical change had profound implications. By allowing every element in a sequence to attend to every other element simultaneously, Transformers enabled the modeling of long-range interactions with unprecedented efficiency. In the context of proteins, this is particularly significant: residues that are far apart in sequence may be proximal in three-dimensional space, and capturing such relationships is essential for accurate structural prediction.

The analogy between protein sequences and language—once a somewhat speculative idea—gained renewed traction in this context. Amino acids can be treated as tokens, sequences as sentences, and evolutionary patterns as a kind of grammar. Building on this perspective, large-scale pre-training strategies were adapted from natural language processing. Models such as BERT, based on masked language modeling (Devlin et al., 2018), and generative frameworks inspired by autoregressive training (Radford et al., 2018), were repurposed to learn representations from vast protein sequence datasets. These protein language models

(PLMs) do not rely explicitly on alignments; instead, they infer statistical regularities directly from raw sequences.

What is perhaps most striking is how much these models appear to capture implicitly. Without being explicitly trained on structural labels, they can encode information about secondary structure motifs, residue contacts, and even functional sites. In some cases, their performance rivals or exceeds traditional methods, particularly in low-homology regimes. And yet, this success raises its own questions. If these models are learning biologically meaningful representations, what exactly are they capturing? And how can those representations be interpreted?

There remains, in other words, a certain opacity—a “black-box” quality—that complicates the integration of deep learning outputs into biological reasoning. Attention maps, for example, can sometimes highlight structurally relevant interactions, but their interpretation is not always straightforward. Moreover, the computational demands of large Transformer models are substantial, often requiring specialized hardware and extensive training data. These practical constraints cannot be ignored, particularly as the field moves toward broader accessibility and application.

Despite these challenges, the trajectory is difficult to overlook. Transformer-based models have not simply improved prediction accuracy; they have reshaped how researchers conceptualize protein analysis. The focus is shifting—from explicit alignment and handcrafted features toward representation learning, from deterministic pipelines toward probabilistic inference, and, perhaps most importantly, from isolated tasks toward integrative frameworks that link sequence, structure, and function.

This review, therefore, aims to synthesize these developments with a degree of cautious reflection. It examines the evolution of computational approaches in protein science, with particular emphasis on the transition to attention-based architectures. It also considers the extent to which these models can generalize beyond well-characterized proteins, offering insights into novel sequences and engineered systems. Finally, it explores the interpretability of learned representations—an area that remains, in many ways, unresolved but increasingly central to the field.

In doing so, the review does not assume that current models represent a final solution. Rather, it treats them as part of an ongoing shift—one that is still unfolding, and whose implications, both practical and conceptual, are

only beginning to be understood.

## 2. Methodology

### 2.1 Conceptual Design and Review Approach

This study adopts a narrative review methodology, designed to synthesize the conceptual and technological evolution of computational protein science, with particular emphasis on Transformer-based deep learning models. Unlike systematic reviews that rely on rigid inclusion criteria and statistical aggregation, the narrative approach allows for a more interpretive and integrative analysis of interdisciplinary developments. This was especially important given the convergence of bioinformatics, structural biology, and artificial intelligence, where insights often emerge not from isolated studies, but from the interaction between computational frameworks and biological data. The scope of the review was intentionally structured to trace a progression—from classical sequence alignment methods to modern representation learning approaches—while maintaining coherence across algorithmic, biological, and evaluative dimensions. Foundational literature was prioritized to ensure that the analysis reflects the underlying principles shaping the field rather than only its most recent outputs.

### 2.2 Literature Identification and Selection Strategy

Relevant literature was identified through a targeted selection of highly cited and methodologically influential studies in bioinformatics and machine learning. Priority was given to works that introduced or significantly advanced key computational paradigms, including sequence alignment algorithms, probabilistic modeling techniques, neural network architectures, and Transformer-based frameworks. Seminal studies such as the development of BLAST and PSI-BLAST were included to represent early homology-based inference (Altschul et al., 1990; Altschul et al., 1997), alongside Hidden Markov Model approaches implemented in HMMER for protein domain identification (Eddy, 1998; Finn et al., 2011). The transition toward machine learning was captured through foundational neural architectures, including recurrent neural networks and long short-term memory models (Elman, 1990; Hochreiter & Schmidhuber, 1997), as well as convolutional networks for motif detection (Fukushima, 1980).

More recent developments were incorporated through studies on representation learning and attention-based

models, particularly the Transformer architecture (Vaswani et al., 2017), and its adaptation into pre-training frameworks such as BERT and generative models (Devlin et al., 2018; Radford et al., 2018). These selections ensured continuity between historical methods and emerging paradigms.

### 2.3 Integration of Biological Data Sources

To contextualize computational developments within biological reality, the review incorporates major curated databases that have served as foundational resources for model training and validation. Structural datasets from the Protein Data Bank (PDB) were used to represent experimentally validated protein conformations (Berman et al., 2000), while large-scale sequence repositories such as UniProt and UniRef provided the extensive unlabeled data required for deep learning pre-training (UniProt Consortium, 2015; Suzek et al., 2015). In addition, domain-specific resources such as Pfam were considered for their role in defining protein families and conserved functional regions (Finn et al., 2013). These datasets were not treated merely as background resources, but as active components in shaping the capabilities of computational models, particularly in the context of large-scale representation learning.

### 2.4 Analytical Framework and Synthesis Strategy

The synthesis process followed a layered analytical framework, integrating four key dimensions: (i) computational architectures, (ii) classical algorithms, (iii) biological data infrastructure, and (iv) evaluation metrics. These dimensions correspond to the structured organization of Tables 1–4 in the manuscript, which collectively provide a scaffold for interpreting the evolution of the field. Computational architectures were analyzed in terms of their ability to capture sequence dependencies and structural relationships, with particular attention to the transition from sequential models to attention-based systems (Vaswani et al., 2017). Classical algorithms were evaluated for their efficiency and limitations in homology-based inference (Remmert et al., 2012), while biological datasets were examined as enabling factors for large-scale model training (Berman et al., 2000; UniProt Consortium, 2015).

Evaluation metrics, including structural similarity measures such as TM-score and RMSD, were incorporated to assess how predictive performance is quantified and validated (Zhang & Skolnick, 2004; Maiorov & Crippen,

1994). Benchmarking initiatives such as CASP were also considered as standardized platforms for comparing computational methods (Moult et al., 2018).

### 2.5 Limitations of the Methodological Approach

While the narrative methodology allows for flexibility and conceptual integration, it also introduces subjectivity in study selection and interpretation. To mitigate this, the review emphasizes widely recognized and foundational references, ensuring that conclusions are grounded in established scientific contributions. Nevertheless, the approach remains interpretive, reflecting the evolving and interdisciplinary nature of computational protein science.

## 3. Large Language Models in Bioinformatics: A Linguistic Turn in Understanding Biological Systems

### 3.1 Reframing Biology as Language: A Conceptual Shift

For a long time, bioinformatics has wrestled—sometimes productively, sometimes not—with a deceptively simple question: how does a linear biological sequence give rise to complex, three-dimensional, functional reality? Proteins, after all, are not merely chains of amino acids; they are dynamic entities, folding, interacting, and responding to their environment with remarkable precision. And yet, despite decades of research grounded in biochemical principles and structural biology, the translation from sequence to function has remained only partially resolved. The difficulty lies not in the lack of data—if anything, the opposite is true—but in how that data is interpreted.

Traditional approaches leaned heavily on evolutionary logic. Tools such as BLAST and HMMER were built on the assumption that similarity implies shared ancestry, and by extension, shared structure and function (Altschul et al., 1997; Eddy, 1998; Finn et al., 2011). In many cases, this assumption holds. But it also introduces a subtle limitation: these methods are, in essence, retrospective. They excel at recognizing patterns that have already been observed but struggle when confronted with novelty—proteins without homologs, or sequences that diverge significantly from known families. As biological datasets expanded, particularly with repositories like the Protein Data Bank (Berman et al., 2000), this limitation became increasingly visible. There was, quite simply, more unknown than known.

It is here—perhaps somewhat unexpectedly—that ideas from natural language processing (NLP) began to reshape the field. The notion that biological sequences could be

treated as a kind of language was not entirely new, but it gained real traction only when computational frameworks matured enough to support it. Amino acids, in this view, become tokens; sequences become sentences; and evolutionary constraints resemble grammar. The analogy is imperfect, certainly, but it is also surprisingly powerful. It allows us to think of biology not just as chemistry, but as information—structured, contextual, and, importantly, learnable.

### 3.2 From Sequential Models to Attention Mechanisms

Early attempts to model biological sequences using deep learning borrowed from architectures designed for text processing. Recurrent neural networks (RNNs), and more specifically long short-term memory (LSTM) networks, were among the first to be adapted for this purpose (Hochreiter & Schmidhuber, 1997). Their appeal was straightforward: they process sequences step by step, maintaining a form of memory that, in principle, captures context.

Yet, in practice, this approach proved limiting. Biological sequences are not just long—they are complex in a way that challenges sequential processing. Interactions between residues may span hundreds of positions, and capturing these long-range dependencies is critical for understanding protein folding and function. RNNs, constrained by their stepwise nature, often fail to maintain such distant relationships effectively. The well-documented vanishing gradient problem exacerbates this issue, gradually diminishing the influence of earlier sequence elements as processing continues.

The introduction of the Transformer architecture marked a decisive departure from this paradigm (Vaswani et al., 2017). By replacing recurrence with self-attention, Transformers allow every element in a sequence to interact with every other element simultaneously. This is not merely a computational convenience—it fundamentally changes what the model can represent. In the context of proteins, it means that distant residues can be linked directly, without the need for sequential propagation of information.

There is, perhaps, a tendency to describe this transition as a clean break. In reality, it feels more like an accumulation of insights reaching a tipping point. The mathematical formalism of attention aligns almost uncannily with the biological reality of protein structures, where spatial proximity often overrides linear distance. The result is a

model architecture that, almost by design, captures the kind of relationships that matter most in molecular systems.

### 3.3 Pre-training and the Emergence of Biological Representations

If the architecture provides the framework, pre-training provides the substance. Large language models in bioinformatics rely on self-supervised learning, a strategy that, at first glance, seems almost paradoxical. Instead of learning from labeled data—structures, functions, or experimental annotations—these models learn from raw sequences alone.

Two primary paradigms have emerged. The first, exemplified by BERT, uses masked language modeling, where portions of a sequence are hidden and the model is trained to predict them (Devlin et al., 2018). The second, associated with generative models like GPT, involves predicting the next token in a sequence (Radford et al., 2018). Both approaches, though conceptually simple, lead to surprisingly rich representations.

Through exposure to massive datasets—often comprising billions of sequences—models begin to internalize patterns that reflect underlying biological principles. Secondary structure motifs, residue co-evolution, and even aspects of protein stability appear to be encoded within the learned representations. What is striking is that this knowledge emerges without explicit supervision. The model is not told what a helix is, or how a binding site functions; it infers these patterns indirectly, through statistical regularities in the data.

This raises an interesting question: what, exactly, are these models learning? The answer is not entirely clear. They are not learning biology in the traditional sense, nor are they simply memorizing sequences. Instead, they appear to construct a latent representation space where biologically meaningful relationships are preserved. It is, in a way, an intermediate form of understanding—one that is useful, even if it is not fully interpretable.

### 3.4 From Sequence to Structure: Bridging a Longstanding Gap

Perhaps the most tangible impact of large language models has been in protein structure prediction. For many years, accurate prediction depended on multiple sequence alignments (MSAs), which provide evolutionary context by comparing related sequences. While effective, this

approach is computationally intensive and inherently limited by the availability of homologous sequences. Recent advances suggest that this dependency may not be as absolute as once thought. Protein language models, trained on single sequences, have demonstrated an ability to infer structural features directly. In some cases, they can predict three-dimensional conformations with remarkable accuracy, drawing on patterns embedded in their learned representations rather than explicit evolutionary comparisons. This development resonates, interestingly, with earlier theoretical insights. Anfinsen's principle—that sequence determines structure (Anfinsen, 1973)—is not new. What is new is the ability to operationalize that principle computationally, at scale. By integrating Transformer architectures with established structural knowledge, these models are narrowing the gap between sequence data and functional understanding.

The implications extend beyond basic science. In virology, for instance, the ability to rapidly assess the structural impact of mutations allows for early detection of potentially concerning variants. In therapeutic design, it enables the engineering of proteins with desired properties—binding specificity, stability, or catalytic efficiency—before they are synthesized in the lab.

### 3.5 Extending the Framework: Genomics and Cellular Systems

While proteins have been the primary focus, the application of large language models is not limited to proteomics. Genomic sequences, with their vast non-coding regions, present a different kind of challenge—one that is, perhaps, even more abstract. Regulatory elements such as promoters and enhancers do not follow simple patterns, and their function depends heavily on context. Transformer-based models, with their capacity for capturing global dependencies, are well-suited to this task. By analyzing entire genomic regions simultaneously, they can identify subtle patterns associated with gene regulation. This has led to improved predictions of transcription factor binding sites and regulatory interactions, offering new insights into gene expression mechanisms.

In transcriptomics, the paradigm shifts again. Here, the focus is not on sequences but on expression profiles—high-dimensional representations of cellular states. Treating these profiles as “sentences” allows language models to classify cell types, identify

subpopulations, and even infer developmental trajectories. The analogy to language is, admittedly, stretched in this context, but it remains conceptually useful.

### 3.6 Toward Rational Drug Discovery

The integration of large language models into drug discovery represents one of the most promising—and perhaps most consequential—applications. Traditional drug development is notoriously inefficient, characterized by high costs and low success rates. A significant portion of this inefficiency stems from the difficulty of predicting how a molecule will interact with its biological target. Language models offer a different approach. By embedding both proteins and chemical compounds into a shared representation space, they enable the prediction of drug–target interactions with increasing accuracy. Representations derived from sequence data can be combined with chemical encodings, such as SMILES strings, to model binding affinity and specificity.

This does not eliminate uncertainty, of course. Biological systems remain complex, and *in silico* predictions must be validated experimentally. But it does shift the balance—from blind screening toward informed design. Molecules can be prioritized based on predicted efficacy, reducing the need for exhaustive testing and accelerating the overall pipeline.

### 3.7 Challenges, Interpretability, and the Limits of Abstraction

Despite their successes, large language models are not without limitations. The most frequently cited concern is interpretability. These models can produce accurate predictions, but the reasoning behind those predictions is often opaque. For experimental scientists, this lack of transparency can be problematic. A prediction, no matter how accurate, is difficult to trust without some understanding of its basis. Efforts to address this issue include attention visualization and attribution methods, which attempt to highlight the parts of a sequence that contribute most to a prediction. While promising, these approaches are still evolving and do not fully resolve the underlying challenge. There are also practical considerations. Training large models requires significant computational resources, raising questions about accessibility and sustainability. Moreover, generative models may produce sequences that are syntactically valid but biologically implausible—a form of “hallucination” that underscores the need for careful validation.

### 3.8 Conclusion: Toward a Unified Biological Intelligence

What emerges from these developments is not simply a new set of tools, but a shift in perspective. Biology, increasingly, is being understood as an information system—one that can be modeled, decoded, and, to some extent, rewritten. Large language models do not replace traditional methods; rather, they complement them, offering new ways to interpret complex data. The idea of a “linguistic turn” in biology may still feel tentative, perhaps even metaphorical. And yet, it captures something real: the growing recognition that biological sequences carry information in ways that resemble language, and that this information can be learned through statistical models. As these approaches mature, they may bring us closer to a more unified understanding of life—one in which sequence, structure, and function are not treated as separate domains, but as interconnected aspects of a single system. Whether this vision will fully materialize remains to be seen. But the direction, at least, seems increasingly clear.

## 4. Synthesizing the Computational Evolution of Protein Science: From Alignment Heuristics to Representation Intelligence

### 4.1 Reconsidering the Algorithmic Foundations: From Retrospective Alignment to Context-Aware Learning

The computational study of proteins, if one looks at it over the past three decades, does not appear to have advanced in a straight line. Rather, it seems to have moved through phases—each defined not only by new tools but by shifting assumptions about what biological data actually represent. Initially, interpretation relied almost entirely on sequence similarity. Methods such as BLAST and its iterative extension PSI-BLAST enabled researchers to identify homologous proteins by aligning sequences against ever-expanding databases (Altschul et al., 1990; Altschul et al., 1997). These tools, as summarized in Table 3, were foundational—not simply because they worked, but because they encoded a powerful idea: that evolutionary proximity implies structural and functional similarity.

Yet, this approach carried a quiet limitation. It was, fundamentally, retrospective. It could only infer properties for proteins that resembled something already known. As databases grew—fueled by initiatives such as UniProt and structural repositories like the Protein Data Bank (PDB)—this limitation became more apparent (Berman et

al., 2000; Suzek et al., 2015). A large fraction of sequences remained without clear homologs, forming what is sometimes described as the “dark proteome.”

The transition away from this constraint began gradually, with the introduction of machine learning architectures capable of learning directly from raw data. As outlined in Table 1, early neural models such as recurrent neural networks (RNNs) and long short-term memory (LSTM) networks attempted to treat protein sequences as temporal signals (Elman, 1990; Hochreiter & Schmidhuber, 1997). These models introduced a degree of flexibility, allowing patterns to be learned rather than explicitly defined. Still, their sequential nature-imposed limitations, particularly in capturing long-range dependencies.

The emergence of the Transformer architecture marked a more decisive conceptual shift. By leveraging self-attention mechanisms, Transformers enabled the simultaneous evaluation of relationships between all elements in a sequence (Vaswani et al., 2017). This change, though technical in implementation, aligned more naturally with the physical reality of proteins, where residues distant in sequence may interact closely in three-dimensional space. As reflected in Table 1, this shift toward global context modeling represents not merely an improvement in performance but a redefinition of how sequence information is interpreted.

### 4.2 Data as Infrastructure: The Quiet Backbone of Computational Progress

If algorithms define how we interpret biological data, then datasets determine what can be learned in the first place. It is difficult—perhaps impossible—to separate the success of modern computational models from the maturation of biological databases over the same period. As detailed in Table 2, repositories such as PDB, UniProt, and Pfam have collectively provided both the structural ground truth and the vast sequence diversity required for model development (Berman et al., 2000; Finn et al., 2013; UniProt Consortium, 2015).

What is interesting, however, is how these datasets have been used. In earlier paradigms, curated databases served primarily as reference points—repositories against which new sequences were compared. In more recent approaches, they function as training corpora. Large-scale pre-training, particularly in Transformer-based models, relies on the availability of millions—often billions—of

**Table 1. Foundational Machine Learning Architectures for Bioinformatics (Pre-2019): Evolution from Sequential Modeling to Attention-Based Learning.** This table summarizes the major neural network architectures applied to biological data prior to the large-scale adoption of Transformer-based models. It highlights how computational paradigms evolved from sequential and local pattern recognition toward global context modeling and representation learning. The progression reflects increasing capability to capture long-range dependencies essential for protein structure and function prediction.

| Architecture | Primary Citation                | Core Mechanism           | Input Type          | Parallelism | Context Range     | Biological Application   | Key Insight                          |
|--------------|---------------------------------|--------------------------|---------------------|-------------|-------------------|--------------------------|--------------------------------------|
| RNN          | Elman (1990)                    | Recurrent hidden states  | Sequential tokens   | Low         | Short-range       | Sequence modeling        | Captured temporal dependencies       |
| LSTM         | Hochreiter & Schmidhuber (1997) | Gated memory units       | Sequential tokens   | Low         | Medium-range      | Protein folding          | Mitigated vanishing gradients        |
| CNN          | Fukushima (1980)                | Convolutional filters    | Grid/spatial        | High        | Local             | Motif detection          | Strong local feature extraction      |
| Word2Vec     | Mikolov et al. (2013)           | Embedding learning       | Token windows       | High        | Semantic          | Feature extraction       | Learned distributed representations  |
| Transformer  | Vaswani et al. (2017)           | Self-attention           | Tokenized sequences | Very High   | Long-range        | Structure prediction     | Modeled global dependencies          |
| BERT         | Devlin et al. (2018)            | Bidirectional attention  | Masked sequences    | Very High   | Deep context      | Function annotation      | Learned biological “grammar”         |
| GPT-1        | Radford et al. (2018)           | Autoregressive attention | Sequential tokens   | Very High   | Long-range        | Sequence generation      | Enabled generative modeling          |
| ResNet       | He et al. (2016)                | Residual learning        | Image/maps          | High        | Deep hierarchical | Contact prediction       | Solved degradation issues            |
| GNN          | Gilmer et al. (2017)            | Message passing          | Graph structures    | Medium      | Local/global      | Interaction prediction   | Modeled molecular interactions       |
| Autoencoder  | Kothari & Oh (1993)             | Latent compression       | Raw features        | High        | Feature-based     | Dimensionality reduction | Unsupervised representation learning |

**Table 2. Foundational Biological Databases and Repositories (Pre-2019): Data Infrastructure for Computational Biology**

This table outlines key biological databases that underpin computational modeling and validation. These repositories provide experimentally derived and curated datasets essential for training, benchmarking, and validating bioinformatics algorithms. Their diversity reflects the multidimensional nature of biological data—from sequences to structures and interactions.

| Database    | Primary Citation          | Data Type           | Size (Approx.) | Data Quality        | Update Frequency | Scope                 | Key Role                             |
|-------------|---------------------------|---------------------|----------------|---------------------|------------------|-----------------------|--------------------------------------|
| PDB         | Berman et al. (2000)      | 3D structures       | ~140k (2018)   | High (experimental) | Weekly           | Structural biology    | Gold standard for protein structures |
| UniProt     | UniProt Consortium (2015) | Protein sequences   | >100M          | Annotated           | Continuous       | Global proteome       | Universal sequence repository        |
| Pfam        | Finn et al. (2013)        | Protein families    | >16k families  | HMM-based           | Regular          | Domain classification | Evolutionary grouping                |
| SCOP        | Murzin et al. (1995)      | Structural folds    | >100k domains  | Manual              | Annual           | Structural hierarchy  | Fold classification                  |
| STRING      | Szklarczyk et al. (2015)  | Interactions        | >2k organisms  | Mixed               | Regular          | Networks              | Protein–protein interactions         |
| BioLiP      | Yang et al. (2013)        | Binding sites       | >30k complexes | High                | Periodic         | Functional sites      | Ligand interaction annotation        |
| CB513       | Cuff & Barton (1999)      | Secondary structure | 513 chains     | Benchmark           | Static           | Structure prediction  | Standard evaluation dataset          |
| CASP        | Moult et al. (2018)       | Prediction targets  | ~100/year      | Experimental        | Biennial         | Benchmarking          | Performance evaluation               |
| InterPro    | Mitchell et al. (2015)    | Functional domains  | >30k entries   | Curated             | Continuous       | Protein annotation    | Integrated signatures                |
| SWISS-MODEL | Waterhouse et al. (2018)  | Predicted models    | >400k          | High                | Continuous       | Homology modeling     | Template-based prediction            |

sequences. These data are not simply labeled examples; they are the raw material from which statistical patterns emerge.

This shift transforms what might have been considered a limitation—the sequence–structure gap—into something closer to an opportunity. By learning from unannotated sequences, models can infer latent relationships that are not explicitly encoded in the data. As suggested by the clustering of functional similarities in learned representation spaces, proteins with divergent sequences may still share underlying properties. In this sense, the role of databases has evolved from passive storage to active participation in model learning.

#### 4.3 Architectural Complementarity: Integrating Global Context with Local Precision

While the narrative of computational evolution often emphasizes the dominance of Transformers, the reality is somewhat more nuanced. Different architectures contribute distinct strengths, and their interplay remains important. As summarized in Table 1, convolutional neural networks (CNNs), for instance, continue to excel at detecting localized patterns such as conserved motifs or active sites (Fukushima, 1980). These features, though small in scale, are often critical for biological function.

At the same time, graph neural networks (GNNs) introduce yet another layer of representation by modeling proteins as relational structures (Gilmer et al., 2017). In this framework, amino acids—or even individual atoms—are treated as nodes, with edges representing chemical interactions. This approach captures spatial and geometric dependencies that are not easily expressed in sequence-based models alone.

The result is not a replacement of one paradigm by another, but a kind of architectural convergence. Transformers provide the global context—the ability to understand sequences holistically—while CNNs and GNNs offer localized and structural precision. This complementary relationship becomes particularly evident in applications such as protein–protein interaction prediction and drug design, where both global sequence understanding and atomic-level detail are required.

#### 4.4 Measuring Progress: From Simple Accuracy to Structural Fidelity

As computational models have become more sophisticated, so too have the metrics used to evaluate

them. Early assessments often relied on straightforward measures such as classification accuracy. However, as tasks shifted toward structural prediction and functional inference, more nuanced metrics became necessary.

Table 4 provides an overview of these evaluation frameworks. Metrics such as root-mean-square deviation (RMSD) and Template Modeling score (TM-score) quantify structural similarity in ways that are both interpretable and biologically meaningful (Maiorov & Crippen, 1994; Zhang & Skolnick, 2004). Similarly, the Global Distance Test (GDT\_TS), widely used in CASP benchmarking exercises, offers a robust measure of structural accuracy across multiple distance thresholds (Zemla, 2003; Moult et al., 2018).

In classification tasks, particularly those involving imbalanced datasets, the Matthews correlation coefficient (MCC) has emerged as a preferred metric due to its balanced treatment of true and false predictions (Matthews, 1975). Meanwhile, area under the ROC curve (AUC-ROC) provides a threshold-independent measure of model discrimination (DeLong et al., 1988).

These metrics do more than quantify performance—they shape how models are developed. By defining what counts as “success,” they influence optimization strategies and, ultimately, research priorities. In this sense, evaluation frameworks are not neutral; they are integral to the evolution of the field.

#### 4.5 Toward Integrated Intelligence: Bridging Sequence, Structure, and Function

Taken together, the developments outlined above suggest a broader shift in how protein science is conceptualized. The field is moving away from isolated analytical techniques toward integrated systems capable of linking sequence, structure, and function within a unified framework. This trajectory is evident when one considers the progression from alignment-based methods (Table 3), to neural architectures (Table 1), to large-scale data repositories (Table 2), and finally to standardized evaluation metrics (Table 4).

There is, perhaps, a temptation to view this progression as linear—a steady march toward increasing sophistication. But it may be more accurate to see it as a layering process, where new approaches build upon, rather than replace, earlier ones. Classical methods such as BLAST remain indispensable, even as Transformer-based models

**Table 3. Classical Algorithms and Computational Methods (Pre-2019): Foundations Before Deep Learning Dominance**

This table summarizes the key computational methods used prior to large-scale deep learning adoption. These approaches, rooted in statistical modeling, alignment, and physics-based simulations, formed the backbone of bioinformatics pipelines and continue to serve as benchmarks and complementary tools.

| Method             | Primary Citation         | Core Logic           | Input             | Application         | Output        | Efficiency | Contribution               |
|--------------------|--------------------------|----------------------|-------------------|---------------------|---------------|------------|----------------------------|
| BLAST              | Altschul et al. (1990)   | Local alignment      | Sequence          | Homology search     | E-value       | High       | Fast database search       |
| PSI-BLAST          | Altschul et al. (1997)   | Iterative alignment  | Sequence          | Profile building    | PSSM          | Medium     | Increased sensitivity      |
| HMMER              | Finn et al. (2011)       | Hidden Markov model  | MSA               | Domain detection    | Score         | High       | Statistical modeling       |
| HHblits            | Remmert et al. (2012)    | HMM-HMM alignment    | Profiles          | Remote homology     | Probability   | High       | Ultra-fast alignment       |
| DSSP               | Kabsch & Sander (1983)   | Geometry rules       | Structure         | Secondary structure | Labels        | High       | Standard annotation        |
| TM-align           | Zhang & Skolnick (2005)  | Structural alignment | Structures        | Comparison          | TM-score      | Medium     | Length-independent metric  |
| CD-HIT             | Fu et al. (2012)         | Clustering           | Sequence database | Redundancy removal  | Clusters      | Very High  | Dataset reduction          |
| Monte Carlo        | Karplus & Kuriyan (2005) | Stochastic sampling  | Energy model      | Folding simulation  | Conformations | Low        | Folding pathway modeling   |
| PSSM               | Jones (1999)             | Scoring matrix       | MSA               | Feature encoding    | Matrix        | Medium     | Secondary prediction basis |
| Molecular Dynamics | Gsponer et al. (2002)    | Physics simulation   | Atomic data       | Folding dynamics    | Trajectories  | Very Low   | High-resolution simulation |

**Table 4. Standard Evaluation Metrics in Bioinformatics (Pre-2019): Quantifying Model Performance**

This table presents commonly used evaluation metrics for assessing predictive models in bioinformatics. These metrics span structural accuracy, classification performance, and statistical robustness, providing standardized benchmarks for comparing computational approaches.

| Metric              | Primary Citation         | Basis                | Range   | Strength                | Limitation            | Benchmark        | Application           |
|---------------------|--------------------------|----------------------|---------|-------------------------|-----------------------|------------------|-----------------------|
| TM-score            | Zhang & Skolnick (2004)  | Structure similarity | 0–1     | Length-independent      | Ignores local errors  | >0.5             | Structure prediction  |
| RMSD                | Maierov & Crippen (1994) | Atomic distance      | ≥0      | Physical interpretation | Sensitive to outliers | <2 Å             | Structural comparison |
| GDT_TS              | Zemla (2003)             | Distance thresholds  | 0–100   | Robust                  | Ignores connectivity  | CASP standard    | Folding accuracy      |
| MCC                 | Matthews (1975)          | Confusion matrix     | -1 to 1 | Balanced metric         | Hard to interpret     | >0.5             | Classification        |
| AUC-ROC             | DeLong et al. (1988)     | ROC curve            | 0.5–1   | Threshold-free          | Skew bias             | High = good      | Binary classification |
| Q3                  | Rost & Sander (1993)     | 3-state accuracy     | 0–100%  | Simple                  | Coarse                | >80%             | Secondary structure   |
| Q8                  | Kabsch & Sander (1983)   | 8-state accuracy     | 0–100%  | Detailed                | Difficult             | DSSP standard    | Structure prediction  |
| Spearman ρ          | Spearman (1904)          | Rank correlation     | -1 to 1 | Nonparametric           | Limited linearity     | High correlation | Regression            |
| Information Content | Shannon (1948)           | Probability entropy  | ≥0      | Specificity measure     | Data-dependent        | —                | Functional annotation |
| Precision           | Standard ML              | TP/(TP+FP)           | 0–1     | Specificity             | Ignores FN            | High desired     | Binding prediction    |

redefine what is possible.

What has changed most fundamentally is not the tools themselves, but the perspective. Proteins are no longer viewed solely as biochemical entities to be experimentally characterized; they are increasingly treated as information systems that can be learned, modeled, and, in some cases, designed. The “language of life” metaphor, while imperfect, captures this shift reasonably well.

#### 4.6 Concluding Synthesis: From Descriptive Models to Generative Understanding

If one were to distill the trajectory of computational protein science into a single observation, it might be this: the field has transitioned from describing biological systems to actively modeling—and even generating—them. Early methods, grounded in alignment and statistical inference, provided valuable insights but were constrained by existing knowledge. Modern approaches, particularly those based on deep learning, extend beyond these limitations by learning directly from data. The integration of architectures such as Word2Vec, ResNet, and Transformer models has enabled the construction of high-dimensional representation spaces where biological meaning is encoded implicitly (Mikolov et al., 2013; He et al., 2016; Vaswani et al., 2017). Within these spaces, relationships between sequences, structures, and functions emerge not through explicit programming, but through learned patterns. This does not imply that the problem is solved. Questions of interpretability, generalizability, and biological validity remain. Yet, the progress is difficult to overlook. For the first time, it seems plausible that the information contained within amino acid sequences can be decoded in a way that is both comprehensive and predictive. In that sense, the field may be approaching something like fluency—not in a literal language, but in the structured complexity of biological systems. And while that fluency is still imperfect, it marks a significant step toward understanding the molecular logic that underlies life itself.

#### 5. Limitations

This review, while comprehensive in scope, is not without its limitations. As a narrative synthesis, it necessarily reflects selective interpretation rather than exhaustive coverage. Certain emerging models and very recent developments may not be fully represented, particularly those evolving beyond the foundational frameworks discussed here. There is also an inherent bias toward

well-cited and historically influential studies, which, while important, may overshadow less prominent yet potentially innovative contributions. Additionally, the rapid pace of advancement in deep learning introduces a temporal limitation; models that are considered state-of-the-art today may quickly become outdated. The review also emphasizes computational perspectives, which may underrepresent experimental validation challenges and biological variability. Finally, the interpretability of Transformer-based models remains an unresolved issue, and this review cannot fully address the gap between predictive accuracy and mechanistic understanding.

#### 6. Conclusion

The evolution of computational protein science reflects more than technological progress—it reveals a gradual shift in how biological systems are conceptualized. What once relied on alignment and inference now leans toward representation and prediction. Transformer-based models, in this context, do not simply improve performance; they redefine the questions we can ask. Yet, their promise is accompanied by uncertainty, particularly regarding interpretability and generalizability. It may be too early to claim full understanding, but it is difficult to ignore the direction. Protein sequences are beginning to look less like static codes and more like dynamic, interpretable systems—ones we are only just learning to read.

#### Author Contributions

S.J.M. conceptualized the study, designed the review framework, conducted literature synthesis and analysis, interpreted the findings, and drafted, reviewed, and finalized the manuscript.

#### References

- Altschul, S. F., Gish, W., Miller, W., Myers, E. W., & Lipman, D. J. (1990). Basic local alignment search tool. *Journal of Molecular Biology*, 215(3), 403–410.
- Altschul, S. F., Madden, T. L., Schaffer, A. A., J. Zhang, Z. Zhang, Miller, W., & Lipman, D. J. (1997). Gapped BLAST and PSI-BLAST: A new generation of protein database search programs. *Nucleic Acids Research*, 25(17), 3389–3402.
- Anfinsen, C. B. (1973). Principles that govern the folding of protein chains. *Science*, 181(4096), 223–230.
- Armenteros, J. J. A., Sonderby, C. K., Sonderby, S. K., Nielsen, H., & Winther, O. (2017). DeepLoc: prediction of protein subcellular localization using deep learning. *Bioinformatics*, 33(21), 3387–3395.
- Berman, H. M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T. N., Weissig, H., ... & Bourne, P. E.

- (2000). The Protein Data Bank. *Nucleic Acids Research*, 28(1), 235–242.
- Cuff, J. A., & Barton, G. J. (1999). Evaluation and improvement of multiple sequence methods for protein secondary structure prediction. *Proteins: Structure, Function, and Bioinformatics*, 34(4), 508–519.
- DeLong, E. R., DeLong, D. M., & Clarke-Pearson, D. L. (1988). Comparing the areas under two or more correlated receiver operating characteristic curves: a nonparametric approach. *Biometrics*, 837–845.
- Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2018). BERT: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.
- Eddy, S. R. (1998). Profile hidden Markov models. *Bioinformatics*, 14(9), 755–763.
- Elman, J. L. (1990). Finding structure in time. *Cognitive Science*, 14(2), 179–211.
- Finn, R. D., Bateman, A., Clements, J., Coghill, P., Eberhardt, R. Y., Eddy, S. R., ... & Punta, M. (2013). Pfam: the protein families database. *Nucleic Acids Research*, 42(D1), D222–D230.
- Finn, R. D., Clements, J., & Eddy, S. R. (2011). HMMER web server: Interactive sequence similarity searching. *Nucleic Acids Research*, 39(suppl\_2), W29–W37.
- Fukushima, K. (1980). Neocognitron: A self-organizing neural network model for a mechanism of pattern recognition unaffected by shift in position. *Biological Cybernetics*, 36(4), 193–202.
- Gilmer, J., Schoenholz, S. S., Riley, P. F., Vinyals, O., & Dahl, G. E. (2017). Neural message passing for quantum chemistry. *arXiv preprint arXiv:1704.01212*.
- He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (pp. 770–778).
- Hochreiter, S., & Schmidhuber, J. (1997). Long short-term memory. *Neural Computation*, 9(8), 1735–1780.
- Jones, D. T. (1999). Protein secondary structure prediction based on position-specific scoring matrices. *Journal of Molecular Biology*, 292(2), 195–202.
- Kabsch, W., & Sander, C. (1983). Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features. *Biopolymers*, 22(12), 2577–2637.
- Levinthal, C. (1968). Are there pathways for protein folding? *Journal de Chimie Physique*, 65, 44–45.
- Maierov, V. N., & Crippen, G. M. (1994). Significance of root-mean-square deviation in comparing three-dimensional structures of globular proteins. *Journal of Molecular Biology*, 235(2), 625–634.
- Matthews, B. W. (1975). Comparison of the predicted and observed secondary structure of T4 phage lysozyme. *Biochimica et Biophysica Acta (BBA)-Protein Structure*, 405(2), 442–451.
- Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013). Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*.
- Moult, J., Fidelis, K., Krysztafowicz, A., Schwede, T., & Tramontano, A. (2018). Critical assessment of methods of protein structure prediction (CASP)—Round XII. *Proteins: Structure, Function, and Bioinformatics*, 86, 7–15.
- Radford, A., Narasimhan, K., Salimans, T., & Sutskever, I. (2018). Improving language understanding by generative pre-training. *OpenAI Technical Report*.
- Radford, A., Narasimhan, K., Salimans, T., & Sutskever, I. (2018). Improving language understanding by generative pre-training. *OpenAI*.
- Remmert, M., Biegert, A., Hauser, A., & Söding, J. (2012). HHblits: lightning-fast iterative protein sequence searching by HMM-HMM alignment. *Nature Methods*, 9(2), 173–175.
- Rocklin, G. J., Chidyausiku, T. M., Goreshnik, I., Ford, A., Houlston, S., Lemak, A., ... & Baker, D. (2017). Global analysis of protein folding using massively parallel design, synthesis, and testing. *Science*, 357(6347), 168–175.
- Sarkisyan, K. S., Bolotin, D. A., Meer, M. V., Usmanova, D. R., Mishin, A. S., Sharonov, G. V., ... & Kondrashov, F. A. (2016). Local fitness landscape of the green fluorescent protein. *Nature*, 533(7603), 397–401.
- Suzek, B. E., Wang, Y., Huang, H., McGarvey, P. B., & Wu, C. H. (2015). UniRef clusters: A comprehensive and scalable alternative for improving sequence similarity searches. *Bioinformatics*, 31(6), 926–932.
- Szklarczyk, D., Franceschini, A., Wyder, S., Forslund, K., Heller, D., Huerta-Cepas, J., ... & von Mering, C. (2015). STRING v10: protein–protein interaction networks, integrated over the tree of life. *Nucleic Acids Research*, 43(D1), D447–D452.
- Transformer-Based Deep Learning for Protein Structure and Function Prediction: From Sequence Understanding to Biological Insight: A Review
- UniProt Consortium. (2015). UniProt: A hub for protein information. *Nucleic Acids Research*, 43(D1), D204–D212.
- Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., & Polosukhin, I. (2017). Attention is all you need. *Advances in Neural Information Processing Systems*, 30, 5998–6008.
- Waterhouse, A., Bertoni, M., Bienert, S., Studer, G., Tauriello, G., Gumienny, R., ... & Schwede, T. (2018). SWISS-MODEL: homology modelling of protein structures and complexes. *Nucleic Acids Research*, 46(W1), W296–W303.
- Yang, J., Roy, A., & Zhang, Y. (2013). BioLiP: A semi-manually curated database for biologically relevant ligand–protein interactions. *Nucleic Acids Research*, 41(D1), D1096–D1103.
- Zemla, A. (2003). LGA: a method for finding 3D structural similarities of macromolecules. *Nucleic Acids Research*, 31(13), 3370–3374.
- Zhang, Y., & Skolnick, J. (2004). Scoring function for automated assessment of protein structure template quality. *Proteins: Structure, Function, and Bioinformatics*, 57(4), 702–710.