Generative Artificial Intelligence for Biomolecule: Toward Unifying Models, Algorithms, and Modalities

Xiner Li[1,7*†] **Xingyu Su**[1*] **Yuchao Lin**[1,11] **Chenyu Wang**[2] **Yijia Xiao**[3] **Tianyu Liu**[4] **Chi Han**[5] **Michael Sun**[2] **Montgomery Bohde**[1] **Anna Hart**[5] **Wendi Yu**[1] **Masatoshi Uehara**[6] **Gabriele Scalia**[7] **Xiao Luo**[8] **Carl Edwards**[7] **Wengong Jin**[9,10] **Jianwen Xie**[11] **Ehsan Hajiramezanali**[7] **Edward De Brouwer**[7] **Qing Sun**[12] **Byung-Jun Yoon**[13,16] **Xiaoning Qian**[1,13,16] **Marinka Zitnik**[14] **Heng Ji**[5] **Hongyu Zhao**[4] **Wei Wang**[3] **Shuiwang Ji**[1,15,17†]

[1]Department of Computer Science and Engineering, Texas A&M University

[2]Computer Science and Artificial Intelligence Laboratory, Massachusetts Institute of Technology

[3]Department of Computer Science, University of California, Los Angeles

[4]Interdepartmental Program of Computational Biology and Bioinformatics, Yale University

[5]Siebel School of Computing and Data Science, University of Illinois Urbana Champaign

[6]Chan Zuckerberg Initiative [7]Genentech, Inc.

[8]Department of Statistics, University of Wisconsin–Madison [9]Broad Institute of MIT and Harvard

[10]Khoury College of Computer Sciences, Northeastern University [11]Lambda, Inc.

[12]Artie McFerrin Department of Chemical Engineering, Texas A&M University

[13]Department of Electrical and Computer Engineering, Texas A&M University

[14]Department of Biomedical Informatics, Harvard Medical School

[15]Department of Materials Science and Engineering, Texas A&M University

[16]Computing and Data Sciences, Brookhaven National Laboratory

[17]J. Mike Walker ’66 Department of Mechanical Engineering, Texas A&M University

***: Equal contribution**

†: Correspondence to: Xiner Li <lxe@tamu.edu>, Shuiwang Ji <sji@tamu.edu>

Abstract:

Rapid advances in generative artificial intelligence have revolutionized biological modeling across domains such as protein, genetics, and single-cell. However, existing works often organize applications by molecule types or specific research tasks, overlooking the methodological convergence and cross-modal innovations. This paper aims to present a unified methodological perspective that highlights the fundamental technical commonalities across biological modalities. We systematically organize recent advances in generative modeling for biology through the lens of core machine learning paradigms, from language models (LMs) and diffusion models to their emerging hybrid architectures. Our work reveals how techniques initially developed for one molecular type (*e.g.*, protein design) can be effectively transferred to others (*e.g.*, RNA engineering), and identifies the convergence trend where discrete diffusion models and iterative language models represent different facets of a unified generative framework. We cover the evolution from domain-specific models to multi-modal biological foundation models and agent-based systems. By emphasizing methodological connections rather than applications, this paper aims to accelerate cross-domain innovation and make the field more accessible to the broader machine learning community. We conclude by identifying promising research directions where successful techniques in one biological domain remain unexplored in others, offering a roadmap for future advances in generative biology.

Contents

1. Introduction

Generative artificial intelligence is transforming how scientists reason about living systems, advancing modern computational biology, moving from static prediction to active design and exploration. In natural language and vision, recent advanced models such as large language models (LLMs) [273, 3, 46, 284, 63, 30, 96, 125] and diffusion models [250, 113, 252] have already reshaped workflows for analysis, synthesis, and interaction. In the realm of AI for science, a similar shift is underway, with recent breakthroughs spawning a new paradigm in computational life sciences [31, 341]. Generative models now propose protein sequences with desired structure [2], design regulatory DNA for precise expression control [149], and explore RNA landscapes that once required painstaking experimental search [15]. At the same time, generative models are beginning to operate over richer biological objects such as three dimensional structures, multi omics profiles, and even experimental protocols, bringing computational design closer to the full complexity of living systems [124].

Despite this rapid progress, the generative biology literature remains highly fragmented. Most work is structured around individual molecular types or specific applications, for example protein design [303], regulatory genomics [13], or RNA structure modeling [293]. Summarization works largely mirror this segmentation and often focus on a single modality or model family, such as LLMs [300, 155, 339, 134, 311, 354, 198, 55]. As a result, it is difficult to see how common methodological ideas reappear across modalities such as DNA, RNA, and proteins or how advances in one domain could be transferred to another. It also obscures a second dimension of fragmentation, where closely related techniques are developed independently within different generative paradigms such as language modeling [194, 336, 57, 6], diffusion and flow matching [120, 328, 56, 116], or hybrid sequence–structure architectures [245, 201].

In this paper we take a technique-first and molecule-agnostic view of generative models for biological sequences and structures. Rather than organizing by domain or task, we ask how core modeling ideas emerge, evolve, and converge across biological settings. We view various modalities as different manifestations of information bearing polymers that share common challenges, including discrete sequence alphabets, long range dependencies, rich structural constraints, and strong selection on function. From this perspective, many architectures and training strategies are naturally reusable across molecular types, even when the original work targets only one domain.

Specifically, we organize the paper around multiple levels of methodological convergence, as illustrated in Figure 1. We first examine how different architectures, such as language models and diffusion or flow-matching models have been adapted to biological settings, including tokenization schemes, conditioning mechanisms, and objective design. We then show how these families increasingly blend into hybrid and unified architectures that couple sequence, structural, and evolutionary information, such as protein structure predictors and long-context genomic models. We also consider how fine-tuning and inference-time techniques impose biological constraints and preferences without retraining models from scratch. This includes supervised and reinforcement style adaptation, training based preference optimization, and training free guidance methods for steering generation toward functional sequences or structures, as well as value based and search based strategies at inference. Finally, we study how recent advancements, such as foundation models and multi-agent systems, connect biological modalities with each other and with natural language. These systems learn joint representations across sequences, structures, omics readouts, and textual descriptions, and increasingly act as agents that orchestrate tools, databases, and wet lab experiments in closed loops for design and discovery.

Across different levels, we highlight a recurring pattern. Techniques first appear in a narrow domain, then re-emerge in other biological settings with minor modifications, and eventually converge into more unified frameworks. Sequence language models for protein design motivate similar architectures for regulatory

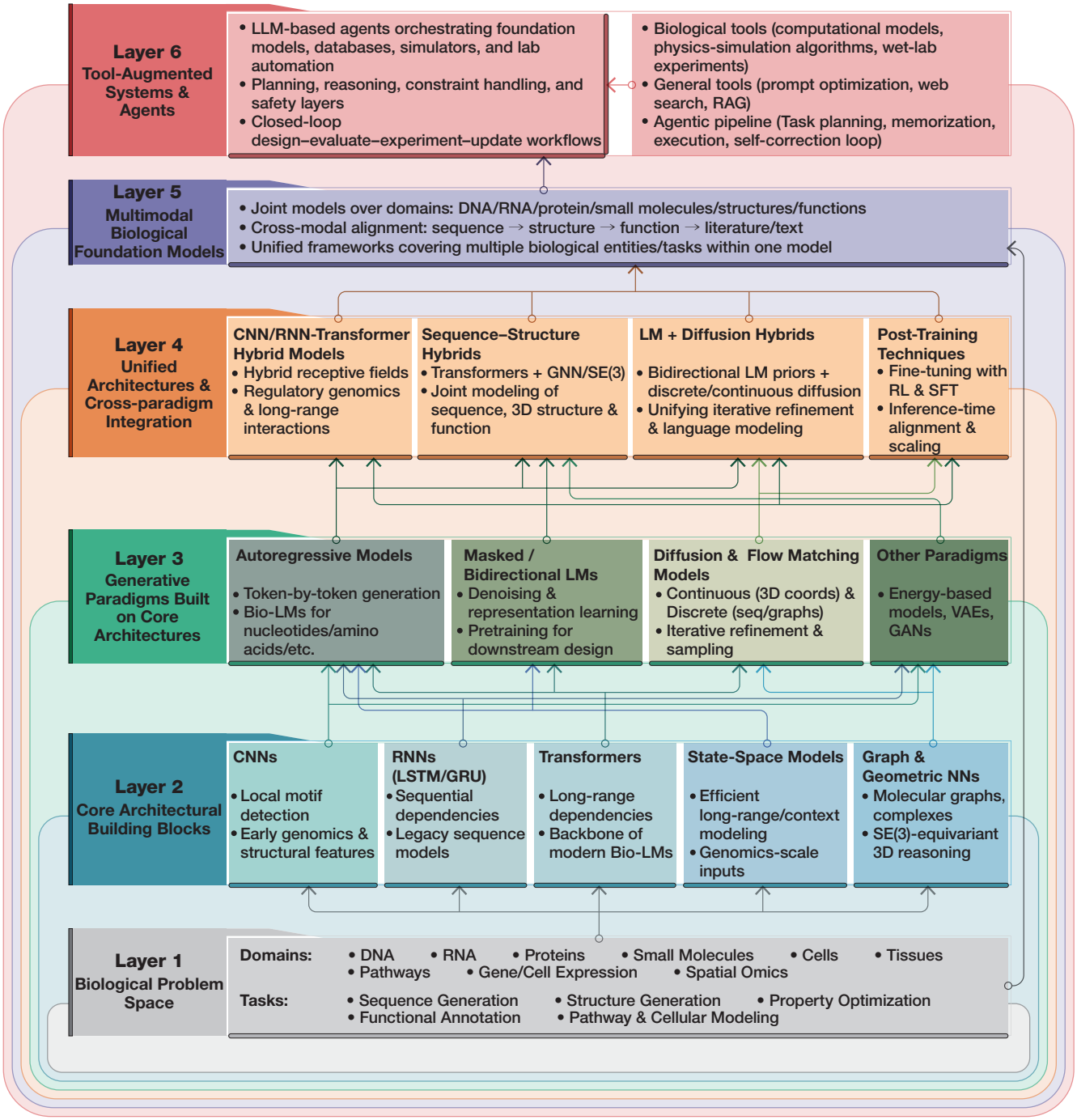

Figure 1: A layered taxonomy of generative AI for biology. Our perspective organizes the field from biological problem spaces and core architectural building blocks, through major generative paradigms such as autoregressive, masked/bidirectional, and diffusion/flow matching, to unified cross-paradigm hybrids, multimodal biological foundation models, and tool-augmented agentic systems for closed-loop discovery.

DNA and RNA. Discrete diffusion and flow matching models initially developed for small molecules are starting to inform protein and RNA work. Structure aware architectures pioneered in protein folding and

design are now being generalized to nucleic acids and multi molecular complexes. Post training methods for controlling natural language generation inspire analogous strategies for steering biological generation toward experimental objectives.

Our paper aims to make these convergences explicit and actionable. For machine learning researchers, we provide a taxonomy that connects familiar concepts from language, vision, and reinforcement learning to biological challenges, and identify opportunities where established techniques have not yet been fully explored in biology. For domain scientists, we offer a unifying view that links apparently disparate models under shared design principles, making it easier to navigate the growing model zoo and to select or adapt methods for concrete applications.

The remainder of the paper is organized as follows. Section 2 introduces biological and modeling preliminaries that recur throughout the paper. Section 3 describes language models for biological sequence generation, including tokenization, architecture choices, and pretraining objectives. Section 4 covers diffusion and flow matching models for both discrete sequences and continuous structures. Section 5 examines hybrid and unified architectures that integrate sequence, structure, and evolutionary information. Section 6 discusses post-training and inference-time techniques that align generative models with biological constraints and design goals. Section 7 discusses foundation models and agentic systems that operate across modalities and interface with external tools and experiments. Section 8 discusses methodological gaps and future directions, and Section 9 concludes with a perspective on how convergent generative modeling may reshape biological discovery and design. By emphasizing the shared technical foundations rather than domain-specific details, we aim to accelerate the emergence of truly general-purpose generative models for biology.

2. Background and Preliminaries

2.1. Biomolecular Generative Tasks

Recent advances in deep generative modeling have enabled a broad spectrum of biomolecular generative tasks spanning molecular and cellular modalities [2, 105, 181, 29, 13, 316]. These tasks aim to learn the underlying distributions of biological systems and to generate novel, plausible, or optimized instances that support downstream applications such as synthetic biology, drug discovery, and functional design.

In the field of DNA sequence design, research efforts mainly focus on two directions. One aims at genome-scale generation, seeking to model large genomic regions to demonstrate the capacity of generative frameworks [206, 236, 29, 192]. The other emphasizes the design of short regulatory elements such as promoters and enhancers, playing critical roles in controlling gene expression [12, 13, 255, 234]. While genome-scale modeling serves as a proof of generative capability, regulatory element design is of greater practical relevance, as it directly relates to synthetic biology applications such as gene expression control and cellular circuit construction.

In the domain of protein generative modeling, recent advances have enabled diverse tasks spanning from sequence creation to 3D structure generation and function optimization. Broadly, these approaches can be categorized into three major paradigms. First, unconditional de novo design aims to explore the vast protein sequence and structure space without external conditioning. This includes both de novo sequence generation [7, 296, 105, 208], which seeks novel and diverse amino acid sequences, and structure generation [303, 130, 306], which constructs plausible backbones or folds directly from latent representations. Such models often capture the intrinsic distribution of natural proteins and serve as priors for downstream conditional tasks. Second, cross-modal conditional generation establishes explicit mappings between sequence and structure. This encompasses sequence-to-structure generation (folding) [16, 166, 2], predicting 3D

conformations from given sequences, and structure-to-sequence generation (inverse folding) [60, 117, 327], inferring compatible sequences for target backbones. These bidirectional formulations jointly model the sequence–structure relationship and enable controllable protein design within learned manifolds. Third, function-conditioned generation integrates biochemical or biophysical constraints into the generative process. Models in this category aim to generate or optimize sequences and structures with specific functional goals, such as enhanced binding affinity, catalytic activity, stability, or solubility. Representative applications include protein and peptide binder design [213, 43, 21], antibody design [271, 186, 20], and protein property optimization [223, 93].

Beyond nucleic acids and proteins, generative modeling is increasingly applied to higher-order biological systems. At the cellular level, models are applied to single-cell generative modeling, simulating realistic cell states, and gene expression landscapes across heterogeneous populations [181, 251, 185]. At the spatial level, spatial omics generation seeks to reconstruct or synthesize spatially resolved molecular profiles such as transcriptomic maps, enabling virtual tissue modeling and data augmentation for spatial assays [22, 153]. At the chemical level, molecular graph generation focuses on designing small molecules or drug-like compounds with desired properties, bridging biomolecular and chemical design spaces [136, 218, 316]. Collectively, these emerging directions indicate that generative models now encompass multiple levels of biological modeling, spanning molecules, cells, and systems, and thereby advancing both our understanding and engineering of living systems.

2.2. Core Modeling Paradigms

Recent advances in biomolecular generative modeling are grounded in several core modeling paradigms. Among them, autoregressive language models, diffusion-based models, and graph-based architectures form the dominant methodological foundations.

Autoregressive language models treat DNA, RNA, or protein sequences as discrete symbol streams and generate them in a token-by-token manner, analogous to natural language modeling [206, 29, 310, 78, 260]. By learning long-range dependencies among nucleotides or amino acids, these models capture the grammatical and evolutionary regularities of biological sequences and implicitly encode structural and functional constraints. Importantly, their token-based formulation makes them naturally compatible with other modalities, such as text or properties, enabling conditional generation across multiple modalities. As a result, autoregressive models not only reproduce natural sequence distributions but also facilitate unconditional generation of novel biomolecules and conditional design guided by prompts, templates, or structural information.

Diffusion-based generative models provide a complementary paradigm for modeling both continuous structural and discrete sequence spaces. Unlike autoregressive models that generate tokens sequentially, diffusion models learn a gradual denoising process that transforms random noise into structured biological data. This formulation enables them to approximate complex, multi-modal distributions and to capture global context that are difficult to model through left-to-right prediction. In the continuous domain, such as protein folding or molecular geometry, diffusion processes operate on 3D coordinates or latent representations, allowing physically consistent and diversity-rich structure generation [303, 130, 56, 316, 328]. In the discrete domain, masked or uniform diffusion variants enable sequence-level generation and controlled mutation sampling [7, 94, 232, 296, 105]. Owing to their flexible conditioning and iterative refinement, diffusion models have become a central tool for controllable generation across different biomolecular modalities such as DNA, RNA, proteins, and small molecules.

Graph-based architectures offer a geometric and relational view of biomolecular modeling [129, 258,

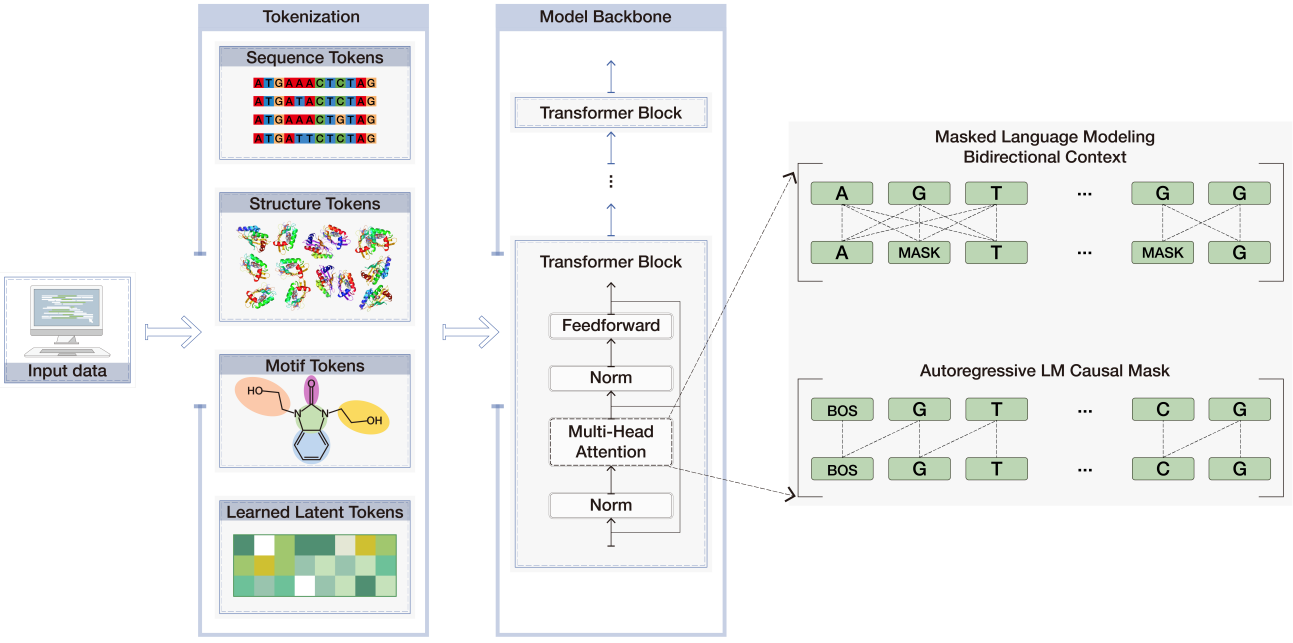

Figure 2: An overview of language-modeling pipelines for biological generation, illustrating how diverse biological inputs are tokenized, processed by a Transformer backbone, and trained under either masked bidirectional-context modeling or autoregressive causal modeling objectives.

235, 60, 98, 159, 171]. By representing residues or atoms as nodes and their spatial or chemical interactions as edges, these models explicitly encode the 3D topology of molecules. Through message passing and geometric equivariance, graph neural networks capture both local structural constraints and long-range dependencies, supporting tasks such as structure-conditioned sequence design or property prediction. Graph-based models offer valuable geometric inductive biases and can complement other paradigms to enhance structural consistency and interpretability in generative modeling.

3. Language Models for Biological Generation

Biological sequences share remarkable similarities with natural language. Both are built from a finite set of building blocks, and the combination of and connections between building blocks enable a combinatorial amount of information to be stored and transmitted. Both languages are mainly sequential with local structure, carrying functional information in local pieces while allowing long-range interactions across the sequence that determine the overall purpose of the sequence. The application of language modeling to biological sequences represents one of the most successful transfers from natural language processing to computational biology. Just as natural language consists of words forming sentences with semantic meaning, biological sequences comprise nucleotides or amino acids that encode functional information. This fundamental analogy has driven the development of increasingly sophisticated language models for DNA, RNA, proteins, and beyond, revealing that the same architectural innovations that revolutionized NLP can unlock new capabilities in biological sequence understanding and generation. Figure 2 summarizes the tokenization–backbone–objective workflow of language models for biological generation, from biological token design to Transformer modeling and masked/autoregressive training.

3.1. Tokenization Strategies

The choice of tokenization, *i.e.*, how continuous sequences are segmented into discrete units for model processing, fundamentally shapes what biological language models can learn and represent. Unlike natural language with clear word boundaries, biological sequences present unique challenges: they lack explicit delimiters, contain overlapping functional units, and encode information at multiple granularities simultaneously.

Single-Character Tokenization. The most straightforward approach treats each nucleotide or amino acid as an individual token. DNA models using single-nucleotide tokenization (A, C, G, T) maintain a vocabulary of just 4 tokens plus special symbols, while protein models require 20 standard amino acids plus additional tokens for ambiguous residues (X), gaps (-), and special markers. This fine-grained approach preserves maximum biological resolution and naturally handles point mutations, but requires models to learn higher-order patterns from scratch. ESM [230] and ProtTrans [72] exemplify successful single amino acid tokenization. For DNA, early models like DNABERT [133] initially avoided single-nucleotide tokenization due to the extreme sparsity of the 4-letter alphabet, though recent work shows this concern may have been overstated.

Motif/Fragment-Based Tokenization. To encode with more biological inductive bias and ensure more informative local pieces, many works consider motifs as tokens, which also improves efficiency with reduced input length and generation steps. K-mer tokenization segments sequences into overlapping or non-overlapping subsequences of length k. For DNA with k=6, this creates a vocabulary of $4^6 = 4096$ possible tokens, providing a richer embedding space while maintaining biological interpretability. DNABERT [133] pioneered k-mer tokenization with k=3,4,5,6, showing improved performance over single-nucleotide approaches on regulatory element prediction. The biological motivation for k-mers is compelling: transcription factors recognize specific 6-12bp motifs, codons comprise 3 nucleotides, and protein domains span characteristic length scales. However, k-mer tokenization introduces complications: overlapping k-mers create artificial dependencies, non-overlapping k-mers can miss patterns at boundaries, and the vocabulary size grows exponentially with k, limiting practical values to $k \leq 8$ for DNA. Several models explicitly incorporate the genetic code into tokenization. CodonBERT [156] treats each three-nucleotide codon as a token, creating a natural 64-token vocabulary that respects the fundamental unit of translation. This approach elegantly handles synonymous mutations and maintains reading frame information, crucial for coding sequence analysis. CodonTransformer [73] extends this concept by jointly modeling codon sequences and their amino acid translations, enabling cross-species optimization while respecting organism-specific codon usage biases. For RNA sequences, UTR-LM [49] employs specialized tokenization for 5' and 3' untranslated regions, recognizing that these regulatory sequences follow different grammars than coding regions. For RNA, OmniGenome includes secondary structural information in tokenization of RNA sequences [319].

Embedding-Based Tokenization and Learned Vocabularies. Targeting more advanced encoding techniques and flexible methods to control the vocabulary size, tokenization with learned vocabularies or embeddings has emerged. Byte-pair encoding (BPE) and similar compression algorithms learn data-driven vocabularies by iteratively merging frequent character pairs. DNABERT-2 [351] shifted from fixed k-mers to BPE, discovering that learned vocabularies better capture species-specific and functional patterns. A benchmark on biological tasks [68] also discover BPE to achieve superior compression and downstream task performance. The Nucleotide Transformer [58] similarly employs BPE with vocabularies ranging from 4,096 to 32,000 tokens, finding optimal performance around 12,000 tokens. It is important to note BPE can still underperform k-mers for nucleotide-resolution tasks [206], highlighting the tradeoff of resolution for efficiency. Furthermore, biological sequences lack the Zipfian distribution of natural language, making vocabulary learning less effective than in NLP. There are inherent linguistic differences between text and

biological sequences. Notably, BPE on amino acid sequences do not agree well with domain boundaries [265], a finding consistent with observations BPE tokens do not overlap with regulatory motifs [168], highlighting the fact learned vocabularies can obscure biological interpretation in the absence of additional context like structure or domain annotations.

Recent work on proteins and small molecules explore learned tokenization schemes that directly process structural information [160, 317, 80]. Vector-quantized autoencoders [281] are a popular class of deep learning models that embed high-dimensional structure into discrete structural vocabularies. ESM3 [105] uses VQ-VAE as a structure encoder to learn a vocabulary of per-residue geometric patterns and jointly model structure alongside text and sequence. FoldSeek [282] uses VQ-VAE to compress local, tertiary structural features into 20 learned discrete codes, called 3Di alphabets, which have found use for efficient search, homology detection, and end-to-end learning with amino acid sequences [108, 259, 154]. GeoBPE [263] invents a geometric analog of BPE for protein structure to automatically discover a hierarchical vocabulary of geometric motifs. This enables conversion of protein *structures* into discrete sequences that integrate naturally with language models for autoregressive modeling of protein geometry.

Special Tokens and Biological Semantics. Beyond sequence elements, biological language models require specialized tokens for unique scenarios. These include tokens for modified bases (methylated cytosine, pseudouridine), ambiguous positions in sequencing data (N for unknown nucleotide, B for not-A), gaps in multiple sequence alignments, and markers for sequence boundaries, domains, or functional regions. The design of these special tokens significantly impacts model behavior—for instance, whether to use a single [MASK] token or amino acid-specific masks affects what models learn during pretraining.

3.2. Autoregressive Language Modeling

Language models (LMs), as their name suggests, are probabilistic models of language. Modern state-of-the-art LMs are generative language models that model the *full probabilistic distribution* of language, meaning they can generate complete sequences of text. In this subsection, we focus on this class of LMs, while leaving those that model only partial (i.e., conditional probability, such as completing partial sequences or predicting middle words) distributions of language to the following subsections.

In generative language models, the more prevalent formulation breaks down the sequence probability for $\mathbf{x} = [x_1, x_2, \cdots, x_n]$ using the chain rule of probability in a left-to-right manner:

$$P(\mathbf{x}) = P(x_0) \prod_{i=1}^{n} P(x_i | x_0, \ldots, x_{i-1}).$$

When the model is represented by parameters θ, we also denote the probability as $P_\theta(\mathbf{x}) \approx P(\mathbf{x})$ to differentiate it from the natural (i.e., real-world) distribution $P(\mathbf{x})$.

Recurrent Neural Networks. The earliest attempts at autoregressive biological sequence modeling began with RNNs and LSTMs. The foundation of biological sequence modeling began with recurrent neural networks. These early architectures, including KEGRU [244] with k-mer embeddings, [183] with an adaptive LSTM framework, LSTM-CNN hybrids [184, 337], and the application of the attention mechanism on LSTM-CNN hybrid models [39], established the feasibility of deep learning for biological sequences. However, they faced significant limitations, such as vanishing gradient problems with long sequences, limited parallelization capabilities, and difficulty capturing very long-range dependencies crucial for biological function. Hybrid

architectures combining CNNs with BiLSTMs [119, 337] partially addressed these issues but remained limited to sequences under 1,000 tokens.

Transformers. The adaptation of transformer architectures fundamentally transformed biological language modeling. ProGen [194], a groundbreaking 1.2 billion parameter GPT-like model trained on 280 million protein sequences, demonstrated that autoregressive transformers could generate artificial lysozymes with catalytic efficiencies comparable to natural enzymes [195]. This success spawned a proliferation of models, including ProtGPT2 [78] with 738 million parameters generating proteins that fold properly according to AlphaFold predictions, and RITA [110], which systematically studied scaling effects up to 1.2 billion parameters. Going beyond single-protein modeling, PoET [275] proposes to model entire protein families as "sequences-of-sequences" using a tiered Transformer architecture. DNA language models followed suit with DNAGPT [336], trained on over 200 billion base pairs from all mammals, and regLM [149], which achieved 10-fold higher selection odds than gradient ascent methods for synthetic promoter design, with more works continuously emerging [128, 196]. Transformers also enabled massive scaling. The field grows towards massive models like xTrimoPGLM [38] with 100 billion parameters, demonstrating that biological language models follow similar scaling laws to NLP models. Transformers have also transformed other biological domains such as single-cell analysis and precision medicine through specialized architectures. scGPT [57], Geneformer [272], and scFoundation [102] are transformer-based foundation models trained with over 10 million cells and demonstrated generative capabilities for cell design and multi-omics integration. BioGPT [6], a GPT-based framework pre-trained on large-scale genomic and electronic health record datasets and fine-tuned for rare disease tasks, is designed to advance personalized genomic medicine and clinical settings. Recently, researchers also developed ensemble-based models such as PaSCient [175] and PULSAR [214] to represent multi-cellular-level interactions and further build foundation models for precision medicine from the perspective of analyzing bimolecular information.

State-Space Models and Sub-Quadratic Architectures. The quadratic complexity of attention limits the sequence length that can be handled by transformers to be insufficient for genomic-scale modeling. State-space models (SSMs) like Mamba [95] address this through recurrent formulations with linear complexity. HyenaDNA [206] leverages implicit convolutions with global receptive fields, processing sequences up to 1 million base pairs. The Evo family [207, 29] employs StripedHyena blocks combining attention with gated convolutions, achieving near-linear scaling for megabase sequences while maintaining generation quality.

These architectures make different trade-offs. Transformers excel at precise local patterns but struggle with very long sequences, while SSMs handle long contexts efficiently but may miss fine-grained dependencies. Hybrid approaches aim to strategically combine both, for example, using attention for local interactions and SSMs for global context.

3.3. Masked Bidirectional Language Modeling

Masked language models (MLMs) corrupt input sequences by masking random tokens, then train models to reconstruct the original sequence from bidirectional context. This paradigm, popularized by BERT [64], proves particularly powerful for biological sequences where functional constraints often depend on both upstream and downstream context.

The standard MLM objective randomly masks 15% of tokens and trains the model to predict them;

$$\mathcal{L} = -\sum_{i \in M} \log P_\theta(x_i | \mathbf{x}_{\setminus M}),$$

where M denotes masked positions and $\mathbf{x}_{\setminus M}$ represents the corrupted sequence. This bidirectional context proves crucial for biological understanding. For example, a protein's active site depends on global fold, enhancers influence distant promoters, and RNA structures form through long-range base pairing.

While standard BERT architectures work well, biological MLMs often require modifications. MSA Transformer [226] extends attention across multiple sequence alignments, allowing the model to leverage evolutionary information directly. The architecture uses tied row attention for sequences and column attention for positions, elegantly encoding coevolution patterns [302]. Biological MLMs explore sophisticated masking beyond random selection. ESM-1b [230] masks contiguous spans to encourage learning of structural motifs. ProteinBERT [27] employs annotation-aware masking, preferentially masking functional sites to focus learning on biologically important regions. DNABERT-2 [351] uses dynamic masking with different corruption rates during training, adapting to sequence complexity.

Masked language models have proven particularly effective for biological applications due to their bidirectional context understanding. The ESM series exemplifies this success trajectory. ESM-1b [230] (650M parameters) first demonstrated that atomic-level protein structure emerges in learned representations, spawning variants like ESM-1v [197] and ESM-IF1 [117], and ESM2 [166]. ESMFold, harnessesing the ESM-2 model, enabled creation of the ESM Metagenomic Atlas containing 617 million predicted structures, processing sequences 65x faster than AlphaFold2 [141]. Additional models include ProtTrans [72] training BERT and AlBERT encoders on millions of protein sequences to build a representation that can be fine-tuned for downstream tasks. ProteinBERT [27] pretrains on both protein sequences and gene ontology annotations to infuse the protein representations with information about the protein function. ProstT5 [108] encodes both protein sequences and protein structures as sequences of tokens to develop a bilingual model that can act as a foundation model, translate between sequence and structural information for tasks such as solving inverse folding problems, and detect remote homologs. MSA Transformer [227] accounts for a unique feature of protein sequences. Protein sequences are often understood in the context of the evolutionary history of the protein across species, commonly modeled through a multiple sequence alignment. MSA Transformer enables the attention mechanism to not only attend to amino acids within the same protein but also attend to corresponding amino acids across species, thus taking evolutionary context into explicit consideration.

For DNA sequences, DNABERT [133] pioneered BERT-style modeling with 110M parameters and k-mer tokenization. DNABERT-2 [351] improved upon this with multi-species training across 850 genomes, using byte-pair encoding instead of k-mers. The Nucleotide Transformer [58], with variants from 50M to 2.5B parameters, demonstrated that multi-species training on 850 genomes and 174B nucleotides consistently outperforms single-species models. For RNA, models like RNA-FM [40] train a foundation model to predict structural and functional tasks, and RNAErnie [294] employs hierarchical masking strategies, corrupting at base and motif levels. CodonBERT [156] represents a specialized masked language model that operates at the codon level rather than individual nucleotides. By treating each three-nucleotide codon as a single token, the model naturally captures the genetic code and synonymous codon usage patterns, enabling the model to perform well on tasks such as recombinant protein expression and mRNA degradation. UTR-LM [49] employs masked language modeling specifically designed for 5' and 3' UTRs, using masking strategies that respect regulatory element boundaries and RNA secondary structure constraints. 3UTR-BERT [323] specializes in 3' untranslated regions that regulate mRNA stability, localization, and degradation. The model employs a multi-scale masking strategy targeting microRNA binding sites, AU-rich elements, and polyadenylation

signals. By training on paired 3'UTR sequences and expression data, the model predicts mRNA half-life, subcellular localization, and the impact of 3'UTR mutations on post-transcriptional regulation.

Language models now span diverse biological applications. scBERT [318] handles over 16,000 genes using Performer attention mechanisms, achieving 95% accuracy on cell type annotation. Geneformer [272], pre-trained on 30 million cells (V1) scaling to 104 million (V2), introduced rank-based gene expression representation enabling zero-shot learning and in silico perturbation analysis. Clinical text integration through specialized models like ClinicalBERT [122] and BioClinicalBERT [8] for adverse drug reaction detection and patient-trial matching. Epigenetic modeling with EpiGePT [86] (71.3M parameters) predicting 8 types of epigenetic modifications. EpiBERT [132] learns from DNA sequence and DNA accessibility jointly and performs various tasks including sequence-level interpretation and fine-tuning-level expression prediction.

Pretrained protein language models can be leveraged for many downstream tasks. [152] trains a linear model for downstream tasks from pretrained representations and finds that model pretraining is well-aligned with structure-related tasks, but less aligned with (and in some cases, does not benefit) non-structural tasks. [237] states that supervised fine-tuning of the model for a downstream task produces better performance than simply training a downstream model on top of a frozen pretrained model. Pretrained protein language models are trained to predict a diverse array of tasks, with the development of multi-task benchmarks such as TAPE [225] and PEER [314].

Masked language models consistently outperform autoregressive models on certain biological tasks due to several key advantages. Bidirectional context crucial for understanding regulatory elements that can be located far from their targets, superior representation learning that captures structural and functional properties more effectively, and better fine-tuning performance across diverse downstream tasks. The ProtTrans suite [72] comprehensively demonstrated that bidirectional models achieve 76-84% Q3 secondary structure prediction accuracy compared to lower performance from unidirectional alternatives.

3.4. Advanced Training Techniques

Pretraining Strategies Beyond Standard MLM/AR. Several models combine generative objectives with contrastive learning. ESM-1v [197] uses a variant prediction objective, contrasting wild-type and mutant sequences to learn mutation effects. ProtST [315] aligns sequence and structure representations through contrastive learning, improving both modalities. Modern biological LMs increasingly combine multiple pretraining objectives. ESM3 [105] jointly models sequence, structure, and function through coordinated masking across modalities. Geneformer [272] combines expression prediction with cell type classification during pretraining.

3.5. Conditional Generation

Conditional language models generate sequences given specific constraints or properties, enabling targeted design for desired functions. This paradigm proves essential for practical applications where random generation is insufficient.

Conditioning Mechanisms. The simplest approach prepends condition tokens to the sequence. ProGen [194] concatenates taxonomic and functional tags (e.g., "[ORGANISM:E.coli][FUNCTION:hydrolase]") before the protein sequence. The model learns to associate these prefixes with sequence patterns, enabling controlled generation. More sophisticated models use separate encoders for conditions. PocketGen [344] encodes ligand molecules with a graph neural network, then uses cross-attention to condition protein se-

quence generation on ligand structure. This architectural separation allows pretrained encoders for different modalities. ProCALM [320] introduces lightweight adapters that modulate pretrained language models for specific conditions. This approach preserves the pretrained model's knowledge while enabling efficient conditioning on new properties without full fine-tuning. Some models guide generation through gradient signals from external predictors. These approaches compute gradients of desired properties with respect to generated sequences, steering generation toward optimal regions of sequence space.

Function-Conditioned Protein Language Models. Conditional language models have revolutionized targeted biological design. Protein language models can be conditioned to achieve a variety of functions. ProGen [194] and ProGen2 [208] (up to 6.4B parameters) pioneered function-guided generation using taxonomic and functional control tags, with ProGen2 achieving 52.4% sequence recovery compared to 32.9% for traditional methods. ZymCTRL [203] specialized in enzyme design, successfully generating carbonic anhydrases and lactate dehydrogenases with <40% sequence identity to natural proteins in zero-shot mode. ProCALM [320] introduced adapter-based conditioning, enabling generation based on enzyme function, taxonomy, and even natural language descriptions. In protein design, often only certain regions of a protein need to be modified. In such cases, the generated region is conditioned on the known regions in the protein. Many models trained with a masked language modeling objective can be adapted for this purpose by masking out the region to be designed while keeping the remaining sequence as context. For example, ESM3 enables both full protein generation based and protein generation based on a partial sequence or structure [105]. GDPL [334] leverages the ESM2 protein language model and folding trunk and structure model from ESMFold [166] to generate a protein scaffold conditioned on a given motif. Additional conditioning often includes a molecule to which the protein should bind. For example, PocketGen leverages both language and graphical modeling to product the sequence and structure of a protein region conditioned on both the desired binding ligand and the protein scaffold [344]. PepMLM generates peptide binders conditioned on a target protein sequence [42].

Structure-Conditioned Protein Language Models. ProteinMPNN [262] transformed structure-based protein design using graph neural networks with message-passing, achieving 52.4% sequence recovery on native backbones and rescuing previously failed designs across multiple protein classes. LigandMPNN [61] extended this to ligand-conditioned design, explicitly modeling non-protein atoms and achieving 63.3% sequence recovery near small molecules with over 100 experimentally validated designs demonstrating nanomolar-to-micromolar binding affinities. ProteinDPO [304] aligns a structure-conditioned language model to generate stable protein sequences by encouraging the model to prefer stabilizing over destabilizing variants given a protein backbone structure.

Function-Conditioned Nucleic Acid Models. While the field of language models for DNA and RNA is generally less developed than that of protein language models, recent work has begun to explore designing conditional DNA and RNA language models, particularly for shorter sequences. For example, LEONINE optimizes enhancer sequences conditioned on the cell type, gene expression profile, and promoter sequence [173]. TACO [324] adapts HyenaDNA [206] through reinforcement-learning based fine-tuning to optimize cis-regulatory elements (short DNA sequences such as promoters and enhancers regulating gene expression). CodonTransformer [73] encodes protein-DNA pairs across different species, and allows optimization of DNA sequence encoding a known amino acid sequence conditioned on the target species.

Multi-Property Optimization. Many factors determine the desirability of protein, DNA, and RNA sequences; thus, multi-property optimization is a new and critical direction in the field. The field has advanced to simultaneous optimization of multiple properties through reinforcement learning frameworks like ProteinRL [257] and the method discussed in [199]. ESM-3 additionally allows the generation of protein sequences, structure, and/or function, each conditioned on the others or parts thereof [105].

4. Diffusion and Flow Matching Models for Biological Generation

4.1. Overview

Diffusion and flow matching models have become the dominant paradigm for image generation and show strong potential for biological generation. Many biological objects (molecular conformations, protein backbones, sequences, complexes) are *high-dimensional* and *multi-modal*, and must satisfy strong structural constraints; diffusion/flow methods handle this by learning a gradual, stable transformation from an easy-to-sample base distribution p_1 (e.g., Gaussian noise) to the data distribution $p_0 = p_{\text{data}}$, while allowing rich conditioning and guidance at sampling time. (Continuous) score-based diffusion [253] specifies a forward noising SDE, learns the score $s_\theta(x,t) \approx \nabla \log p_t(x)$, and samples either with the reverse-time SDE (stochastic) or the probability-flow ODE (deterministic), which yields identical marginals. (Continuous) flow matching [169] instead posits an ODE and directly regresses the time-dependent velocity $u_t(x)$ that transports probability mass according to the continuity equation under a chosen probability path. And when this path matches the Gaussian conditional path commonly used in diffusion [113], the learned velocity corresponds to the probability-flow vector field derived from the score, and the resulting training objectives coincide up to constants, explaining why architectures and conditioning tricks often transfer between diffusion and flow matching. In practice, flow matching often enables deterministic generation with fewer function evaluations, while diffusion naturally supports stochastic sampling and annealed noise schedules that can improve diversity and robustness. In their discrete counterparts, the state space is finite and lacks a standard Euclidean gradient with respect to tokens, so one [11, 32] typically learns the time-reversal process by approximating the clean-state posterior $p_\theta(x_0|x_t)$. The discrete-time reverse transition kernel can then be obtained by marginalizing $p(x_{t-1}|x_t) = \sum_{x_0} q(x_{t-1}|x_t,x_0)p_\theta(x_0|x_t)$, where $q(x_{t-1}|x_t,x_0)$ is completely determined by the chosen forward corruption kernel via the Bayes' rule conditioned on x_0. For the continuous-time counterpart, the training target remains the same, while the sampling uses either the velocity field [88] or, equivalently, the rate matrix [33] by marginalization using $p_\theta(x_0|x_t)$. Figure 3 provides a unified view of discrete-time discrete and continuous diffusion processes for biomolecular generation, contrasting their forward transitions and corresponding learned reverse dynamics for sampling.

4.2. Modeling Considerations

Successfully transferring diffusion models and flow matching architectures to biological data requires reconciling the divergence of biological data types from the data type they were developed for. Unlike images, biological data combine continuous representations with complex structural constraints. Furthermore, the same data (e.g. proteins) can have multiple views (e.g. 1D sequence, 2D contact map, or 3D geometry) which require drastically different modeling assumptions and architectures. Therefore, we frame diffusion/flow methods along two axes that map cleanly onto these modalities. First, **continuous diffusion / flow matching** operates on $\mathbb{R}^d$ or geometry-aware manifolds. Second, **discrete diffusion / flow matching** operates on categorical spaces (tokens, residues, motifs) natural for protein/DNA/RNA sequence generation. To introduce these methods, we outline the essential concepts underlying diffusion and flow matching in these methods as below, especially for protein modeling.

Equivariance and Invariance. Let G be a symmetry group acting on a data space X. A map $f : X \to Y$ is *equivariant* to G if $f(g \cdot x) = g \cdot f(x)$ for all $g \in G$; it is *invariant* if $f(g \cdot x) = f(x)$. For proteins, the relevant symmetries are rigid motions E(3) and often only SE(3) due to chirality, since some E(3)-equivariant GNNs are reflection-aware and can encourage non-physical, left-handed geometries; SE(3)-aware attention (IPA) avoids that by being sensitive to orientation [328, 9]. Enforcing these symmetries improves data efficiency, prevents spurious pose memorization, and encodes physically correct inductive bias.

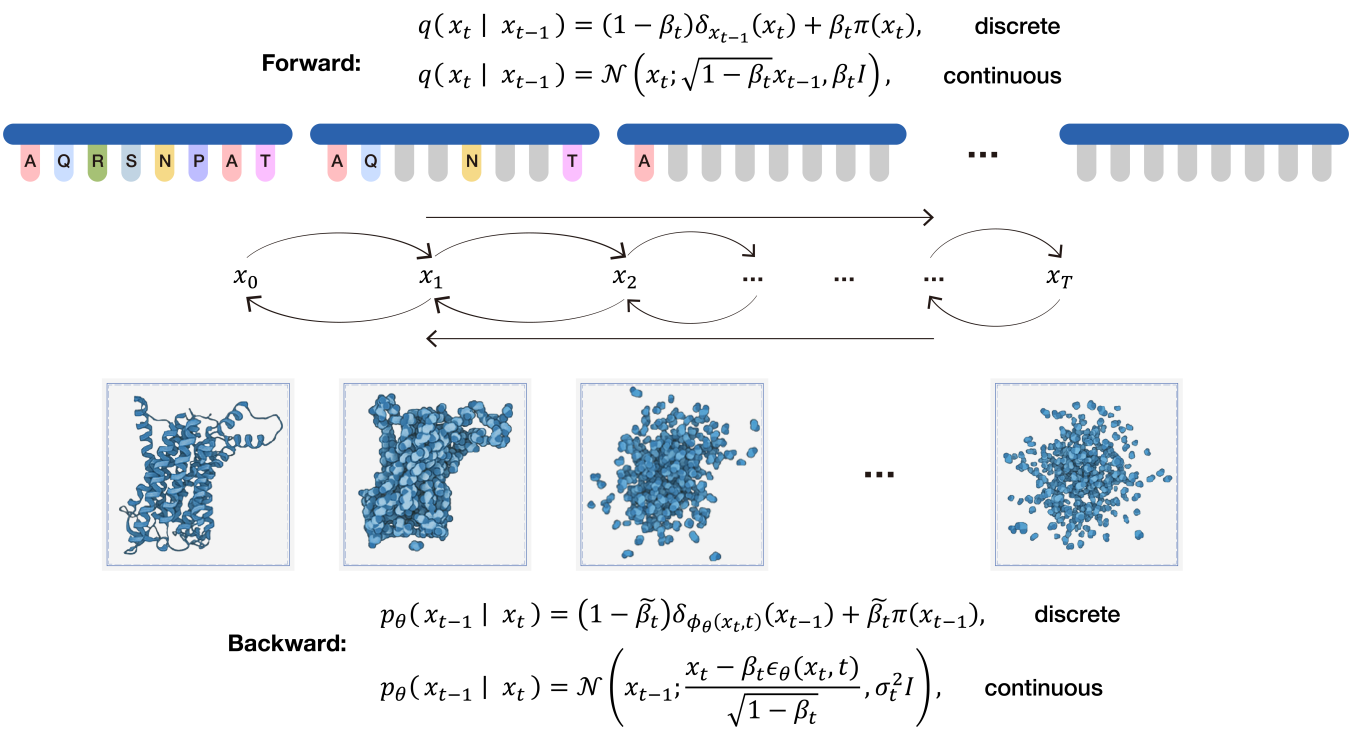

Figure 3: Overview of the forward–reverse diffusion pipeline for biological generation, showing the forward noising process and the learned reverse denoising process, in both discrete (token-level) and continuous (geometry-level) formulations.

How Symmetry Enters Diffusion. In coordinate spaces, the forward corruption is typically *isotropic* Gaussian noise, whose density is rotation-symmetric; combined with centering (subtracting the center of mass), it yields an SE(3)-*invariant* noise distribution and therefore an SE(3)-*equivariant* diffusion path for the structure. The denoiser must then be SE(3)-equivariant to recover an equivariant time-reversal process [328, 48, 274, 332].

Geometric Parameterization. A standard geometric representation stores all-atom or C_α-backbone Cartesian coordinates $X \in \mathbb{R}^{3N}$ for amino acids. Alternatively, one can generate directly over backbone and side-chain dihedral torsion angles. Each torsion is treated as a coordinate on a torus $\mathbb{T}$ and is noised by rescaled Brownian motion whose perturbation kernel is proportional to wrapped Gaussian noise [137]. This parameterization removes rigid-body degrees of freedom by construction and is therefore SE(3)-invariant [306]. Its main advantage is pose-free modeling, while its main challenge is that kinematic and steric constraints induce long-range, non-local couplings between angles. Another common choice is to represent each residue by a rigid frame $\mathrm{SE}(3)$ [328, 332], optionally modeled with side-chain torsions. In protein frame spaces $\mathrm{SE}(3)^N = (\mathbb{R}^3 \rtimes \mathrm{SO}(3))^N$, translations are noised by Brownian motion or OU process in $\mathbb{R}^3$ and rotations by Brownian motion on $\mathrm{SO}(3)$ (e.g., $\mathrm{IG}_{\mathrm{SO}(3)}$ heat kernel), with matched time-reversal SDE/ODE; training often uses frame-aligned losses (e.g., FAPE) that are invariant to rigid motion.

Categorical Spaces and Discrete Diffusion. In sequence settings, the state space is a finite alphabet $\mathcal{A}$ (e.g., amino acids plus `BOS/EOS/MASK/PAD` for proteins; nucleotides for DNA/RNA) or a higher-level symbolic space of *motifs* or *subwords* (e.g., k-mers, BPE/WordPiece tokens, structural alphabets, VQ-VAE codes). Discrete diffusion specifies a corruption kernel $q_t(x_t \mid x_{t-1})$ on $\mathcal{A}^L$ that gradually forgets the original tokens. Common choices are (i) *multinomial diffusion* [115] that keeps a token with probability

$1-\beta_t$ and otherwise resamples from a stationary distribution π (uniform or empirical prior), (ii) *masked diffusion* [9] that replaces tokens by a special `[MASK]` with increasing rate, and (iii) *biologically informed* kernels that mix toward π through substitution matrices (e.g., BLOSUM-derived transition probabilities) so corruption remains biologically plausible. Length can be handled by fixing L, or by adding *edit operations* (including insertion and deletion) via a birth–death process over sequence length. The reverse model learns $p_\theta(x_{t-1} \mid x_t)$ [11, 157] or a time-dependent rate matrix [32], trained by approximating the posterior $p_\theta(x_0|x_t)$, with per-step reweighting to match the noise schedule. Classifier-free guidance is implemented by toggling conditioning features including functions, domains, active-site residues, or MSA columns, while classifier guidance reweights logits with an external property predictor. Other key considerations include (1) *Alphabet design*: coarser tokens (motifs or subwords) shorten sequences and ease long-range coupling but risk losing fine constraints; fine tokens preserve fidelity but require stronger inductive bias. (2) *Stationary target* π: using background amino-acid frequencies or column-wise MSA marginals stabilizes training and sampling relative to uniform. (3) *Constraints and conditioning*: hard constraints are enforced by masking illegal transitions, while soft constraints use potential energies (e.g., under structure predictors, folding energies, or function classifiers) during sampling. (4) *Non-local dependence*: biological syntax is long-range; architectures (bidirectional transformers/GNNs over sequence graphs, MSA-aware attention) must capture couplings beyond Markov neighborhoods. (5) *Sampling cost*: many diffusion steps can be amortized by schedule truncation, knowledge-distillation/consistency training, or flow matching formulations that reduce steps while retaining parallel, non-autoregressive generation.

4.3. Continuous Diffusion and Flow Matching for Structures

The development of continuous diffusion and flow-matching models for biological structure generation is arguably the field's most mature achievement.

Small Molecules. For small molecules, the central challenge is modeling *multi-modal conformer ensembles* while remaining invariant/equivariant to rigid motions. Geometry GNNs [71] are natural and commonly used backbones for this setting. In 3D coordinate space, GeoDiff [316] constructs diffusion dynamics with SE(3)-equivariant Markov kernels to obtain SE(3)-invariant conformer distributions conditioned on the molecular graph, while EDM [116] uses an E(3)-equivariant diffusion network to jointly denoise atomic positions and atom types. In contrast, MCF [298] diffuses directly on 3D positions via a conformer-field parameterization, explicitly aiming to reduce reliance on hand-crafted inductive biases such as rotational equivariance. For finer control over the true continuous degrees of freedom in molecular conformer generation, torsional diffusion [137] performs diffusion on the hypertorus of rotatable bonds, improving conformer ensemble quality by operating in internal coordinates. On the flow side, equivariant flow matching provides simulation-free objectives for equivariant CNFs [144], and has been instantiated for molecular generation via deterministic ODE transports (e.g., ET-Flow [103] and EquiFM [254]), often reducing sampling steps and aligning well with symmetry priors. Finally, ShEPhERD [4] extends beyond geometry by jointly diffusing 3D molecular graphs and explicit interaction-profile representations such as shape, electrostatic potential, and directional pharmacophores for *bioisosteric* drug design.

Proteins. For proteins, the backbone generation problem highlights why diffusion and flow models are compelling, as they can model *distributions* over valid folds instead of a single prediction while supporting constraints with motifs, binders, and symmetry through conditional generation. RFdiffusion [303] fine-tunes the RoseTTAFold structure-prediction backbone for denoising and has been validated across binders, symmetric assemblies, and motif scaffolding; combined with ProteinMPNN [60] and AlphaFold [141], it underpins a common *de novo* protein design workflow. Chroma [130] models the joint space of protein structures and sequences with programmable generation, including generating large complexes. FrameDiff [328]

formalizes diffusion on per-residue rigid frames to generate designable monomers without relying on a pretrained predictor, while latent diffusion approaches compress structures into lower-resolution codes and run generation efficiently in latent space [81]. In parallel, flow-matching variants learn ODE vector fields on the $\mathrm{SE}(3)^N$ manifolds to reduce the number of sampling steps, including FoldFlow [26] and its sequence-conditioned extension FoldFlow-2 [126], as well as FrameFlow [329].

Beyond backbones, continuous models increasingly target *all-atom and conditional* generation. RFdiffusion All-Atom [146] expands the framework to generate biomolecular interactions in the presence of ligands and other contexts. For conformational ensembles, AlphaFlow and ESMFlow fine-tune AlphaFold and ESMFold under flow matching to sample sequence-conditioned landscapes with MD-like statistics [138]. For local editing, FrameDiPT [335] performs structure inpainting via SE(3)-equivariant diffusion conditioned on surrounding context. For side chains, DiffPack [342] diffuses on the χ-angle (torsion-angle) hypertorus, while FlowPacker [150] adopts torsional flow matching for efficient, constraint-respecting packing.

Biomolecular Complexes. On the biomolecular complexes interaction generation, TargetDiff [97] uses pocket-conditioned joint diffusion of coordinates and atom types to generate the ligand, DiffDock [56] defines diffusion over ligand translation, rotation, and torsion manifold to generate bound poses and report calibrated confidence, and DiffDock-PP [143] extends this to rigid protein-protein docking. In structure-based drug design, pocket-conditioned diffusion models generate ligands directly in 3D binding sites, e.g., DiffSBDD [238]; DiffLinker [127] targets fragment linking via E(3)-equivariant pocket-conditioning diffusion; NeuralPLexer3 [222] is a physics-inspired flow-based generative model for complex structure prediction; FlowDock [200] uses conditional flow matching to map apo to holo complexes while outputting confidence and affinity estimates. More broadly, diffusion-style architectures have also become central in large-scale complex prediction (e.g., AlphaFold 3 [2]), underscoring the generalization ability of these generative formulations.

Nucleic Acids. Compared to proteins, nucleic-acid generative modeling is earlier-stage due to greater conformational flexibility and scarcer 3D training data, but progressing rapidly. RiboDiffusion [120] models RNA inverse folding by learning the conditional distribution of RNA sequences given 3D backbone structures using a geometric GNN module coupled to a Transformer sequence module. RNAFlow [210] applies flow matching to protein-conditioned RNA sequence and structure co-design by combining an RNA inverse folding model with a frozen RosettaFold2NA-style predictor [15] inside the loop. RNA-FrameFlow [10] extends FrameFlow [329] to RNA backbone structure generation by creating a local frame for each nucleotide, allowing capturing larger RNA nucleotides while still supporting the modeling of all backbone atoms implicitly.

4.4. Discrete Diffusion and Flow Matching for Sequences

Discrete diffusion and discrete flow models have emerged as a powerful alternative to autoregressive generators for biological sequences and molecular graphs, since many design problems are naturally framed as *constrained completion* rather than left-to-right decoding, while diffusion and flow matching formulations supporting bidirectional context, parallel updates, and plug-in guidance for optimization objectives during sampling. Most approaches start from either D3PMs [11] or CTMC formulations [32], and parameterize reverse dynamics via clean-state posterior prediction.

Small Molecules. Discrete diffusion and flow methods for small molecules span *(i) graph generators* that model topology and chemistry in a permutation-invariant way and *(ii) token generators* that leverage sequence models often with SMILES/SAFE notations and conditional guidance. In graph generation, DiGress [287] performs discrete denoising diffusion directly on molecular graphs by noising and denoising categorical node and edge types with a graph Transformer, enabling conditional generation and scaling to large drug-like

corpora; DiffMS [24] develops a formula-constrained discrete diffusion model that generates molecular graphs conditioned on mass spectra for structure elucidation. Complementary score-based approaches such as GDSS [139] model coupled node features and adjacency through a continuous-time diffusion system, explicitly capturing dependencies between topology and attributes. Beyond pure discrete topology, hybrid 3D generators jointly model categorical atom types and continuous coordinates. EquiFM [254] and FlowMol [70] extend the flow matching to mixed continuous–categorical spaces by transporting the probability of both atom types and 3D coordinates, yielding efficient, deterministic sampling and strong 3D generative performance. For string-based generation, TGM-DLM [91] introduces a diffusion language model for text-guided SMILES generation, while GenMol [151] brings masked discrete diffusion to specialized SAFE molecular sequences and introduces fragment re-masking and task-specific guidance to handle a wide range of drug-discovery tasks in a single model.

Proteins. Discrete diffusion and flow methods for protein sequence design model a categorical state space with denoising directly over amino acid tokens, enabling conditional guided generation and structure-conditioned inverse folding. Foundation-scale diffusion language models such as DPLM [295] and EvoDiff [7] cast protein language modeling as discrete diffusion and support controllable generation in sequence space. For family- or motif-conditioned generation, CPDiffusion [348] conditions on secondary structure and conserved residues, while GraDe-IF [327] performs backbone-conditioned diffusion for inverse folding with BLOSUM-derived transitions. Complementing these, NOS [94] enables multi-objective sequence design by applying guidance through gradients in the shared hidden states of the denoising network, and LaMBO-2 extends this with Bayesian optimization under uncertainty and edit constraints. Finally, sequence–structure co-design systems can couple sequence diffusion with a structure predictor (e.g., ESMFold) used as a differentiable guide or training scaffold [9], often combining sequence cross-entropy with a frame-aligned coordinate loss (e.g., FAPE). DFM [33] provides a protein co-design model with the flexibility of multimodal protein generation based on CTMC [32], jointly generating protein structure and sequence. ProteinGenerator [170] is a sequence-space diffusion model based on RoseTTAFold that simultaneously generates sequences and structures with iterative refinement. For multi-property peptide optimization, PepTune [267] introduces Monte Carlo Tree Guidance (MCTG) to perform inference-time scaling toward Pareto-efficient trade-offs across multiple therapeutic objectives.

Nucleic Acids. DNA and RNA sequence generation remains less mature than protein design but is accelerating. Regulatory DNA diffusion models, including D3 [234], DNA-diffusion [59], and TR2-D2 [268] using Monte Carlo Tree Search (MCTS) for multi-objective discrete diffusion fine-tuning, can conditionally generate sequences with targeted functional behavior, while RNAdiffusion [121] performs controllable latent diffusion for RNA sequences with reward-guided sampling. Complementing discrete diffusion, discrete flow matching emphasizes stable probability transport on categorical simplices. Dirichlet Flow Matching [255] constructs mixture-of-Dirichlet probability paths for DNA sequence design, and Fisher-Flow [62] follows Fisher-Rao geometry to define numerically stable flows with strong promoter/enhancer performance. MOG-DFM [44] further provides a general framework to steer pretrained discrete flow matching generators toward Pareto-efficient trade-offs across multiple objectives.

5. Toward Unifying Models

The use of generative AI within computational biology has long shown a convergence where multiple generative modeling paradigms are unified into powerful hybrid architectures. This synthesis represents the need to build sophisticated frameworks that combine the strengths of transformers, convolutional networks, and other architectures in order to tackle complex and unique biological challenges. Recently, the boundaries between

language models and diffusion models are increasingly blurred as studies discover that these seemingly distinct paradigms can be unified and may share theoretical connections. This convergence has led to a new generation of hybrid models that combine the sequential reasoning capabilities of language models with the iterative refinement and uncertainty modeling of diffusion processes. These unified architectures represent the cutting edge of generative modeling for biology, offering unprecedented flexibility and performance across diverse biological tasks.

5.1. Hybrid Architectures: Domain-Specific Integration

The necessity for unified architectures in biological AI emerged from fundamental limitations of purely data-driven approaches when confronted with biological complexity. Unlike natural language or images, biological sequences encode hierarchical information across multiple scales, from local motifs to long-range interactions spanning thousands of positions, while simultaneously satisfying physical constraints that pure sequence models struggle to capture.

Successful unified models employ sophisticated architectural innovations that go far beyond simple ensemble approaches. AlphaFold2's [141] revolutionary accuracy arose not from a single innovation but from its sophisticated unification of multiple computational streams. The Evoformer architecture processes evolutionary information through a Multi-Sequence Alignment (MSA) representation while simultaneously maintaining a pair representation encoding spatial relationships. These two information streams communicate bidirectionally through 48 blocks, where MSA row attention incorporates pair biases and pair representation updates integrate information from the MSA column attention. This creates an iterative refinement process allowing evolutionary patterns to inform structural hypotheses and vice versa. The structure module further unifies this representation with an equivariant transformer operating directly on 3D coordinates, demonstrating that biological problems require simultaneous reasoning across sequence, evolutionary, and geometric spaces. RoseTTAFold [14] pioneered a complementary three-track architecture that processes 1D sequence features, 2D distance maps, and 3D coordinate information in parallel tracks with continuous information exchange. The key insight was that biological information naturally exists at multiple levels of abstraction, where sequence patterns capture evolutionary constraints, distance maps encode contact preferences, and 3D coordinates must satisfy geometric consistency. By allowing these representations to evolve jointly through alternating updates, the model achieves simultaneous optimization across biological constraints that no single-track architecture could.

Another line of work shows how CNN/RNN/transformer/SSM hybrids enable genomic-scale modeling, particularly for DNA sequences where extreme lengths, often hundreds of thousands to millions of bases, demand architectures that combine local and long-range processing. Early hybrid models emerged because biological regulation intrinsically involves both local cis-regulatory motifs and long-range chromatin interactions. Enformer [12] demonstrated this need by combining convolutional layers for local pattern detection with transformer blocks for long-range interactions, processing sequences up to 100–200 kb at 128 bp resolution to predict gene expression, histone modifications, TF binding, and DNA accessibility across many experimental contexts. Borzoi [167] extended Enformer through multi-modal integration of CAGE, RNA-seq, DNase-seq, and ChIP-seq data to achieve cell-type and tissue-specific gene expression prediction, while Sei [41] used regulatory sequence modeling to interpret human genetics and disease. A rich ecosystem of task-specific convolutional and hybrid models has also emerged, including SpliceAI [131], Pangolin [331], and DeltaSplice [313] for splice-site prediction; scTCA [330] for single-cell DNA sequencing; DeepFormer [325] for DNA function prediction; TransCrispr [288] for guide-RNA design; and Orca for large-scale genome organization including chromatin compartments, interchromosomal interactions, and structural variant analysis. Beyond transformer-based hybrids, new architectures have pushed long-context

modeling even further by replacing attention with sub-quadratic operators rooted in state-space models. HyenaDNA [206] achieves a receptive field of up to one million base pairs using learned convolution filters and position-dependent modulations that maintain computational tractability while supporting both local motif recognition and global context integration. The Nucleotide Transformer [58] family further demonstrated that optimal performance depends on architectural adaptation: transformers model coding sequences with clear syntactic structure, while convolutional preprocessing remains essential for regulatory regions where transcription factor binding motifs must be detected. The Evo1 and Evo2 models [207, 29] exemplify the power of this design space through the StripedHyena architecture [147], which combines grouped-query attention, gated convolutions, and state-space layers to process sequences of 131,072 to one million tokens with near-linear scaling. These models train 30–100% faster than pure transformers while maintaining superior performance, highlighting how state-space layers enable efficient transitions between convolutional and recurrent representations for megabase-scale modeling. Generative models have also begun to adopt these long-context hybrids, with Evo2 enabling one-million-base-pair generative modeling across many species. AlphaGenome [13] further extends the convolutional-hybrid paradigm to integrate DNA, epigenomic, and RNA modalities over one-megabase context windows to predict gene expression, splicing, chromatin state, and chromatin contacts at base-pair resolution.

Shared embedding architectures represent another key in unified modeling, widely applied in works discussed in this sections and beyond. The UnitedNet [270] framework employs modality-specific encoders that map DNA, RNA, and protein sequences to unified latent spaces of identical dimensionality. Adaptive fusion mechanisms use trainable weights with contrastive loss functions that align latent codes from different modalities, ensuring biological coherence across molecular types. UnitedNet's alternating training scheme alternates between group identification and cross-modal prediction, with combined loss functions that balance competing objectives through adaptive weighting. This multi-objective training creates models that can seamlessly switch between generation modes, specifically one-shot prediction when confidence is high, iterative refinement when uncertainty demands it, and guided sampling when specific constraints must be satisfied. The ICAN [148] architecture demonstrates how drug SMILES serve as queries while protein sequences provide keys and values, generating context matrices that capture drug-protein sub-sequence interactions. Multi-head cross-attention extends this concept with separate attention heads for different modality pairs (DNA-RNA, RNA-protein, DNA-protein), enabling task-specific attention masks that regulate modality interactions. LucaOne [106] demonstrates unified representational spaces across DNA, RNA, and proteins from 169,861 species, achieving superior performance in many downstream tasks using few-shot learning, despite no explicit DNA-protein pairing during training. CD-GPT [353] introduces a 1B parameter unified architecture with shared multi-molecule vocabulary across DNA, RNA, and proteins. Its three-stage pretraining strategy progresses from mono-sequence modeling to central dogma pairs to protein structure integration, enabling comprehensive understanding of molecular relationships. Life-Code [177] unifies DNA, RNA, and protein and maps all sequences to nucleotides, enabling multi-omics integration with a symmetric convolution backbone to capture long-range genomic context and distilling knowledge from large protein LMs to keep training efficient.

These hybrid models established a crucial principle that biological AI requires not just powerful general architectures but thoughtful integration of domain-specific inductive biases. The MSA processing in AlphaFold2 encodes evolutionary constraints impossible to learn from single sequences alone. The $SE(3)$-equivariant layers in structure modules enforce physical symmetries that would require astronomical amounts of data to learn from scratch. The multi-scale processing in genomic models reflects the hierarchical organization of regulatory logic. These designs demonstrate that unification in biological AI means not homogenization but principled integration of complementary computational strategies, each suited to different aspects of biological information.

5.2. Unified Architecture Trend: Diffusion Meets Language Modeling

One of the most profound shifts in biological generative modeling is the growing recognition that autoregressive language models and diffusion models, once seen as fundamentally different, are complementary parts of a broader framework. This convergence is clearest in architectures that combine sequential generation with iterative refinement.

AlphaFold3 [2] reimagined structure prediction by replacing AlphaFold2's sequential structure module with a diffusion-based generator that directly produces atomic coordinates. The model parameterizes structure generation as a reverse diffusion process $p_\theta(x_{t-1}|x_t, s) = \mathcal{N}(x_{t-1}; \mu_\theta(x_t, t, s), \sigma_t^2 I)$, where s represents sequence and evolutionary conditioning. This shift enables new capabilities, where the model can handle uncertainty through stochastic sampling, generate diverse conformational ensembles by varying the noise seed, and jointly model all molecular types (proteins, nucleic acids, ligands) within a unified coordinate space. The diffusion framework provides a principled way to incorporate physical constraints through the learned score function $\nabla_x \log p(x|s)$, which implicitly encodes the energy landscape of molecular conformations. Coincidentally, ESM3 [105] extends the masked modeling of the previous ESM series with a diffusion iterative learning, treating sequence, structure, and function as three views of the same underlying biological reality. The model employs a sophisticated tokenization scheme where structures are discretized through vector quantization of local geometric patterns, creating a vocabulary $\mathcal{V}_{\text{struct}}$ that complements the amino acid vocabulary $\mathcal{V}_{\text{seq}}$ and functional annotation vocabulary $\mathcal{V}_{\text{func}}$. During generation, the model performs masked diffusion across all three modalities $p(x^{\text{seq}}, x^{\text{struct}}, x^{\text{func}}) = \prod_t p(x_t|x^{\text{seq}}_{<t}, x^{\text{struct}}_{<t}, x^{\text{func}}_{<t})$. This allows the model to condition any modality on any combination of the others, generating sequences given structure, structures given function, or complete molecules given partial constraints across all three domains.

The D3PM framework [11] with absorbing states (MASK tokens) creates a direct mathematical connection to masked language modeling, with the transition matrix designed to mimic masking behavior. The DPLM [295] series explicitly bridges discrete diffusion and language modeling by showing that masked language models can be viewed as a special case of discrete diffusion with an absorbing state. The forward process corrupts sequences by replacing tokens with a special [MASK] token according to a schedule $q(x_t|x_0) = \text{Cat}(x_t; \mathbf{Q}_t x_0)$, where $\mathbf{Q}_t$ is a transition matrix that gradually increases masking probability. The reverse process learns to predict the original tokens $p_\theta(x_0|x_t) = \prod_i \text{Cat}(x_{0,i}; f_\theta(x_t)_i)$. This formulation unifies BERT-style masked language modeling (single-step unmasking) with iterative diffusion (progressive unmasking over many steps), enabling controllable generation quality through the number of denoising steps. EvoDiff [7] demonstrates another unification strategy through its OADM (Order-Agnostic Autoregressive Diffusion Model) framework, which combines the benefits of autoregressive ordering with diffusion's flexibility. The model maintains a partial ordering over sequence positions but allows this ordering to vary during training. During generation, the model can dynamically choose which positions to generate based on confidence scores, effectively performing adaptive computation that focuses refinement where uncertainty is highest. All-atom Diffusion Transformer (ADiT) [140] exemplifies diffusion-transformer integration through a two-stage architecture where autoencoders map unified all-atom representations to shared latent spaces, followed by diffusion models generating new latent embeddings for molecular sampling.

Unified training objectives combine the benefits of different paradigms. Training requires sophisticated multi-task learning frameworks. Models increasingly use composite losses, e.g., $\mathcal{L} = \lambda_1 \mathcal{L}_{\text{MLM}} + \lambda_2 \mathcal{L}_{\text{diff}} + \lambda_3 \mathcal{L}_{\text{struct}}$, where masked language modeling provides strong sequence priors, diffusion enables iterative refinement, and structure prediction ensures biological validity.

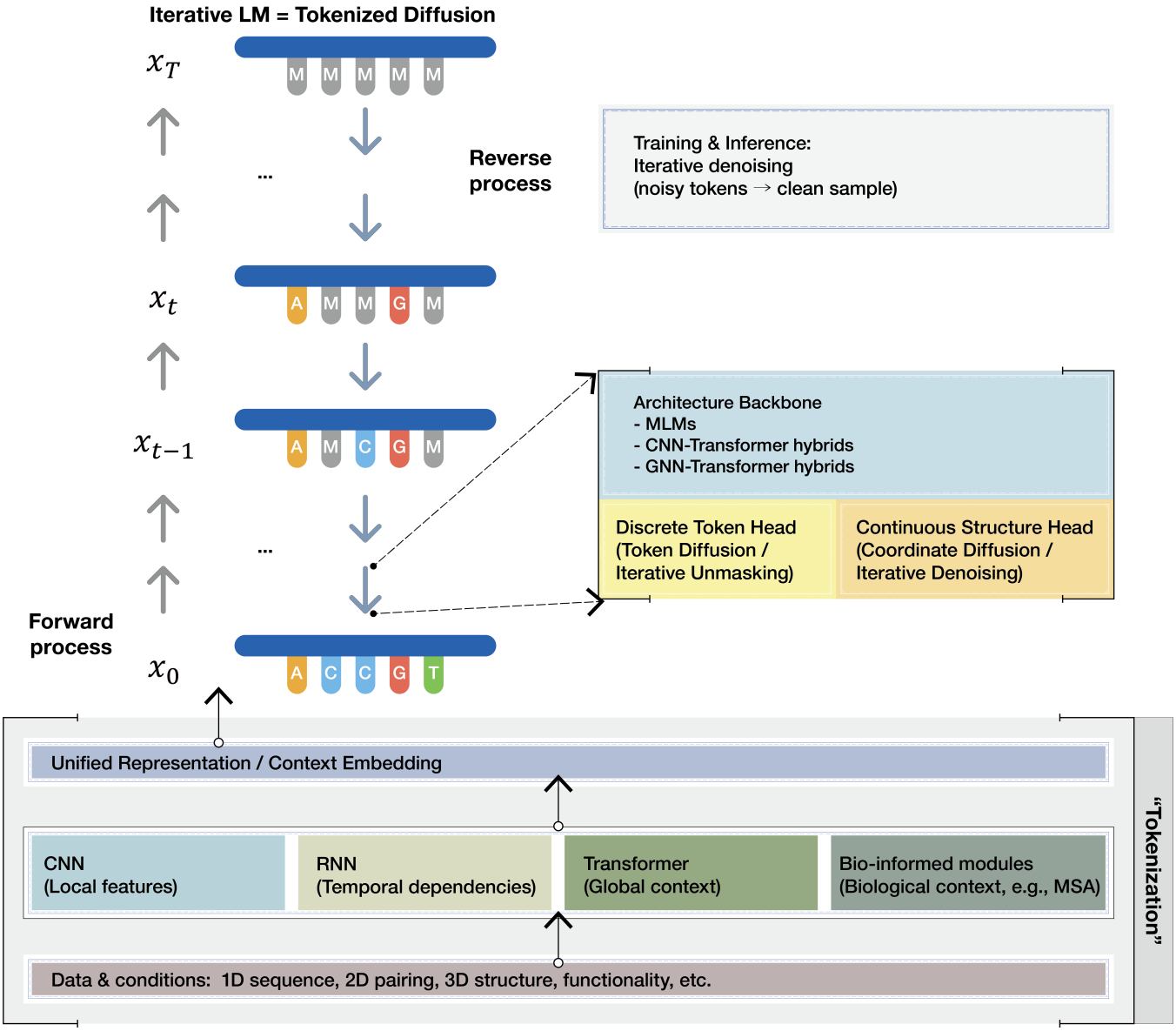

Figure 4: A unified generative-model view integrating classic architectures, language models, and diffusion models. This illustrates the perspective that iterative language modeling can be viewed as tokenized diffusion, where training and inference correspond to iterative denoising, supported by shared backbones and modality-specific (discrete/continuous) output heads.

5.3. Theoretical Analysis of Unified Frameworks

The convergence of autoregressive and iterative generation paradigms in biology can be understood through a unified mathematical framework that reveals their deep connections and complementary advantages. As presented in Figure 4, we discuss a unified generative-model perspective for biological sequence–structure generation, connecting classic architectures, language models, and diffusion models by interpreting iterative token prediction (unmasking) as a tokenized diffusion-style denoising process with discrete-token and continuous-structure heads. The theoretical foundation in this subsection discusses why biological domains

particularly benefit from hybrid approaches that combine sequential and iterative generation.

Consider a biological sequence or structure $\mathbf{x} \in \mathcal{X}$ that we aim to generate from some conditioning information c (which might be empty for unconditional generation). Both autoregressive and diffusion models can be viewed as defining a series of intermediate distributions $p_t(\mathbf{x}^{(t)}|c)$ that bridge between a simple initial (non-informative) distribution p_0 and the target data distribution p_T, where $\mathbf{x}^{(t)}$ denotes the sequence state at time step t.

Autoregressive LM as Sequential Diffusion. For autoregressive models, the intermediate states correspond to partial sequences $\mathbf{x}^{(t)} = (x_1, \ldots, x_t, \varnothing_{t+1}, \ldots, \varnothing_L)$ where x_i denotes the token at position i and $\varnothing$ represents undefined positions. We can reformulate this as a diffusion process with a specific corruption kernel $q(\mathbf{x}^{(t)}|\mathbf{x}^{(t+1)}) = \delta(x_{1:t} = x_{1:t}) \cdot \mathbb{1}[x_{t+1} = \varnothing]$, which deterministically masks position $t+1$ while preserving all previous positions. The reverse process then becomes $p_\theta(\mathbf{x}^{(t+1)}|\mathbf{x}^{(t)}, c) = \delta(x_{1:t}) \cdot p_\theta(x_{t+1}|x_{1:t}, c)$, recovering the standard autoregressive factorization.

Diffusion as Order-Agnostic Autoregression. Conversely, diffusion models can be expressed through an autoregressive lens by marginalizing over all possible generation orders. Define a random permutation $\sigma \in S_L$ where L is the sequence length, and generate tokens according to

$$p(\mathbf{x}|c) = \mathbb{E}_{\sigma \sim p(\sigma)} \left[\prod_{i=1}^{L} p(x_{\sigma(i)}|x_{\sigma(1)}, \ldots, x_{\sigma(i-1)}, c) \right].$$

When $p(\sigma)$ is uniform over permutations, this recovers the diffusion objective, as the model must learn to predict any token given any subset of other tokens, which is precisely what the denoising objective accomplishes.

Discrete Diffusion as Iterative Masked Language Modeling. The connection becomes particularly clear for discrete diffusion with absorbing states. Consider a forward process that gradually masks tokens $q(\mathbf{x}^{(t)}|\mathbf{x}^{(t-1)}, \mathbf{x}^{(0)}) = \prod_{i=1}^{L} q(x_i^{(t)}|x_i^{(t-1)}, x_i^{(0)})$. This creates a corruption process identical to BERT-style masking, but applied progressively over time steps. The reverse process learns to unmask tokens $p_\theta(\mathbf{x}^{(t-1)}|\mathbf{x}^{(t)}) = \prod_i p_\theta(x_i^{(t-1)}|\mathbf{x}^{(t)})$, where the model predicts original tokens for masked positions. The training objective becomes

$$\mathcal{L} = \sum_{t=1}^{T} \mathbb{E}_{q(\mathbf{x}^{(t)}|\mathbf{x}^{(0)})} \left[\sum_{i: x_i^{(t)} = \text{[MASK]}} -\log p_\theta(x_i^{(0)}|\mathbf{x}^{(t)}) \right].$$

This is precisely the masked language modeling objective, but summed over multiple masking ratios. BERT corresponds to $T = 1$ with a single masking step, while discrete diffusion extends this to multiple iterations, enabling progressive refinement.

The observation that masked diffusion models are time-agnostic masked models establishes a theoretical bridge where the continuous-time variational objective of masked diffusion reduces to weighted mixtures of classical masked language modeling losses. This insight also explains why pre-trained BERT-style models serve as excellent initialization for discrete diffusion and why approaches like DiffusionBERT [107] successfully combine both paradigms.

Unified Variational Framework. Both paradigms optimize the same evidence lower bound (ELBO) with different parameterizations of the approximate posterior. Let the joint distribution be defined as

$p(\mathbf{x}^{(0)}, \mathbf{x}^{(1:T)}) = p(\mathbf{x}^{(0)}) \prod_{t=1}^{T} q(\mathbf{x}^{(t)}|\mathbf{x}^{(t-1)})$. The marginal likelihood can be lower bounded with

$$\log p(\mathbf{x}^{(T)}) \geq \mathbb{E}_{q(\mathbf{x}^{(0:T-1)}|\mathbf{x}^{(T)})} \left[\log \frac{p_\theta(\mathbf{x}^{(T)}|\mathbf{x}^{(0)}) \prod_{t=1}^{T} q(\mathbf{x}^{(t-1)}|\mathbf{x}^{(t)}, \mathbf{x}^{(0)})}{q(\mathbf{x}^{(0)}|\mathbf{x}^{(T)}) \prod_{t=1}^{T} q(\mathbf{x}^{(t-1)}|\mathbf{x}^{(t)}, \mathbf{x}^{(T)})} \right].$$

For autoregressive models, the forward process q deterministically extends partial sequences, and the reverse process p_θ predicts the next token. For diffusion models, q adds noise (continuous) or masks tokens (discrete), while p_θ denoises or unmasks. We observe that both optimize the same objective with different choices of the forward process q and its corresponding optimal reverse process.

Continuous-Discrete Unification. Recent theoretical work has unified continuous and discrete diffusion under a single framework with certain theories [182, 104]. We also briefly elaborate here using Markov kernels on general state spaces. For a state space $\mathcal{X}$ (continuous $\mathbb{R}^d$ or discrete token space), define a forward kernel $\mathcal{K}_t : \mathcal{X} \times \mathcal{X} \to [0,1]$ and its reverse $\overline{\mathcal{K}}_t$. The forward process follows $p_t(\mathbf{x}) = \int p_0(\mathbf{x}^{(0)}) \mathcal{K}_t(\mathbf{x}^{(0)}, \mathbf{x}) d\mathbf{x}^{(0)}$. For continuous diffusion $\mathcal{K}_t(\mathbf{x}^{(0)}, \mathbf{x}) = \mathcal{N}(\mathbf{x}; \sqrt{\alpha_t}\mathbf{x}^{(0)}, (1-\alpha_t)I)$. For discrete diffusion $\mathcal{K}_t(\mathbf{x}^{(0)}, \mathbf{x}) = (\mathbf{x}^{(0)})^T Q_t \mathbf{x}$, where Q_t is a stochastic matrix. Both satisfy the same score matching objective in their respective spaces $\mathcal{L} = \mathbb{E}_{t \sim U[0,T], p_t(\mathbf{x})} \left[\lambda(t) || \nabla_{\mathbf{x}} \log p_t(\mathbf{x}) - s_\theta(\mathbf{x}, t) ||^2 \right]$, where for discrete spaces, the "score" becomes the difference in log-probabilities $\nabla_{\mathbf{x}} \log p_t(\mathbf{x}) \to \log p_t(\mathbf{x}) - \log p_t(\mathbf{x}')$.

5.4. Unified Learning Techniques

The unification of generative modeling techniques extends beyond individual architectures to encompass shared technical innovations that transcend molecular boundaries. These convergent techniques demonstrate that similar computational principles apply across both model architectures and tasks, e.g., whether generating DNA regulatory elements, RNA structures, or protein sequences.

Remasking strategies, originally developed for natural language models, have been adapted across models and biological domains with remarkable consistency. Language models use re-decoding strategies [89], diffusion models employ iterative refinement [280, 278], and masked models iterate through unmasking, all achieving progressive improvement through repeated processing. Intuitively, the remasking strategy performs another decoding for decoded parts of the output to achieve better generation. Remasking strategies in discrete diffusion models directly parallel BERT-style masking but with principled schedules for progressive unmasking. The ReMDM framework [292] introduces custom remasking processes that enable true iterative refinement rather than single-pass prediction. In protein generation, ESM-IF employs confidence-based remasking where the model iteratively refines low-confidence predictions $m_{t+1,i} = \mathbb{1}[\text{confidence}(x_{t,i}) < \tau_t]$, with threshold τ_t decreasing over iterations. The same principle appears in DNA sequence design, where RegLM uses gradient-based remasking to optimize regulatory sequences, where positions with highest gradient magnitude $|\nabla_{x_i} \mathcal{L}_{\text{expression}}|$ are remasked and regenerated. RNA structure prediction models like RNA-FM employ structure-aware remasking, prioritizing positions involved in base pairing $p(\text{remask}_i) \propto \sum_j \mathbb{1}[\text{pair}(i,j)] \cdot \text{uncertainty}(i,j)$.

Confidence-based generation ordering/refinement is also emerging as a shared principle across modalities and approaches, manifesting differently but serving similar purposes. The generative framework maintains a belief distribution over possible values at each position, $b_t(x_i) = p(x_i | x_{t,\neg i}, c)$, and decode/denoise positions by the entropy score order or refines positions where entropy $H[b_t(x_i)]$ exceeds a threshold. This manifests in RITA's protein design as iterative unmasking of amino acids based on predicted stability, in HyenaDNA's genome generation as progressive revelation of conserved regions before variable ones, and in RNAformer's structure prediction as refinement of base pairs from high to low confidence.

Curriculum learning strategies have converged on similar progressions across domains. Recent models employ progressive difficulty curricula that begin with simple sequences and gradually increase complexity through exponential pacing functions. This biological-inspired curriculum learning progresses in sequence length (from short to long), structural complexity (from simple motifs to complex domains), and evolutionary distance (from similar species to distant species), etc.. Models typically follow (1) length curriculum, i.e., training on short sequences before long ones, with length $L_t = L_{\min} \cdot (L_{\max}/L_{\min})^{t/T}$; (2) complexity curriculum, with simple structures before complex ones, measured by secondary structure content for proteins, regulatory element density for DNA, or base-pairing complexity for RNA; or (3) homology curriculum, with similar sequences before distant ones, gradually increasing evolutionary distance.

Hierarchical generation strategies decompose the problem into multiple scales. The generation process follows $p(x) = p(x_{\text{coarse}}) \prod_i p(x_{\text{fine},i}|x_{\text{coarse}})$. In protein design, this means first generating secondary structure topology, then backbone geometry, then side-chain conformations. For DNA, models first generate GC content and repeat patterns, then specific sequences. RNA generation proceeds from secondary structure motifs to full tertiary structure. This hierarchical decomposition reduces the effective dimension of the generation problem from $O(L)$ to $O(\sqrt{L})$ where L is sequence length.

Temperature scheduling provides fine-grained control over generation diversity. Modern models use adaptive temperature $T_t = T_0 \cdot \exp(-\lambda t) + T_{\min}$, starting with high temperature for diversity and annealing for quality. Biological models further adapt temperature per-position based on evolutionary conservation $T_{t,i} = T_t \cdot (1 + \alpha \cdot \text{conservation}_i)$, allowing variable regions to explore more while conserved regions remain stable.

These technical convergences reveal that biological sequence generation, regardless of the specific molecular type, shares fundamental computational challenges: managing uncertainty across multiple scales, satisfying complex constraints, and navigating vast combinatorial spaces. The success of unified techniques across DNA, RNA, and proteins suggests that future advances in any one domain will likely transfer to others, accelerating progress toward truly universal biological generation models. The convergence is not mere technical convenience but reflects deep commonalities in how biological information is encoded, processed, and functionalized across the molecular foundations of life.

5.5. Advantages of Unified Architectures for Biological Generation

The unified view of modeling we analyzed in this section reveals several theoretical advantages for biological generation. First, biological sequences encode information at multiple scales, local motifs, secondary structures, and long-range interactions. The hybrid framework can implicitly and naturally handles this through $p(x|c) = p_{\text{local}}(x_{\text{motifs}}|c) \cdot p_{\text{global}}(x_{\text{structure}}|x_{\text{motifs}}, c)$, where autoregressive generation excels at capturing local syntax (e.g., conserved motifs) while iterative refinement enforces global consistency (e.g., structural constraints). Second, biological molecules must satisfy numerous constraints, valid chemistry, thermodynamic stability, functional requirements. Iterative models can enforce these through projection operations $x_{t-1} = \Pi_{\mathcal{C}}[\mu_\theta(x_t, t, c) + \sigma_t \epsilon]$, where $\Pi_{\mathcal{C}}$ projects onto the constraint set $\mathcal{C}$. This is difficult for purely autoregressive models, which would need to backtrack upon constraint violations. Also, the iterative nature enables principled uncertainty estimation through the variance of the generative process $\text{Var}[x] = \mathbb{E}_{p(x_T)}[\text{Var}[x|x_T]] + \text{Var}_{p(x_T)}[\mathbb{E}[x|x_T]]$. This decomposition separates aleatoric uncertainty (inherent variability) from epistemic uncertainty (model confidence), crucial for biological applications where we must know when predictions are reliable. Furthermore, the unified framework enables dynamic allocation of computational resources based on problem difficulty. Define a difficulty measure $d(x_t) = -\log p(x_t|c)$ representing the model's uncertainty. Then, adaptive generation allocates more iterations to difficult regions

$T(x) = T_{\text{base}} + \alpha \cdot d(x)$. This mirrors biological evolution, which spends more time optimizing functionally critical regions. Finally, biological systems are inherently modular, where proteins contain domains, DNA has regulatory modules, RNA forms structural motifs. The unified framework naturally supports compositional generation $p(x) = \prod_{i=1}^{M} p(x^{(i)}|x^{(<i)}, c) \cdot \prod_{i,j} \psi(x^{(i)}, x^{(j)})$, where $x^{(i)}$ are modules, the first term generates modules autoregressively, and ψ are compatibility potentials refined through iteration.

These theoretical benefits manifest practically in improved performance metrics. The effective sample complexity for learning biological distributions scales as $N_{\text{eff}} \sim O(d/\epsilon^2)$ for autoregressive models but $N_{\text{eff}} \sim O(\sqrt{d}/\epsilon)$ for iterative models when constraints are present, where d is the sequence dimension and ϵ is the desired accuracy. This square-root advantage explains why hybrid models consistently outperform pure autoregressive approaches on structure prediction and design tasks.

Beyond performance improvements, unified or hybrid models enable entirely new capabilities impossible with single-paradigm approaches. Cross-modal tasks represent the most significant emergent capability. For example, IsoFormer [87] became the first model capable of predicting RNA transcript isoform expression from DNA, RNA, and protein modalities simultaneously. ChatNT [229] demonstrates zero-shot generalization across 27 different biological tasks through natural language interfaces. Attention specialization emerges from unified training, with IsoFormer showing different RNA encoder layers specializing for 3'UTR, 5'UTR, and CDS regions when trained in multi-modal settings. This biological motif recognition enables ChatNT to identify TATA boxes and splice sites without explicit programming for these features.

Computational efficiency gains are also shown to be substantial. For example, ChatNT's single 7B parameter model replaces 27 separate specialized models while maintaining or exceeding performance on individual tasks. This parameter efficiency translates to reduced deployment complexity, lower maintenance overhead, and improved scaling efficiency for adding new biological tasks.

The convergence of generative modeling approaches in biology represents more than technical innovation, but also signals a fundamental shift in how we understand and engineer biological systems. Unified models consistently outperform specialized approaches while offering better computational efficiency, transfer learning capabilities, and biological interpretability. The success stories from AlphaFold3's universal biomolecular modeling to Evo1 and 2's megabase-scale sequence generation demonstrate that biological complexity demands integrated approaches that mirror the interconnected nature of life itself. As these unified frameworks continue to evolve, we can expect further integration of reinforcement learning, causal modeling, and multi-scale approaches that span from molecules to organisms. The convergence of generative models in biology is not merely a trend but a necessary evolution toward AI systems capable of understanding and designing life at its full complexity.

6. Toward Unifying Post-Training Algorithms

The power of large-scale pretrained models can only be fully realized through effective adaptation to specific biological tasks. This section examines how fine-tuning and adaptation techniques, originally developed for natural language models, have been successfully transferred and modified for biological applications. We explore both the commonalities across molecular domains and the unique challenges each presents, highlighting how cross-domain insights can improve adaptation strategies.

6.1. Inference-Time Strategies

One essential aspect when applying generative AI models to biological applications is to control their generation behavior so that the generations satisfy the practical use cases. In particular, the generations need to not only be valid in-distribution samples, but also align with certain conditions and objectives. For example, we want to generate small molecules that can bind to a given protein, DNA sequences that can express in certain cell types, and proteins that have desirable druggable properties [279].

To enable effective controllability, a variety of inference-time strategies have been applied to leverage both the generation capability of generative models and the guidance signals. Among them, one straightforward method is to model it as a conditional generation problem, where the specific environment, desirable properties, or guidance signals are input to the model as a condition and incorporated into model training. During inference time, classifier-free guidance (CFG) [112], originally proposed for image diffusion models, is widely applied to enhance generation quality. To be specific, the estimated conditional score is combined with the unconditional score during diffusion sampling, and a weighting hyperparameter is utilized to control the tradeoff between sample quality and diversity. CFG, along with an improved variant AutoGuidance [142], is widely applied to biological design applications, such as protein backbone conditions, essential functional motifs, binders, symmetry, and structural constraint for protein design [90, 303, 5, 286, 256]. Despite its strong performance, training such conditional generative models requires paired data containing the corresponding labels, which is expensive to obtain under many biological settings. Further, for reward optimization tasks, while it is feasible to train conditional generative models conditioning on reward values, it is shown to be suboptimal since high-reward samples are rare [277, 291].

Alternatively, training-free methods can be applied for more flexible guidance. First, gradient based methods, specifically classifier guidance [253, 65] and its variants [19, 51, 114], utilize the gradient of a predictor model during inference. Specifically, classifier-guidance decompose the conditional score function into the unconditional score and the gradient of a predictor model with respect to samples in intermediate diffusion timestep. This is further extended to the biological settings for both continuous and discrete diffusion models [209, 312]. Such method largely depends on the gradient of the predictor model, while can have large approximation error especially for diffusion models in the discrete state space.

Another type of inference-time methods is the value based sampling strategy, especially sequential Monte Carlo (SMC) based methods and its variants [109, 307, 274, 69, 36, 220, 161, 346]. To be specific, SMC combines the pretrained (unconditional) generative model $p_0(x)$ with the task specific objective function $r(x)$ to formulate an exponential reward weighted target distribution, and leverages twisting to modify proposals and weighting schemes to approach the optimal target distribution $p^{(\alpha)}(x)$:

$$p^{(\alpha)}(x) \propto \exp(r(x)/\alpha)p_0(x).$$

SMC decomposes the inference into easier subproblems via iterative rounds of an importance sampling step and a selection step, where the weights are in proportion to the ratio of the exponential value function of the next step and that of the current step. Notably, a key component in value based decoding methods is the value function estimation for an intermediate state in the diffusion process. One simple way is to utilize a training-free posterior mean estimation [161, 51]. However, this relies on the one-step prediction of a noisy sample to the clean sample, and an additional reward evaluation based on the prediction. This can be inaccurate when the sample is very noisy and the one-step prediction is thus far out of the reward function input distribution. Also, the prediction can not fully represent the actual data distribution generated from the current noisy sample, and thus leads to bias to the true value. Alternative approaches include various value function learning algorithms, such as Monte Carlo regression, soft Q-learning [204], contrastive twist learning [346], etc.

In addition to gradient-based and sampling-based algorithms, search-based methods have been applied for optimization under a variety of settings, for example designing promoter or enhancer sequences that express in certain cell types. Moreover, such method can be combined with sampling algorithms to enable leveraging the data distribution from the pretrained generative models and efficiently explore a broad space [162].

In summary, these inference-time strategies provide powerful tools to steer the generative models to certain directions and enable generations that have desirable properties or satisfy given conditions. This plays an important role in biological design tasks, where we want to control the generations in a flexible and data efficient way.

6.2. Task-Specific Fine-tuning

Apart from inference time adaptations, task-specific fine-tuning is another powerful approach that can both leverage the strong pretrained generative models and alter it to specialized generators. While inference time strategies have their own advantages, they generally leads to higher inference time computational cost compared to the fine-tuning methods. Fine-tuning algorithms of generative models can be applied in a variety of use cases, in particular reward optimization.

Reinforcement learning algorithms have been widely applied to fine-tune large language models to align them with human feedback [212, 356] or improve task-specific rewards (i.e., inducing reasoning behavior) [243, 99, 215] during the post-training stage. Specifically, the model is fine-tuned with either a reward model trained from human preferences or process reward labels [165] or a verifiable reward based on the final output, using different reinforcement learning algorithms, including policy gradient [264], PPO [240], TRPO [239], and GRPO [243], or preference optimization methods like DPO [224]. These methods have been applied to language models for biological generation. For example, ProteinDPO [304] fine-tunes the protein inverse folding model ESM-IF [117] by constructing preference pairs based on stability data of protein sequences and utilizing the DPO algorithm. After fine-tuning, the model is able to generate protein sequences that can fold to the given structure while achieving high stability.

Such fine-tuning algorithms, originally proposed for autoregressive language models, can be generalized to diffusion models. Numerous studies have explore RL-based fine-tuning for continuous diffusion models [75, 23, 53, 221, 276]. The denoising trajectory in the continuous diffusion process is formulated as a multi-step Markov decision process (MDP), with the state defined as the current noisy sample x_t, timestep t and the condition c, and the transition kernel defined as the one-step denoising distribution $p_\theta(x_{t-1}|x_t, c)$. The task-specific reward is generally calculated based on the final clean output x_0. Further, these fine-tuning algorithms have been extended to diffusion models in the discrete space, where the diffusion process is instead formulated as a continuous time Markov chain (CTMC) [290, 34, 345, 100, 285, 261]. These fine-tuning algorithms have been applied to a broad set of biological applications, including generating enhancer sequences with high expression level in certain cell types, designing proteins with high stability, generating small molecules with desirable properties, etc.

6.3. Computation and Data Efficient Adaptation

One of the key challenges when applying generative models to biological applications is the limited amount of data, typically obtained with high cost. Importantly, while task-specific data may be limited, there exists a larger set of related data that can be helpful to the task of interest. For example, when designing proteins for a specific target, general protein folding data provides helpful knowledge that can be transferred to the specific task. This requires using a pretrained model and adapting it for specific usage.

In practice, few-shot learning and transfer learning techniques can be applied to enhance the model's ability on various tasks. Essentially, few-shot learning is designed to enable models to generalize to new tasks with only a few task specific samples, via meta learning, domain adaptation, or multi-task learning. Meta learning [79], i.e., learning to learn, represents a paradigm where the model is trained on a wide distribution of different subtasks, and a generalizable learning strategy is developed to deal with new, unseen tasks with limited data. Another widely used techniques in transfer learning is domain adaptation [179, 180, 355], where the model is trained on data-rich source domains, while evaluated on a data-sparse target domain. This requires the model to learn features that are irrelevant to domain specific information and transferrable to the new target domain. Additionally, multi-task learning, where a single model is trained to conduct multiple tasks simultaneously, is an alternative way to improve generalizability and training efficiency under the scenario with limited task specific data [37, 241, 343]. The model extracts shared representations from the given data and utilize the data label from different tasks to enhance the overall performance. These techniques greatly improve the data efficiency, and have been applied in biological applications, including predicting regulatory element activity [233], learning genetic perturbation effect [205], modeling protein fitness landscapes [352], predicting multiple molecular properties [178], etc.

In addition, pretrained generative models can be adapted to follow complex task specifications via instruction tuning (also known as supervised fine-tuning), a widely applied fine-tuning technique for large language models on a labeled dataset of instructional prompts and the corresponding outputs [212, 172, 217, 340]. Specifically, instruction tuning enables the pretrained model to follow certain prompts or conditions and bridge the gap between the pretraining objective and the downstream use cases of users. Instruction tuning can be applied to biological design tasks, e.g. proteins, small molecules and regulatory DNA sequences, where a pretrained model is fine-tuned to follow certain design instructions. For example, [299] aligns human and protein language via instruction tuning to enable text-protein joint tasks like text-based protein function prediction and sequence design. [190] develops a multitask LLM with protein domain expertise using a comprehensive instruction dataset to enable protein language understanding and generation. Similar strategies can be potentially applied to small molecules [35] and regulatory DNA/RNA sequences [164].

Apart from the challenge of data limitation, computation efficiency is another essential aspect to consider given the emerging large-scale models and the computational resource constraint in biological labs. Parameter-efficient fine-tuning (PEFT) [101] is a technique developed for large language models to adapt pretrained LLMs to new tasks by freezing most of the model parameters and only allow for the training of a few parameters. One common method among them is Low-Rank Adaptation (LoRA) [118], which introduces trainable low-rank matrices to the transformer layers and only trains these matrices as a low-rank approximation of the original weight updates. Another way is to conduct modular fine-tuning, which introduces individual modules for separate tasks or objectives over the shared neural network components, for example soft prompt tuning [158] or multitask adapter fine-tuning [219, 297, 47]. Such techniques can be applied to large-scale models in the biological setting, such as Proteína [90].

6.4. Reinforcement Learning from Biological Feedback

As mentioned above, reinforcement learning algorithms can be applied to align generative models with human feedback or human preference [212, 356]. Similarly, in biological applications, the biological feedback obtained from wet-lab experiments, biological prior knowledge, or specific constraint can be incorporated to enhance model capabilities. For example, [324] designs high-fitness *Cis*-regulatory elements (CREs) via RL fine-tuning of a pretrained autoregressive model using fitness reward as the feedback. [45] designs regulatory DNA sequences to achieve cell-type specific gene expression by fine-tuning the genomic language models to optimize reward under expression constraints.

Importantly, this feedback can be obtained in cycle to continuously improve the ability of the generative models, via active learning [83, 242]. Active learning algorithms aim to select the most uncertain unlabeled samples to query for label from human or oracles. Considering the cost of obtaining labels, active learning makes the process more efficient by attaining the label with the highest gain. Considering the time cost of the label querying process, batch active learning [52] is proposed to conduct querying in a large batch at scale. Such techniques have a broad application to the biological domain, where the cost of obtaining experimental feedback is high and the optimization landscape is hard to predict with limited data. Machine learning guided directed evolution [321, 135] designs novel protein sequences by combining protein language model with a property predictor, iteratively updated by newly acquired wet-lab experimental feedback. [123] conducts optimal experimental design of perturbation screens to select the most informative perturbations at each step and enables an efficient exploration of the perturbation space. Compared with text and image domains, biological domain has the unique challenge in terms of the limited types of feedback, noise in the data and the high cost during labeling. Therefore, it is essential to develop algorithms that can incorporate useful feedback from various resources, such as a combination of cheap while less accurate data with accurate but expensive data.

6.5. Scaling Laws in Post-Training

Recent advances have shown that the performance of large models can be systematically improved not only by increasing model size during pretraining, but also by scaling resources during post-training stage. In practice, performance beyond pretraining can be improved through both fine-tuning and inference-time scaling.

Representative examples of fine-tuning scaling include InstructGPT [212], which demonstrates performance gains with increased instruction and preference data. Subsequent work [50] further demonstrates that instruction fine-tuning exhibits strong scaling behavior, where increasing task diversity, model size, and chain-of-thought supervision leads to substantial and consistent performance gains across a wide range of models and benchmarks. Beyond instruction supervision, another study [333] systematically analyzes fine-tuning scaling across multiple factors and finds that performance follows joint scaling laws governed by model size and fine-tuning data, with model scaling playing a more dominant role than pretraining data or parameter-efficient tuning in data-limited regimes. Complementary to these results on autoregressive language models, recent work [326] demonstrates that diffusion-based language models also exhibit clear scaling behavior with respect to data, model size, and instruction fine-tuning, leading to consistent gains across downstream tasks and enabling strong zero-shot and few-shot generalization.

Beyond fine-tuning, many works show that model performance can also scale systematically with increased test-time computation. In particular, it is shown that in large language models, optimally allocating inference-time compute can be more effective than scaling model parameters, enabling smaller models to outperform substantially larger ones on sufficiently challenging prompts [249]. Importantly, such test-time scaling behavior is not limited to autoregressive language models. Diffusion models have also been shown to exhibit clear performance gains as additional inference-time computation is allocated beyond standard denoising schedules [193]. Further supporting this perspective, subsequent studies provide empirical evidence that inference-time computation in LLMs follows systematic scaling laws, where more advanced inference strategies allow smaller models to achieve Pareto-optimal trade-offs under fixed compute budgets [308]. Moving beyond general generative models, recent work on protein foundation models demonstrates that increased test-time computation can likewise be leveraged to yield systematic performance improvements, with gains scaling predictably as verification budgets are intensified during inference [189].

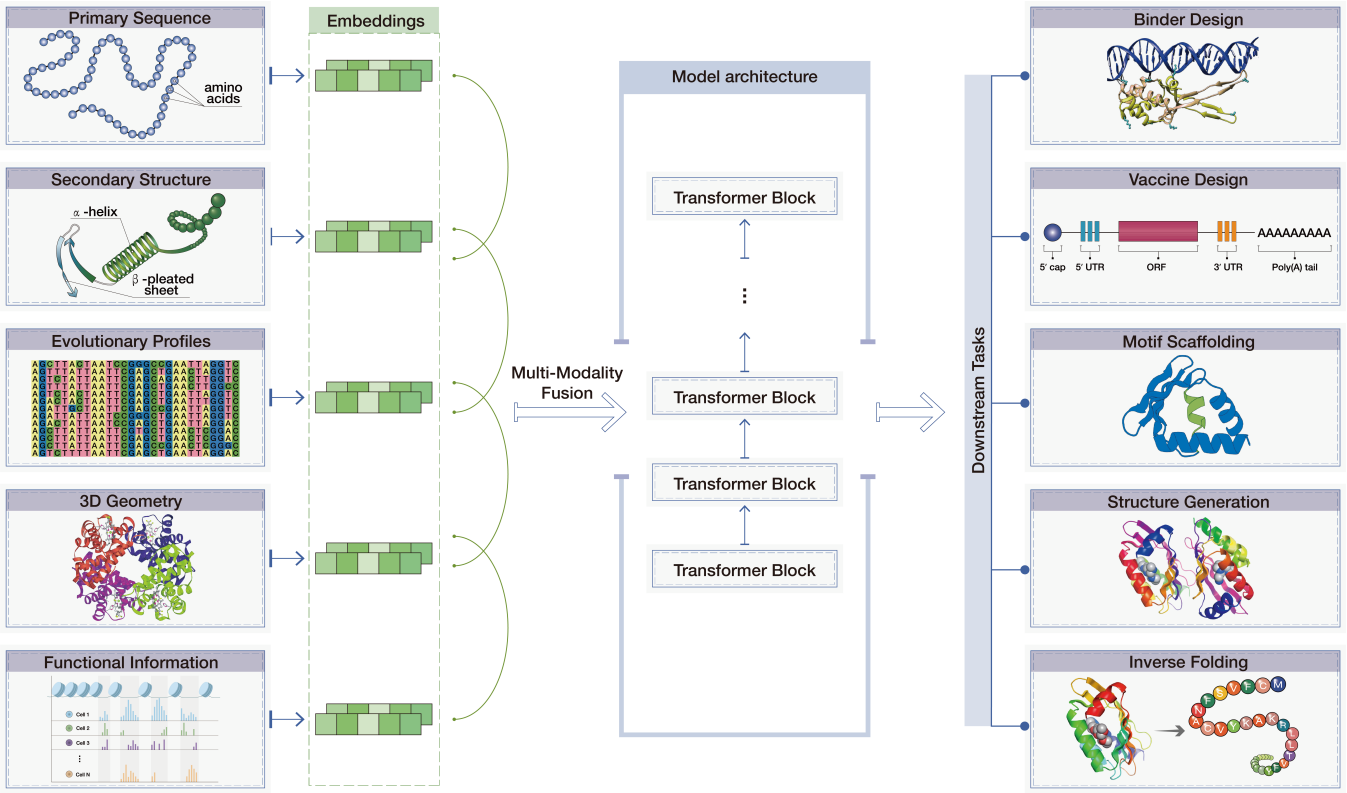

Figure 5: A multi-modal foundation model for biomolecules, where heterogeneous inputs are embedded and fused into a backbone to support diverse downstream design tasks such as binder/vaccine design, motif scaffolding, structure generation, and inverse folding.

7. Toward Unifying Modalities and Capabilities

7.1. Foundation Models and Multi-Modal Systems

The biological world is inherently multi-modal: DNA, RNA, and proteins represent distinct yet interconnected molecular types, cellular phenotypes emerge from their interactions, and all of these are contextualized through natural language in scientific discourse. Traditional models, which treated each data type in isolation, have struggled to capture the full spectrum of biological relationships. Recent advances in foundation models are beginning to bridge across these modalities, e.g., linking nucleotide sequences with protein function, or connecting molecular data with natural language, enabling joint representation learning and cross-domain reasoning that hold the promise of reshaping biological discovery and design. Figure 5 presents the unified-representation paradigm in biomolecular modeling, where multi-source biological signals are encoded into modality-specific embeddings, fused, and processed by a shared foundation models to enable both predictive and generative downstream objectives.

In the following subsections, we discuss representative multi-modal biological foundation models (Section 7.1.1), examine how scale and corpus diversity drive emergent capabilities (Section 7.1.2), explore the integration of foundation models with biological knowledge bases (Section 7.1.3), and highlight the role of natural language interfaces in bridging human intent with computational design (Section 7.1.4).

7.1.1. Multi-Modal Biological Foundation Models

In this section, we focus on cross-modality within biology, where a model conditions on one molecular modality to generate another, without relying on natural language.

AlphaFold3 [2] unifies proteins, nucleic acids, ligands, ions, and modified residues in a single framework to generate multi-molecular complexes and predict their interactions. It employs diffusion-style generative refinement together with long-range coordination modules to produce coherent all-atom assemblies. RoseTTAFoldNA [17] extends RoseTTAFold [14] to assemblies that pair proteins with DNA or RNA and improves modeling when homologous templates are scarce. It introduces nucleic-acid–aware representations that better capture base pairing and protein–nucleic acid interface geometry. The model also supports binding-site localization and interface-quality assessment for downstream analysis. RoseTTAFold All-Atom [146] generalizes to heterogeneous all-atom complexes that include proteins with small molecules, metals, and chemical modifications, enabling complex generation and evaluation in one stack. Its SE(3)-equivariant all-atom architecture jointly reasons over covalent chemistry and noncovalent contacts. The framework supports both forward modeling and inverse design within a unified pipeline. Chai-1 [25] couples structure-aware modeling with sequence generation to design and assess binders across molecular targets in an all-atom setting. It is trained on heterogeneous complex data and provides differentiable objectives and fast re-scoring to prioritize candidates before wet-lab testing. The model integrates geometric consistency checks to enhance interface plausibility. Chai-2 [271] advances this pipeline with zero-shot de novo antibody and mini-protein generation and reports strong wet-lab hit rates under target-conditioned design. It conditions on target epitopes or surfaces to co-sample structures and sequences that satisfy interface constraints. The method demonstrates generalization across antigen classes with minimal task-specific tuning. Evo2 [29] provides a unified backbone over DNA, RNA, and protein sequences and supports conditional generation and zero-shot transfer across these sequence modalities. It learns shared latent codes that enable few-shot adaptation between sequence types. The model further supports editing scenarios in which changes in genomic or transcript sequences propagate to protein design constraints. IsoFormer [87] aggregates DNA, RNA, and protein encoders into a unified architecture for cross-modal transfer learning. It enables prediction of transcript expression across tissues and facilitates adaptation between sequence types with limited supervision.

Together these works outline a common recipe for biological cross-modality: learn a shared latent space across heterogeneous molecular types, enforce inter-molecular constraints during conditional generation, and apply fine-grained all-atom scoring for selection, setting up the scale-driven behaviors discussed next.

7.1.2. Cross-Modal Emergent Properties

Progress in biological foundation models has closely tracked increases in pretraining corpus size, diversity, and context length. As models ingest broader and more heterogeneous sequence collections, their representations capture richer structural and functional regularities that support downstream generalization [230, 166, 58].

Empirical “scaling law” trends have been observed in protein and genomic language models: performance improves predictably with parameters, tokens, and compute, and the gains persist across evaluation families rather than a single benchmark [230, 166]. Long-context training compounds these effects by enabling models to integrate signals spanning kilobases to megabases, which is essential for regulatory genomics [206].

At larger scales, models exhibit capabilities not explicitly optimized for during pretraining. Protein LMs trained only on sequences nonetheless encode contact and secondary-structure information and can support near-atomistic folding with lightweight heads [166]. Generative models produce functional proteins in the

absence of task-specific supervision, highlighting a shift from pattern recognition to actionable design [195]. Unified backbones that cover DNA, RNA, and protein further reveal cross-modality transfer in zero-shot settings [29].

In-context learning has begun to appear in biological settings when context windows are sufficiently long and optimization is stable. Genomic LMs can use few-shot exemplars supplied in the input to perform motif discovery, variant prioritization, and locus-specific reasoning without parameter updates [206]. This behavior suggests that part of the benefit of scale arises from improved amortization of task specification into the context.

Zero-shot generalization provides an additional lens on emergence. Sequence-only models predict mutational effects on held-out deep mutational scanning assays without assay-specific training [28]. Cross-modality sequence models transfer between DNA, RNA, and protein tasks under unified tokenization and objectives [29]. Single-cell foundation models pretrained on large atlas-scale corpora also show limited but measurable zero- or few-shot transfer to new cell types and perturbations, clarifying both the promise and the limits of scale in cellular settings [272, 57].

Taken together, these observations motivate scaling along three axes—data volume and diversity, model capacity, and context length—while emphasizing the need for robust calibration, standardized cross-dataset evaluations, and systematic wet-lab validation to distinguish true emergence from benchmark saturation.

7.1.3. Integration with Knowledge Bases

While biological foundation models have demonstrated remarkable capacity in learning from raw sequences, structures, and natural language, they often lack access to structured biological priors such as curated pathways, ontologies, and interaction networks. Knowledge bases (KBs), which organize facts and mechanistic relationships across molecular entities, provide a complementary resource that can enhance the reasoning, interpretability, and generalization capabilities of foundation models. Recent research highlights several paradigms for integrating KBs with foundation models, bridging the gap between data-driven pattern recognition and knowledge-grounded inference.

OntoProtein [338] integrates knowledge graph (KG) information from Gene Ontology into protein language model pretraining. It constructs a large-scale GO–protein graph, where nodes are proteins or ontology terms described by sequences or textual annotations. The model jointly optimizes protein and ontology embeddings through contrastive learning with knowledge-aware negative sampling, enabling structured biological knowledge to inform sequence representations. BioBridge [301] uses knowledge graphs as a bridge across independently trained single-modality models, aligning their representations without full fine-tuning. This approach improves cross-modal retrieval and supports applications such as multi-modal question answering and drug design. MolFM [188] embeds molecular structures, scientific text, and KB entities into a shared space through cross-attention, linking structural features with curated graph relations. This joint embedding enhances performance in cross-modal retrieval under both zero-shot and fine-tuned settings. BioT5 [216] incorporates both structured KBs and unstructured literature during pretraining, using SELFIES to encode valid chemical knowledge while linking it to natural language. This design enriches cross-modal understanding and boosts performance on biochemical prediction tasks. BioReason [74] integrates pathway databases such as KEGG with foundation models, coupling DNA-based encoders and LLMs for multi-step, interpretable reasoning. It achieves improved disease and variant prediction while generating explicit knowledge-grounded reasoning chains.

Together, these methods show how bridging, joint embedding, knowledge-informed pretraining, and

ontology-guided reasoning can expand the scope and trustworthiness of biological foundation models.

7.1.4. Natural Language Interfaces

Chroma [130] is a programmable diffusion-based generative model for protein sequence and structure co-design. It formulates protein generation as Bayesian inference under external constraints, enabling conditioning on diverse inputs such as structural motifs, symmetry, or even natural-language prompts to guide functional or semantic properties. ProteinDT [174] introduces a multimodal framework that connects natural language with protein representations through a contrastive alignment module, ProteinCLAP, enabling text-guided protein design. To support this cross-modal learning, the authors construct SwissProtCLAP, a large-scale dataset comprising 441,000 text–protein pairs that links functional descriptions to corresponding sequences, establishing a foundation for language-conditioned protein generation and analysis. ProtST [315] introduces a multimodal pre-training framework that aligns protein sequences with their corresponding biomedical text descriptions through tasks such as masked sequence prediction, cross-modal representation alignment, and multimodal mask prediction. The authors also compile the ProtDescribe dataset, consisting of over 550,000 paired protein and text entries from Swiss-Prot [18] to train this framework. Prot2Text [1] frames protein-function annotation as a text-generation task by jointly encoding protein sequence and structure via a Graph Neural Network and a pretrained protein language model, and decoding free-text functional descriptions with an LLM. It also releases a large multimodal dataset of 256,690 proteins to support this cross-modal learning. ProtCLIP [349] proposes a CLIP-style multimodal framework that aligns protein sequences with textual biological descriptions through contrastive pre-training. To incorporate fine-grained biological semantics, the model introduces two segment-wise pre-training objectives that explicitly model static and dynamic functional segments, thereby injecting function-informed supervision into protein representations. The authors further construct ProtAnno, a large-scale protein–text paired dataset combining high-quality and weakly labeled samples, to support this function-aware multimodal alignment. ProteinCLIP [305] also applies a CLIP-style contrastive learning framework that pairs protein amino acid sequences with curated natural-language descriptions of their functions. By refining embeddings from pre-trained protein language models toward functional alignment with textual representations, it produces function-centric protein embeddings. ProtT3 [176] introduces a protein-to-text generation framework that bridges protein sequence understanding and natural language modeling. It incorporates a protein language model (PLM) to encode amino acid sequences and a language model (LM) to generate descriptive text. ProtChatGPT [289] introduces an interactive, LLM-based system for protein research that combines frozen protein encoders and a projection adapter to align multi-level protein embeddings with a large language model. Users can upload protein sequences or structures and pose questions; and the system responds with natural-language explanations or design suggestions by bridging the modality gap between protein data and text. TextSMOG [187] introduces a text-guided diffusion framework that bridges natural-language prompts and three-dimensional molecular structure generation. It conditions an equivariant 3D diffusion model on textual descriptions via a multi-modal conversion module that translates language into reference geometry, enabling the model to generate chemically valid and structurally novel small molecules aligned with rich functional language directives.

7.2. Agent-Based Systems and Tool Use

Beyond standalone models, the integration of generative AI into agent-based systems represents the next frontier in computational biology. These systems combine generative models with reasoning capabilities, external tools, and iterative experimentation cycles. This section explores how agent architectures are enabling more sophisticated biological design workflows that mirror the iterative nature of real laboratory research.

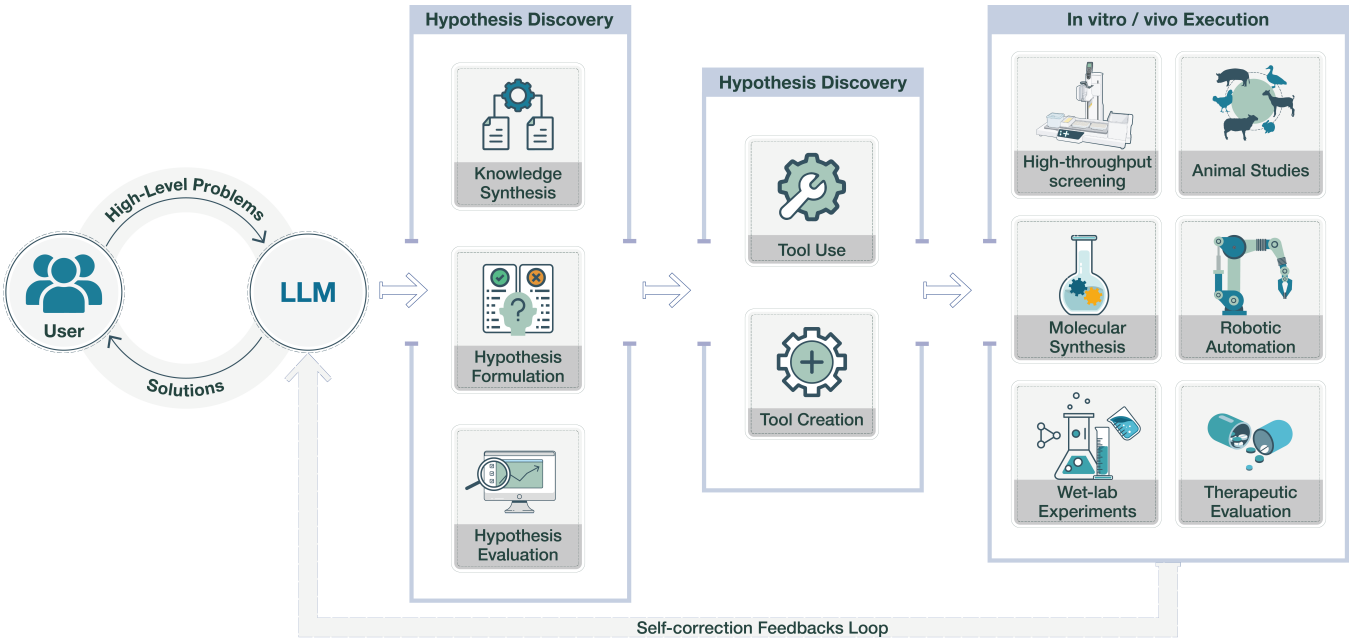

Figure 6: The biological agent paradigm that couples hypothesis generation with tool use/creation and laboratory automation, enabling iterative refinement from computational reasoning to experimental validation.

Figure 6 illustrates the agent framework for scientific discovery, where user-defined high-level problems are translated into hypotheses (synthesis, formulation, and evaluation), augmented by tool use/creation, and iteratively validated through in vitro/in vivo execution (e.g., screening, synthesis, wet-lab experiments, and therapeutic evaluation) via a self-correction feedback loop.

7.2.1. Biological Design Agents

The emergence of LLM-based agents has transformed experimental planning in biology [84, 350]. BioDiscoveryAgent achieved 21% improvement in predicting genetic perturbations by integrating Claude 3.5 Sonnet with specialized biological tools, demonstrating fully interpretable AI decision-making for experimental design [231]. The integration with laboratory automation reached commercial viability through platforms like the Autonomous Enzyme Engineering Platform, which eliminated human intervention entirely while achieving 16-fold improvement in AtHMT ethyltransferase activity and 95% mutagenesis success rates [247].

Multi-step reasoning capabilities advanced significantly through systems like BioReason, the first deep integration of DNA foundation models with LLMs, improving KEGG-based disease pathway prediction from 88% to 97% accuracy [74]. EVOLVEpro system exemplified AI-based protein optimization beyond evolutionary constraints, successfully enhancing stability and efficiency across six different proteins [135]. These systems represent a paradigm shift from traditional computational tools to autonomous experimental agents capable of generating novel biological insights.

7.2.2. Tool Integration and APIs

Structure prediction integration has matured into standardized platforms, with the AlphaFold ecosystem providing seamless access through Google Cloud Vertex AI for protein interactions with DNA, RNA, and ligands [2]. Performance benchmarks demonstrate practical scalability: 100-residue proteins in 4.9 seconds

on A100 GPUs. Levitate Bio's unified AI Folding API standardized interfaces for AlphaFold2, ESMFold, and RoseTTAFold, eliminating integration complexity while maintaining specialized capabilities [141, 166, 14].

Database integration has evolved beyond simple queries into sophisticated AI ecosystems. The BioServices framework provides unified access to 40+ bioinformatics web services [54], while the AlphaFold Database API offers 214+ million protein structure predictions [283]. Workflow management systems like Nextflow [66] and Snakemake [145] have become essential infrastructure, enabling reproducible biological analyses across cloud and HPC environments with automatic parallelization.

ToolUniverse pushes this evolution one step further by treating tools as a first-class interface for AI scientists, not just for humans writing pipelines. It provides a scientific environment that lets an AI model discover and invoke scientific tools through a standardized protocol, reducing one-off glue code and brittle wrappers [85]. ToolUniverse integrates 700+ tools spanning models, datasets, APIs, retrieval systems, and analysis utilities, and it emphasizes tool specifications so heterogeneous tools can be composed [85]. Beyond access, it automatically refines tool interfaces for correct use, supports generating new tools from natural-language descriptions, and composes tools into agentic workflows [85]. In effect, it shifts integration from running workflows to orchestrating a scientific ecosystem where tool selection and execution are explicit parts of the scientific loop [85].

7.2.3. Iterative Design Cycles

Generate-test-refine workflows have transformed biological design by shifting from static screening to dynamic, information-maximal experimental planning. The SAMPLE platform [228] discovered enzymes 12°C more thermostable while searching less than 2% of combinatorial space, achieving 83% accuracy in activity classification.

To handle complex search spaces, recent frameworks have incorporated optimal experimental design (OED) principles. For instance, active learning agents for perturbation screens can now dynamically select the most informative perturbations at each step, maximizing information gain while minimizing experimental cost [123]. In drug discovery, this approach has yielded substantial efficiency gains; AstraZeneca's reinforcement learning system demonstrated 5–66-fold increases in hits for fixed oracle budgets by continuously updating its selection strategy based on feedback [67].

7.2.4. Multi-Agent Collaboration

Specialized agent architectures proved superior to single-system approaches. Google's AI Co-Scientist deployed seven specialized agents including generation, reflection, and meta-review agents, successfully validating discoveries across drug repurposing and antimicrobial resistance [92]. Harvard's comprehensive framework identified four multi-agent configurations handling structural biology, genetics, and chemistry tasks, demonstrating autonomous literature synthesis and novel hypothesis generation [84].

The NSF's $100 million investment in programmable cloud laboratories envisions national networks of AI-enabled laboratories with custom workflows. FutureHouse operationalized specialized agents including Crow for literature retrieval and Phoenix for chemistry planning, enabling natural language interactions while maintaining coordination across biological subdisciplines [248, 191].

The convergence of biological design agents, tool integration, iterative cycles, and multi-agent collaboration has created validated systems with 16–90x performance improvements in protein engineering and 50–100x acceleration in materials synthesis [228, 266]. As standardization matures and integration challenges resolve, these systems promise to transform biological research, enabling previously impossible

experimental approaches across biotechnology and synthetic biology applications.

8. Challenges, Opportunities, and Future Directions

Despite the remarkable convergence of generative modeling techniques across biological domains, significant gaps remain in both methodological coverage and practical implementation. This section identifies critical unexplored intersections where successful techniques await cross-domain application, outlines the path toward truly universal biological models, addresses the need for standardized evaluation frameworks, and highlights promising directions that could reshape the field.

8.1. Unexplored Intersections: Gaps and Opportunities

The methodological landscape of generative biology reveals asymmetries in technique adoption across molecular domains. While protein generation has benefited from sophisticated diffusion-based approaches like RFdiffusion [303] and FrameDiff [328] that operate directly on 3D coordinates with $SE(3)$-equivariance, analogous methods for nucleic acid structure generation remain nascent. RNA tertiary structure generation lags significantly behind protein capabilities, with RiboDiffusion [120] and RNAFlow [210] representing early attempts that have yet to match the designability and experimental validation rates achieved in protein engineering. The disparity is even more pronounced for DNA, where structural generation through DiscDiff [163] remains limited to simplified representations rather than the all-atom precision now standard in protein design.

Inverse folding, a cornerstone technique in protein design through models like ProteinMPNN [60] and ESM-IF [117], has seen minimal adaptation to nucleic acids despite the clear analogies between protein backbone design and RNA/DNA structure-conditioned sequence generation. The few attempts at RNA inverse folding lack the sophisticated graph neural network architectures and extensive experimental validation that characterize protein applications. This gap is particularly striking given that RNA secondary structure prediction has long been a computational biology success story, suggesting that the building blocks for effective RNA inverse design already exist but await proper integration.

The remarkable success of discrete flow matching for DNA sequence generation, exemplified by Dirichlet Flow Matching [255] and Fisher-Flow [62], has not been matched by similar innovations in protein or RNA domains. These methods leverage smooth probability transport paths specifically tailored for categorical data, yet protein sequence generation continues to rely primarily on autoregressive models or standard discrete diffusion. The potential for Fisher-Rao metric geodesics to improve protein sequence generation remains entirely unexplored, despite proteins sharing the fundamental categorical nature that makes these methods effective for DNA.

Multi-agent reinforcement learning approaches have transformed protein engineering workflows, as demonstrated by systems achieving 16-fold improvements in enzyme activity, yet similar agent-based approaches for DNA regulatory element design or RNA therapeutic development remain rudimentary. The absence is particularly notable given that regulatory genomics involves complex multi-objective optimization problems ideally suited for agent-based exploration. Similarly, the sophisticated value function learning and sequential Monte Carlo methods developed for protein design have seen limited application to nucleic acid engineering, despite these molecules presenting analogous design challenges.

Scaling challenges unique to biology further complicate cross-domain technique transfer. DNA sequences routinely span millions of base pairs, requiring architectural innovations like the StripedHyena blocks in

Evo [207] that achieve near-linear complexity. Yet attempts to scale protein models to similar context lengths for modeling entire proteomes or multi-protein systems remain limited by quadratic attention complexity. Conversely, the atomic-level precision required for protein function, now routine in models processing thousands of atoms, has proven difficult to achieve for large nucleic acid structures where computational requirements explode with sequence length.

8.2. Toward Universal Biological Models

The vision of universal biological models that seamlessly handle any molecular entity represents both the ultimate goal and greatest challenge in the field. Current multi-modal systems like AlphaFold3 [2] and ESM3 [105] demonstrate the feasibility of unified frameworks, yet they remain fundamentally limited in scope. A truly universal model would need to span from individual atoms to entire organisms, incorporating not just sequences and structures but also dynamics, interactions, and cellular context.

The central challenge lies in reconciling the vastly different scales, symmetries, and constraints that characterize different biological entities. Proteins exhibit $SE(3)$ symmetry and fold into compact structures, DNA forms long polymers with local regulatory logic and long-range chromatin interactions, while RNA adopts complex secondary structures with pseudoknots and tertiary contacts. A universal architecture must elegantly handle these disparate geometries without sacrificing the specialized inductive biases that make current domain-specific models successful. Early attempts at unification through shared tokenization schemes, as seen in Life-Code [177], show promise but struggle to maintain performance parity with specialized models.

Multi-agent alignment emerges as a promising paradigm for achieving universality without architectural compromises. Rather than forcing all biological modalities through a single model, specialized agents could maintain domain expertise while coordinating through learned communication protocols. This approach mirrors the modular organization of biological systems themselves, where specialized molecular machines collaborate to achieve complex functions. The BioDiscoveryAgent and similar systems demonstrate the viability of this approach, yet current implementations rely on rigid, hand-designed coordination mechanisms rather than learned, adaptive protocols.

Integration with existing biological knowledge systems presents another critical frontier. Current models largely ignore decades of accumulated biological knowledge encoded in databases, ontologies, and mechanistic models. While efforts like BioReason [74] begin to bridge this gap by incorporating pathway databases, the vast majority of structured biological knowledge remains inaccessible to generative models. A universal biological model must seamlessly integrate symbolic reasoning about pathways and interactions with generative modeling of sequences and structures.

The path toward artificial biological intelligence requires fundamental advances in how models represent and reason about biological causality. Current generative models excel at pattern recognition and interpolation within training distributions but struggle with extrapolation to novel biological contexts or reasoning about perturbations. True biological intelligence would require models that understand not just correlation but causation, enabling them to predict the effects of mutations, design entirely novel protein folds, or engineer synthetic biological circuits with predictable behavior.

8.3. Standardized Benchmarking and Evaluation

The limitations of unified benchmarks across biological modalities severely hampers progress assessment and method comparison. While protein folding has CASP [202]and antibody design has therapeutic bench-

marks [82], no equivalent comprehensive evaluations exist for DNA regulatory element design or RNA structure generation. This fragmentation makes it impossible to determine whether performance gaps between domains reflect fundamental difficulty differences or simply disparate research attention. The field urgently needs equivalents to ImageNet or GLUE that span all biological modalities with standardized tasks, metrics, and baselines.

Current evaluation metrics often poorly correlate with biological function [111], focusing on sequence similarity or structural accuracy while ignoring critical properties like stability, specificity, and evolvability. Perplexity and RMSD tell us little about whether a generated protein will fold or function in vivo. The community needs metrics that directly measure biological utility, such as expression levels for regulatory elements, catalytic efficiency for enzymes, binding affinity for aptamers. These functional metrics must be efficiently computable to enable large-scale evaluation, suggesting the need for learned surrogate models validated against experimental data.

Biological robustness evaluation [77] remains almost entirely absent from current benchmarking. Models may achieve high accuracy on test sets while remaining fragile to minor perturbations or distribution shifts that would be tolerated by natural biological systems. Systematic evaluation should include robustness to mutations, performance across organisms and conditions, and behavior under experimental noise. The lack of such evaluation is particularly concerning for models intended for therapeutic applications where safety and reliability are paramount.

8.4. Potential Innovations and Further Directions

Foundation agents for cell biology represent an ambitious frontier where generative models move beyond molecules to entire cells [309]. Such systems would integrate models of gene expression, protein interaction networks, metabolic pathways, and spatial organization to simulate and design at the cellular level. Early work on virtual cell models [269] provides a glimpse of this future, but current approaches lack the generative capabilities needed for de novo cell design or systematic perturbation analysis.

The combination of symbolic and generative modeling could unlock new capabilities in biological logic design [76, 322]. While current models excel at generating sequences that match statistical patterns, they struggle with the discrete logical operations that govern biological regulation. Hybrid neurosymbolic approaches could enable the design of genetic circuits with guaranteed behavioral properties, combining the creativity of generative models with the precision of formal verification. This fusion becomes particularly relevant for synthetic biology applications requiring predictable, modular biological components.

Continuous learning systems that improve through interaction with experimental data could accelerate the design-build-test-learn cycle fundamental to biological engineering [246, 347]. Current models are typically frozen after pretraining, missing opportunities to learn from the thousands of experimental validations performed by users. Online learning frameworks that safely incorporate experimental feedback while maintaining performance on established benchmarks could create ever-improving biological design assistants.

8.5. Interdisciplinary Collaboration

The realization of generative biology's full potential requires unprecedented collaboration between computational and experimental communities. Open science initiatives must extend beyond code sharing to include standardized experimental protocols, shared cell lines and reagents, and coordinated validation efforts. The success of AlphaFold [2] resulted not just from algorithmic innovation but from decades of structural biology creating training data, while similar community efforts are needed across all biological domains [211].

Educational initiatives must bridge the growing gap between cutting-edge AI methods and biological domain expertise. As models become more sophisticated, the barrier to entry rises for biologists seeking to apply these tools. Conversely, computer scientists often lack the biological intuition needed to identify meaningful problems and evaluate solutions. Integrated training programs that produce genuinely bilingual researchers, equally fluent in both computation and experimentation, will be essential for the field's continued progress.

The path forward requires not just technical innovation but a fundamental rethinking of how we organize biological research. The convergence documented throughout this study suggests that the traditional boundaries between genomics, proteomics, and other biological subdisciplines are becoming obsolete. As we move toward universal biological models and integrated experimental platforms, the field must evolve corresponding institutional structures, funding mechanisms, and evaluation criteria that support this interdisciplinary future. Only through such comprehensive transformation can generative biology fulfill its promise of revolutionizing our understanding and engineering of life.

9. Conclusion and Outlook

The convergence of generative AI techniques across biological domains represents both a tremendous opportunity and a call to action for the research community. By aligning the discussion around technical paradigms rather than molecular domains, we uncover unifying techniques, shared trends, and transferable insights that can catalyze progress across subfields. By adopting a methodological perspective that transcends traditional domain boundaries, we can accelerate progress toward truly general-purpose biological AI systems. The future of generative biology lies not in isolated advances within one biomolecular type, but in the recognition that these are simply different manifestations of a common computational challenge: understanding and generating the sequences and structures that encode life itself. We hope this work encourages collaborative model development and helps new researchers enter the field through a principled methodological lens.

Acknowledgments

The authors thank Dr. Aviv Regev for insightful discussions and valuable feedback. The authors thank Ms. Xiaomeng Fu for her professional graphic design support. S.J. gratefully acknowledges support from the National Science Foundation under grants IIS-2243850, MOMS-2331036, and CNS-2328395; the Advanced Research Projects Agency for Health under grant 1AY1AX000053; and the National Institutes of Health under grant U01AG070112. Additional support to S.J. was provided by the Texas A&M Institute of Data Science, the Truchard Family Endowed Chair, the Presidential Impact Fellowship, the Chancellor EDGES Fellowship at Texas A&M University, and Water Exceptional Item through Texas A&M AgriLife Research facilitated by the Texas Water Resources Institute. Y.X. and W.W. are supported by National Science Foundation (2106859, 2200274, 2312501), National Institutes of Health (U54HG012517, U24DK097771, U54OD036472, OT2OD038003, R01HL175135), Amazon, NEC, and Optum AI. B.-J.Y and X.Q. are supported by Advanced Research Projects Agency for Health under grant 1AY1AX000053. A.H. acknowledges support by the DOE Center for Advanced Bioenergy and Bioproducts Innovation (U.S. Department of Energy, Office of Science, Biological and Environmental Research Program under Award Number DE-SC0018420). Any opinions, findings, and conclusions or recommendations expressed in this publication are those of the author(s) and do not necessarily reflect the views of the U.S. Department of Energy.

References

[1] Hadi Abdine, Michail Chatzianastasis, Costas Bouyioukos, and Michalis Vazirgiannis. Prot2text: Multimodal protein's function generation with gnns and transformers. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 10757–10765, 2024.

[2] Josh Abramson, Jonas Adler, Jack Dunger, Richard Evans, Tim Green, Alexander Pritzel, Olaf Ronneberger, Lindsay Willmore, Andrew J Ballard, Joshua Bambrick, et al. Accurate structure prediction of biomolecular interactions with alphafold 3. *Nature*, 630(8016):493–500, 2024.

[3] Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. GPT-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023.

[4] Keir Adams, Kento Abeywardane, Jenna Fromer, and Connor W. Coley. ShEPhERD: Diffusing shape, electrostatics, and pharmacophores for bioisosteric drug design, 2024.

[5] Woody Ahern, Jason Yim, Doug Tischer, Saman Salike, Seth M Woodbury, Donghyo Kim, Indrek Kalvet, Yakov Kipnis, Brian Coventry, Han Raut Altae-Tran, et al. Atom level enzyme active site scaffolding using rfdiffusion2. *bioRxiv*, pages 2025–04, 2025.

[6] Ghada Al-Kateb, Emine Cengiz, and Murat Gök. Biogpt: A generative transformer-based framework for personalized genomic medicine and rare disease diagnosis. *Mesopotamian Journal of Artificial Intelligence in Healthcare*, 2025:154–164, 2025.

[7] Sarah Alamdari, Nitya Thakkar, Rianne van den Berg, Neil Tenenholtz, Bob Strome, Alan Moses, Alex Xijie Lu, Nicolo Fusi, Ava Pardis Amini, and Kevin K Yang. Protein generation with evolutionary diffusion: sequence is all you need. *BioRxiv*, pages 2023–09, 2023.

[8] Emily Alsentzer, John R Murphy, Willie Boag, Wei-Hung Weng, Di Jin, Tristan Naumann, and Matthew McDermott. Publicly available clinical bert embeddings. *arXiv preprint arXiv:1904.03323*, 2019.

[9] Namrata Anand, Raphael Eguchi, Irimpan I Mathews, Carla P Perez, Alexander Derry, Russ B Altman, and Po-Ssu Huang. Protein sequence design with a learned potential. *Nature communications*, 13(1):746, 2022.

[10] Rishabh Anand, Chaitanya K Joshi, Alex Morehead, Arian R Jamasb, Charles Harris, Simon V Mathis, Kieran Didi, Rex Ying, Bryan Hooi, and Pietro Liò. Rna-frameflow: Flow matching for de novo 3d rna backbone design. *ArXiv*, pages arXiv–2406, 2025.

[11] Jacob Austin, Daniel D Johnson, Jonathan Ho, Daniel Tarlow, and Rianne Van Den Berg. Structured denoising diffusion models in discrete state-spaces. *Advances in neural information processing systems*, 34:17981–17993, 2021.

[12] Žiga Avsec, Vikram Agarwal, Daniel Visentin, Joseph R Ledsam, Agnieszka Grabska-Barwinska, Kyle R Taylor, Yannis Assael, John Jumper, Pushmeet Kohli, and David R Kelley. Effective gene expression prediction from sequence by integrating long-range interactions. *Nature methods*, 18(10):1196–1203, 2021.

[13] Žiga Avsec, Natasha Latysheva, Jun Cheng, Guido Novati, Kyle R. Taylor, Tom Ward, Clare Bycroft, Lauren Nicolaisen, Eirini Arvaniti, Joshua Pan, Raina Thomas, Vincent Dutordoir, Matteo Perino, Soham De, Alexander Karollus, Adam Gayoso, Toby Sargeant, Anne Mottram, Lai Hong Wong, Pavol Drotár, Adam Kosiorek, Andrew Senior, Richard Tanburn, Taylor Applebaum, Souradeep Basu, Demis Hassabis, and Pushmeet Kohli. Alphagenome: advancing regulatory variant effect prediction with a unified dna sequence model. *ArXiV*, 2025. doi: https://doi.org/10.1101/2025.06.25.661532.

[14] Minkyung Baek, Frank DiMaio, Ivan Anishchenko, Justas Dauparas, Sergey Ovchinnikov, Gyu Rie Lee, Jue Wang, Qian Cong, Lisa N Kinch, R Dustin Schaeffer, et al. Accurate prediction of protein structures and interactions using a three-track neural network. *Science*, 373(6557):871–876, 2021.

[15] Minkyung Baek, Ryan McHugh, Ivan Anishchenko, David Baker, and Frank DiMaio. Accurate prediction of nucleic acid and protein-nucleic acid complexes using rosettafoldna. *bioRxiv*, pages 2022–09, 2022.

[16] Minkyung Baek, Ivan Anishchenko, Ian R Humphreys, Qian Cong, David Baker, and Frank DiMaio. Efficient and accurate prediction of protein structure using rosettafold2. *BioRxiv*, pages 2023–05, 2023.

[17] Minkyung Baek, Ryan McHugh, Ivan Anishchenko, Hanlun Jiang, David Baker, and Frank DiMaio. Accurate prediction of protein–nucleic acid complexes using rosettafoldna. *Nature methods*, 21(1):117–121, 2024.

[18] Amos Bairoch and Rolf Apweiler. The swiss-prot protein sequence database and its supplement trembl in 2000. *Nucleic acids research*, 28(1):45–48, 2000.

[19] Arpit Bansal, Hong-Min Chu, Avi Schwarzschild, Soumyadip Sengupta, Micah Goldblum, Jonas Geiping, and Tom Goldstein. Universal guidance for diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 843–852, 2023.

[20] Nathaniel R Bennett, Joseph L Watson, Robert J Ragotte, Andrew J Borst, DéJenaé L See, Connor Weidle, Riti Biswas, Yutong Yu, Ellen L Shrock, Russell Ault, et al. Atomically accurate de novo design of antibodies with rfdiffusion. *Nature*, pages 1–11, 2025.

[21] Suhaas Bhat, Kalyan Palepu, Lauren Hong, Joey Mao, Tianzheng Ye, Rema Iyer, Lin Zhao, Tianlai Chen, Sophia Vincoff, Rio Watson, et al. De novo design of peptide binders to conformationally diverse targets with contrastive language modeling. *Science Advances*, 11(4):eadr8638, 2025.

[22] Tommaso Biancalani, Gabriele Scalia, Lorenzo Buffoni, Raghav Avasthi, Ziqing Lu, Aman Sanger, Neriman Tokcan, Charles R Vanderburg, Åsa Segerstolpe, Meng Zhang, et al. Deep learning and alignment of spatially resolved single-cell transcriptomes with tangram. *Nature methods*, 18(11):1352–1362, 2021.

[23] Kevin Black, Michael Janner, Yilun Du, Ilya Kostrikov, and Sergey Levine. Training diffusion models with reinforcement learning. *arXiv preprint arXiv:2305.13301*, 2024.

[24] Montgomery Bohde, Mrunali Manjrekar, Runzhong Wang, Shuiwang Ji, and Connor W Coley. DiffMS: Diffusion generation of molecules conditioned on mass spectra. In *Proceedings of the International Conference on Machine Learning*, pages 4737–4756, 2025.

[25] Jacques Boitreaud, Jack Dent, Matthew McPartlon, Joshua Meier, Vinicius Reis, Alex Rogozhnikov, and Kevin Wu. Chai-1: Decoding the molecular interactions of life. *BioRxiv*, 2024.

[26] Avishek Joey Bose, Tara Akhound-Sadegh, Guillaume Huguet, Kilian Fatras, Jarrid Rector-Brooks, Cheng-Hao Liu, Andrei Cristian Nica, Maksym Korablyov, Michael Bronstein, and Alexander Tong. Se (3)-stochastic flow matching for protein backbone generation. *arXiv preprint arXiv:2310.02391*, 2023.

[27] Nadav Brandes, Dan Ofer, Yam Peleg, Nadav Rappoport, and Michal Linial. Proteinbert: a universal deep-learnign model of protein sequence and function. *Bioinformatics*, 2022.

[28] Nadav Brandes, Grant Goldman, Charlotte H Wang, Chun Jimmie Ye, and Vasilis Ntranos. Genome-wide prediction of disease variant effects with a deep protein language model. *Nature Genetics*, 55(9):1512–1522, 2023.

[29] Garyk Brixi, Matthew G Durrant, Jerome Ku, Michael Poli, Greg Brockman, Daniel Chang, Gabriel A Gonzalez, Samuel H King, David B Li, Aditi T Merchant, et al. Genome modeling and design across all domains of life with evo 2. *BioRxiv*, pages 2025–02, 2025.

[30] Tom B Brown. Language models are few-shot learners. *arXiv preprint arXiv:2005.14165*, 2020.

[31] Charlotte Bunne, Yusuf Roohani, Yanay Rosen, Ankit Gupta, Xikun Zhang, Marcel Roed, Theo Alexandrov, Mohammed AlQuraishi, Patricia Brennan, Daniel B Burkhardt, et al. How to build the virtual cell with artificial intelligence: Priorities and opportunities. *Cell*, 187(25):7045–7063, 2024.

[32] Andrew Campbell, Joe Benton, Valentin De Bortoli, Thomas Rainforth, George Deligiannidis, and Arnaud Doucet. A continuous time framework for discrete denoising models. *Advances in Neural Information Processing Systems*, 35:28266–28279, 2022.

[33] Andrew Campbell, Jason Yim, Regina Barzilay, Tom Rainforth, and Tommi Jaakkola. Generative flows on discrete state-spaces: Enabling multimodal flows with applications to protein co-design. *arXiv preprint arXiv:2402.04997*, 2024.

[34] Hanqun Cao, Haosen Shi, Chenyu Wang, Sinno Jialin Pan, and Pheng-Ann Heng. Glid^2 e: A gradient-free lightweight fine-tune approach for discrete sequence design. In *ICLR 2025 Workshop on Generative and Experimental Perspectives for Biomolecular Design*, 2025.

[35] He Cao, Zijing Liu, Xingyu Lu, Yuan Yao, and Yu Li. Instructmol: Multi-modal integration for building a versatile and reliable molecular assistant in drug discovery. *arXiv preprint arXiv:2311.16208*, 2023.

[36] Gabriel Cardoso, Yazid Janati El Idrissi, Sylvain Le Corff, and Eric Moulines. Monte carlo guided diffusion for bayesian linear inverse problems. *arXiv preprint arXiv:2308.07983*, 2023.

[37] Rich Caruana. Multitask learning. *Machine learning*, 28(1):41–75, 1997.

[38] Bo Chen, Xingyi Cheng, Pan Li, Yangli-ao Geng, Jing Gong, Shen Li, Zhilei Bei, Xu Tan, Boyan Wang, Xin Zeng, et al. xtrimopglm: unified 100b-scale pre-trained transformer for deciphering the language of protein. *arXiv preprint arXiv:2401.06199*, 2024.

[39] Chen Chen, Jie Hou, Xiaowen Shi, Hua Yang, James A Birchler, and Jianlin Cheng. Deepgrn: prediction of transcription factor binding site across cell-types using attention-based deep neural networks. *BMC bioinformatics*, 22(1):38, 2021.

[40] Jiayang Chen, Zhihang Hu, Siqi Sun, Qingxiong Tan, Yixuan Wang, Qinze Yu, Licheng Zong, Liang Hong, Jin Xiao, Tao Shen, et al. Interpretable rna foundation model from unannotated data for highly accurate rna structure and function predictions. *arXiv preprint arXiv:2204.00300*, 2022.

[41] Kathleen M. Chen, Aaron K. Wong, Olga G. Troyanskaya, and Jian Zhou. A sequence-based global map of regulatory activity for decifering human genetics. *Nature Genetics*, 2022.

[42] Leo Chen, Zachary Quinn, Madeleine Dumas, Christina Reng, auren Hong, Moises Lopez-Gonzalez, Alexander Mestre, Rio Watson, Sofia Vincoff, Lin Zhao, Jianli Wu, Audrey Stavrand, Mayumi Schaepers-Cheu, Tian Zi Wang, Divya Srijay, Connor Monticello, Pranay Vure, Rishab Pulugurta, Sarah Pertsemlidis, Kseniia Kholina, Shrey Goel, Matthew P DeLisa, Jen-Tsan Ashley Chi, Ray Truant, Hector C Aguilar, and Pranam Chatterjee. Target sequence-conditioned design of peptide binders using masked language modeling. *Nature Biotechnology*, 2025.

[43] Leo Tianlai Chen, Zachary Quinn, Madeleine Dumas, Christina Peng, Lauren Hong, Moises Lopez-Gonzalez, Alexander Mestre, Rio Watson, Sophia Vincoff, Lin Zhao, et al. Target sequence-conditioned design of peptide binders using masked language modeling. *Nature Biotechnology*, pages 1–9, 2025.

[44] Tong Chen, Yinuo Zhang, Sophia Tang, and Pranam Chatterjee. Multi-objective-guided discrete flow matching for controllable biological sequence design, 2025. URL https://arxiv.org/abs/2505.07086.

[45] Xingyu Chen, Shihao Ma, Runsheng Lin, Jiecong Lin, and Bo Wang. Ctrl-dna: Controllable cell-type-specific regulatory dna design via constrained rl. *arXiv preprint arXiv:2505.20578*, 2025.

[46] Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, et al. PaLM: Scaling language modeling with pathways. *Journal of Machine Learning Research*, 24(240):1–113, 2023.

[47] Alexandra Chronopoulou, Matthew E Peters, Alexander Fraser, and Jesse Dodge. Adaptersoup: Weight averaging to improve generalization of pretrained language models. *arXiv preprint arXiv:2302.07027*, 2023.

[48] Alexander E Chu, Jinho Kim, Lucy Cheng, Gina El Nesr, Minkai Xu, Richard W Shuai, and Po-Ssu Huang. An all-atom protein generative model. *Proceedings of the National Academy of Sciences*, 121(27):e2311500121, 2024.

[49] Yanyi Chu, Dan Yu, Yupeng Li, Kaixuan Huang, Yue Shen, Le Cong, Jason Zhang, and Mengdi Wang. A 5'utr language model for decoding untranslated regions of mrna and function predictions. *Nature Machine Intelligence*, 6(4):449–460, 2024.

[50] Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Yunxuan Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, et al. Scaling instruction-finetuned language models. *Journal of Machine Learning Research*, 25(70):1–53, 2024.

[51] Hyungjin Chung, Jeongsol Kim, Michael T Mccann, Marc L Klasky, and Jong Chul Ye. Diffusion posterior sampling for general noisy inverse problems. *arXiv preprint arXiv:2209.14687*, 2022.

[52] Gui Citovsky, Giulia DeSalvo, Claudio Gentile, Lazaros Karydas, Anand Rajagopalan, Afshin Rostamizadeh, and Sanjiv Kumar. Batch active learning at scale. *Advances in Neural Information Processing Systems*, 34: 11933–11944, 2021.

[53] Kevin Clark, Paul Vicol, Kevin Swersky, and David J Fleet. Directly fine-tuning diffusion models on differentiable rewards. *arXiv preprint arXiv:2309.17400*, 2023.

[54] Thomas Cokelaer, Dennis Pultz, Lea M Harder, Jordi Serra-Musach, and Julio Saez-Rodriguez. Bioservices: a common python package to access biological web services programmatically. *Bioinformatics*, 29(24):3241–3242, 2013.

[55] Ana Conesa, Pedro Madrigal, Sonia Tarazona, David Gomez-Cabrero, Alejandra Cervera, Andrew McPherson, Michał Wojciech Szcześniak, Daniel J Gaffney, Laura L Elo, Xuegong Zhang, et al. A survey of best practices for rna-seq data analysis. *Genome biology*, 17(1):13, 2016.

[56] Gabriele Corso, Hannes Stärk, Bowen Jing, Regina Barzilay, and Tommi Jaakkola. Diffdock: Diffusion steps, twists, and turns for molecular docking. *arXiv preprint arXiv:2210.01776*, 2022.

[57] Haotian Cui, Chloe Wang, Hassaan Maan, Kuan Pang, Fengning Luo, Nan Duan, and Bo Wang. scgpt: toward building a foundation model for single-cell multi-omics using generative ai. *Nature methods*, 21(8):1470–1480, 2024.

[58] Hugo Dalla-Torre, Liam Gonzalez, Javier Mendoza-Revilla, Nicolas Lopez Carranza, Adam Henryk Grzywaczewski, Francesco Oteri, Christian Dallago, Evan Trop, Bernardo P de Almeida, Hassan Sirelkhatim, et al. Nucleotide transformer: building and evaluating robust foundation models for human genomics. *Nature Methods*, 22(2):287–297, 2025.

[59] Lucas Ferreira DaSilva, Simon Senan, Zain Munir Patel, Aniketh Janardhan Reddy, Sameer Gabbita, Zach Nussbaum, César Miguel Valdez Córdova, Aaron Wenteler, Noah Weber, Tin M Tunjic, et al. Dna-diffusion: leveraging generative models for controlling chromatin accessibility and gene expression via synthetic regulatory elements. *Biorxiv*, 2024.

[60] Justas Dauparas, Ivan Anishchenko, Nathaniel Bennett, Hua Bai, Robert J Ragotte, Lukas F Milles, Basile IM Wicky, Alexis Courbet, Rob J de Haas, Neville Bethel, et al. Robust deep learning–based protein sequence design using proteinmpnn. *Science*, 378(6615):49–56, 2022.

[61] Justas Dauparas, Gyu Rie Lee, Robert Pecoraro, Linna An, Ivan Anishchenko, Cameron Glasscock, and David Baker. Atomic context-conditioned protein sequence design using ligandmpnn. *Nature Methods*, pages 1–7, 2025.

[62] Oscar Davis, Samuel Kessler, Mircea Petrache, Ismail Ceylan, Michael Bronstein, and Joey Bose. Fisher flow matching for generative modeling over discrete data. *Advances in Neural Information Processing Systems*, 37: 139054–139084, 2024.

[63] Jacob Devlin. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

[64] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies, volume 1 (long and short papers)*, pages 4171–4186, 2019.

[65] Prafulla Dhariwal and Alexander Nichol. Diffusion models beat gans on image synthesis. *Advances in neural information processing systems*, 34:8780–8794, 2021.

[66] Paolo Di Tommaso, Maria Chatzou, Evan W Floden, Pablo Prieto Barja, Emilio Palumbo, and Cedric Notredame. Nextflow enables reproducible computational workflows. *Nature biotechnology*, 35(4):316–319, 2017.

[67] Michael Dodds, Jeff Guo, Thomas Löhr, Alessandro Tibo, Ola Engkvist, and Jon Paul Janet. Sample efficient reinforcement learning with active learning for molecular design. *Chemical Science*, 15(11):4146–4160, 2024.

[68] Edo Dotan, Gal Jaschek, Tal Pupko, and Yonatan Belinkov. Effect of tokenization on transformers for biological sequences. *Bioinformatics*, 40(4):btae196, 2024.

[69] Zehao Dou and Yang Song. Diffusion posterior sampling for linear inverse problem solving: A filtering perspective. In *The Twelfth International Conference on Learning Representations*, 2024.

[70] Ian Dunn and David Ryan Koes. Mixed continuous and categorical flow matching for 3d de novo molecule generation. *ArXiv*, pages arXiv–2404, 2024.

[71] Alexandre Duval, Simon V Mathis, Chaitanya K Joshi, Victor Schmidt, Santiago Miret, Fragkiskos D Malliaros, Taco Cohen, Pietro Lio, Yoshua Bengio, and Michael Bronstein. A hitchhiker's guide to geometric gnns for 3d atomic systems. *arXiv preprint arXiv:2312.07511*, 2023.

[72] Ahmed Elnaggar, Michael Heinzinger, Christian Dallago, Ghalia Rehawi, Yu Wang, Llion Jones, Tom Gibbs, Tamas Feher, Christoph Angerer, Martin Steinegger, et al. Prottrans: Toward understanding the language of life through self-supervised learning. *IEEE transactions on pattern analysis and machine intelligence*, 44(10): 7112–7127, 2021.

[73] Adibvafa Fallahpour, Vincent Gureghian, Guillaume J Filion, Ariel B Lindner, and Amir Pandi. Codontransformer: a multispecies codon optimizer using context-aware neural networks. *Nature Communications*, 16(1):3205, 2025.

[74] Adibvafa Fallahpour, Andrew Magnuson, Purav Gupta, Shihao Ma, Jack Naimer, Arnav Shah, Haonan Duan, Omar Ibrahim, Hani Goodarzi, Chris J Maddison, et al. Bioreason: Incentivizing multimodal biological reasoning within a dna-llm model. *arXiv preprint arXiv:2505.23579*, 2025.

[75] Ying Fan, Olivia Watkins, Yuqing Du, Hao Liu, Moonkyung Ryu, Craig Boutilier, Pieter Abbeel, Mohammad Ghavamzadeh, Kangwook Lee, and Kimin Lee. DPOK: Reinforcement learning for fine-tuning text-to-image diffusion models. *arXiv preprint arXiv:2305.16381*, 2023.

[76] Meng Fang, Shilong Deng, Yudi Zhang, Zijing Shi, Ling Chen, Mykola Pechenizkiy, and Jun Wang. Large language models are neurosymbolic reasoners. In *Proceedings of the AAAI conference on artificial intelligence*, volume 38, pages 17985–17993, 2024.

[77] Marie-Anne Félix and Michalis Barkoulas. Pervasive robustness in biological systems. *Nature Reviews Genetics*, 16(8):483–496, 2015.

[78] Noelia Ferruz, Steffen Schmidt, and Birte Höcker. Protgpt2 is a deep unsupervised language model for protein design. *Nature communications*, 13(1):4348, 2022.

[79] Chelsea Finn, Pieter Abbeel, and Sergey Levine. Model-agnostic meta-learning for fast adaptation of deep networks. In *International conference on machine learning*, pages 1126–1135. PMLR, 2017.

[80] Cong Fu, Xiner Li, Blake Olson, Heng Ji, and Shuiwang Ji. Fragment and geometry aware tokenization of molecules for structure-based drug design using language models. *arXiv preprint arXiv:2408.09730*, 2024.

[81] Cong Fu, Keqiang Yan, Limei Wang, Wing Yee Au, Michael Curtis McThrow, Tao Komikado, Koji Maruhashi, Kanji Uchino, Xiaoning Qian, and Shuiwang Ji. A latent diffusion model for protein structure generation. In *Learning on graphs conference*, pages 29–1. PMLR, 2024.

[82] Aline Fuchs, Chantal Csajka, Yann Thoma, Thierry Buclin, and Nicolas Widmer. Benchmarking therapeutic drug monitoring software: a review of available computer tools. *Clinical pharmacokinetics*, 52(1):9–22, 2013.

[83] Yarin Gal, Riashat Islam, and Zoubin Ghahramani. Deep bayesian active learning with image data. In *International conference on machine learning*, pages 1183–1192. PMLR, 2017.

[84] Shanghua Gao, Ada Fang, Yepeng Huang, Valentina Giunchiglia, Ayush Noori, Jonathan Richard Schwarz, Yasha Ektefaie, Jovana Kondic, and Marinka Zitnik. Empowering biomedical discovery with ai agents. *Cell*, 187(22): 6125–6151, 2024.

[85] Shanghua Gao, Richard Zhu, Pengwei Sui, Zhenglun Kong, Sufian Aldogom, Yepeng Huang, Ayush Noori, Reza Shamji, Krishna Parvataneni, Theodoros Tsiligkaridis, et al. Democratizing AI scientists using ToolUniverse. *arXiv:2509.23426*, 2025.

[86] Zijing Gao, Qiao Liu, Wanwen Zeng, Rui Jiang, and Wing Hung Wong. Epigept: a pretrained transformer model for epigenomics. *bioRxiv*, pages 2023–07, 2024.

[87] Juan Jose Garau-Luis, Patrick Bordes, Liam Gonzalez, Maša Roller, Bernardo de Almeida, Christopher Blum, Lorenz Hexemer, Stefan Laurent, Maren Lang, Thomas Pierrot, et al. Multi-modal transfer learning between biological foundation models. *Advances in Neural Information Processing Systems*, 37:78431–78450, 2024.

[88] Itai Gat, Tal Remez, Neta Shaul, Felix Kreuk, Ricky TQ Chen, Gabriel Synnaeve, Yossi Adi, and Yaron Lipman. Discrete flow matching. *Advances in Neural Information Processing Systems*, 37:133345–133385, 2024.

[89] Itai Gat, Neta Shaul, Uriel Singer, and Yaron Lipman. Corrector sampling in language models. *arXiv preprint arXiv:2506.06215*, 2025.

[90] Tomas Geffner, Kieran Didi, Zuobai Zhang, Danny Reidenbach, Zhonglin Cao, Jason Yim, Mario Geiger, Christian Dallago, Emine Kucukbenli, Arash Vahdat, et al. Proteina: Scaling flow-based protein structure generative models. *arXiv preprint arXiv:2503.00710*, 2025.

[91] Haisong Gong, Qiang Liu, Shu Wu, and Liang Wang. Text-guided molecule generation with diffusion language model. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 109–117, 2024.

[92] Juraj Gottweis, Wei-Hung Weng, Alexander Daryin, Tao Tu, Anil Palepu, Petar Sirkovic, Artiom Myaskovsky, Felix Weissenberger, Keran Rong, Ryutaro Tanno, et al. Towards an ai co-scientist. *arXiv preprint arXiv:2502.18864*, 2025.

[93] Casper A Goverde, Martin Pacesa, Nicolas Goldbach, Lars J Dornfeld, Petra EM Balbi, Sandrine Georgeon, Stéphane Rosset, Srajan Kapoor, Jagrity Choudhury, Justas Dauparas, et al. Computational design of soluble and functional membrane protein analogues. *Nature*, 631(8020):449–458, 2024.

[94] Nate Gruver, Samuel Stanton, Nathan Frey, Tim GJ Rudner, Isidro Hotzel, Julien Lafrance-Vanasse, Arvind Rajpal, Kyunghyun Cho, and Andrew G Wilson. Protein design with guided discrete diffusion. *Advances in neural information processing systems*, 36:12489–12517, 2023.

[95] Albert Gu and Tri Dao. Mamba: Linear-time sequence modeling with selective state spaces. *arXiv preprint arXiv:2312.00752*, 2023.

[96] Albert Gu, Karan Goel, and Christopher Ré. Efficiently modeling long sequences with structured state spaces. *arXiv preprint arXiv:2111.00396*, 2021.

[97] Jiaqi Guan, Wesley Wei Qian, Xingang Peng, Yufeng Su, Jian Peng, and Jianzhu Ma. 3d equivariant diffusion for target-aware molecule generation and affinity prediction. *arXiv preprint arXiv:2303.03543*, 2023.

[98] Shurui Gui, Xiner Li, Limei Wang, and Shuiwang Ji. GOOD: A graph out-of-distribution benchmark. In *The 36th Annual Conference on Neural Information Processing Systems (Track on Datasets and Benchmarks)*, pages 2059–2073, 2022.

[99] Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, et al. Deepseek-r1: Incentivizing reasoning capability in LLMs via reinforcement learning. *arXiv preprint arXiv:2501.12948*, 2025.

[100] Jiaqi Han, Austin Wang, Minkai Xu, Wenda Chu, Meihua Dang, Yisong Yue, and Stefano Ermon. Discrete diffusion trajectory alignment via stepwise decomposition. *arXiv preprint arXiv:2507.04832*, 2025.

[101] Zeyu Han, Chao Gao, Jinyang Liu, Jeff Zhang, and Sai Qian Zhang. Parameter-efficient fine-tuning for large models: A comprehensive survey. *arXiv preprint arXiv:2403.14608*, 2024.

[102] Minsheng Hao, Jing Gong, Xin Zeng, Chiming Liu, Yucheng Guo, Xingyi Cheng, Taifeng Wang, Jianzhu Ma, Xuegong Zhang, and Le Song. Large-scale foundation model on single-cell transcriptomics. *Nature methods*, 21 (8):1481–1491, 2024.

[103] Majdi Hassan, Nikhil Shenoy, Jungyoon Lee, Hannes Stärk, Stephan Thaler, and Dominique Beaini. Et-flow: Equivariant flow-matching for molecular conformer generation. *Advances in Neural Information Processing Systems*, 37:128798–128824, 2024.

[104] Etrit Haxholli, Yeti Z Gürbüz, Oğul Can, and Eli Waxman. Efficient perplexity bound and ratio matching in discrete diffusion language models. *arXiv preprint arXiv:2507.04341*, 2025.

[105] Thomas Hayes, Roshan Rao, Halil Akin, Nicholas J Sofroniew, Deniz Oktay, Zeming Lin, Robert Verkuil, Vincent Q Tran, Jonathan Deaton, Marius Wiggert, et al. Simulating 500 million years of evolution with a language model. *Science*, 387(6736):850–858, 2025.

[106] Yong He, Pan Fang, Yongtao Shan, Yuanfei Pan, Yanhong Wei, Yichang Chen, Yihao Chen, Yi Liu, Zhenyu Zeng, Zhan Zhou, et al. Generalized biological foundation model with unified nucleic acid and protein language. *Nature Machine Intelligence*, pages 1–12, 2025.

[107] Zhengfu He, Tianxiang Sun, Kuanning Wang, Xuanjing Huang, and Xipeng Qiu. Diffusionbert: Improving generative masked language models with diffusion models. *arXiv preprint arXiv:2211.15029*, 2022.

[108] Michael Heinzinger, Konstantin Weissenow, Joaquin Gomez Sanchez, Adrian Henkel, Milot Mirdita, Martin Steinegger, and Burkhard Rost. Bilingual language model for protein sequence and structure. *NAR Genomics and Bioinformatics*, 6(4):lqae150, September 2024. ISSN 2631-9268. doi: 10.1093/nargab/lqae150. URL https://academic.oup.com/nargab/article/doi/10.1093/nargab/lqae150/7901286.

[109] Jeremy Heng, Adrian N Bishop, George Deligiannidis, and Arnaud Doucet. Controlled sequential monte carlo. *The Annals of Statistics*, 48(5):2904–2929, 2020.

[110] Daniel Hesslow, Niccoló Zanichelli, Pascal Notin, Iacopo Poli, and Debora Marks. Rita: a study on scaling up generative protein sequence models. *arXiv preprint arXiv:2205.05789*, 2022.

[111] Steven A Hicks, Inga Strümke, Vajira Thambawita, Malek Hammou, Michael A Riegler, Pål Halvorsen, and Sravanthi Parasa. On evaluation metrics for medical applications of artificial intelligence. *Scientific reports*, 12 (1):5979, 2022.

[112] Jonathan Ho and Tim Salimans. Classifier-free diffusion guidance. *arXiv preprint arXiv:2207.12598*, 2022.

[113] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020.

[114] Jonathan Ho, Tim Salimans, Alexey Gritsenko, William Chan, Mohammad Norouzi, and David J Fleet. Video diffusion models. *Advances in Neural Information Processing Systems*, 35:8633–8646, 2022.

[115] Emiel Hoogeboom, Didrik Nielsen, Priyank Jaini, Patrick Forré, and Max Welling. Argmax flows and multinomial diffusion: Learning categorical distributions. *Advances in neural information processing systems*, 34:12454–12465, 2021.

[116] Emiel Hoogeboom, Vıctor Garcia Satorras, Clément Vignac, and Max Welling. Equivariant diffusion for molecule generation in 3d. In *International conference on machine learning*, pages 8867–8887. PMLR, 2022.

[117] Chloe Hsu, Robert Verkuil, Jason Liu, Zeming Lin, Brian Hie, Tom Sercu, Adam Lerer, and Alexander Rives. Learning inverse folding from millions of predicted structures. In *International conference on machine learning*, pages 8946–8970. PMLR, 2022.

[118] Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*, 2021.

[119] Siquan Hu, Ruixiong Ma, and Haiou Wang. An improved deep learning method for predicting dna-binding proteins based on contextual features in amino acid sequences. *PLoS one*, 14(11):e0225317, 2019.

[120] Han Huang, Ziqian Lin, Dongchen He, Liang Hong, and Yu Li. Ribodiffusion: tertiary structure-based rna inverse folding with generative diffusion models. *Bioinformatics*, 40(Supplement_1):i347–i356, 2024.

[121] Kaixuan Huang, Yukang Yang, Kaidi Fu, Yanyi Chu, Le Cong, and Mengdi Wang. Latent diffusion models for controllable rna sequence generation. *arXiv preprint arXiv:2409.09828*, 2024.

[122] Kexin Huang, Jaan Altosaar, and Rajesh Ranganath. Clinicalbert: Modeling clinical notes and predicting hospital readmission. *arXiv preprint arXiv:1904.05342*, 2019.

[123] Kexin Huang, Romain Lopez, Jan-Christian Hütter, Takamasa Kudo, Antonio Rios, and Aviv Regev. Sequential optimal experimental design of perturbation screens guided by multi-modal priors. In *International Conference on Research in Computational Molecular Biology*, pages 17–37. Springer, 2024.

[124] Kexin Huang, Serena Zhang, Hanchen Wang, Yuanhao Qu, Yingzhou Lu, Yusuf Roohani, Ryan Li, Lin Qiu, Junze Zhang, Yin Di, et al. Biomni: A general-purpose biomedical ai agent. *bioRxiv*, pages 2025–05, 2025.

[125] Yue Huang, Lichao Sun, Haoran Wang, Siyuan Wu, Qihui Zhang, Yuan Li, Chujie Gao, Yixin Huang, Wenhan Lyu, Yixuan Zhang, et al. Trustllm: Trustworthiness in large language models. *arXiv preprint arXiv:2401.05561*, 2024.

[126] Guillaume Huguet, James Vuckovic, Kilian Fatras, Eric Thibodeau-Laufer, Pablo Lemos, Riashat Islam, Chenghao Liu, Jarrid Rector-Brooks, Tara Akhound-Sadegh, Michael Bronstein, et al. Sequence-augmented se (3)-flow matching for conditional protein generation. *Advances in neural information processing systems*, 37:33007–33036, 2024.

[127] Ilia Igashov, Hannes Stärk, Clément Vignac, Victor Garcia Satorras, Pascal Frossard, Max Welling, Michael Bronstein, and Bruno Correia. Equivariant 3d-conditional diffusion models for molecular linker design. *arXiv preprint arXiv:2210.05274*, 2022.

[128] Musa Nuri İhtiyar and Arzucan Özgür. Generative language models on nucleotide sequences of human genes. *Scientific Reports*, 14(1):22204, 2024.

[129] John Ingraham, Vikas Garg, Regina Barzilay, and Tommi Jaakkola. Generative models for graph-based protein design. *Advances in neural information processing systems*, 32, 2019.

[130] John B Ingraham, Max Baranov, Zak Costello, Karl W Barber, Wujie Wang, Ahmed Ismail, Vincent Frappier, Dana M Lord, Christopher Ng-Thow-Hing, Erik R Van Vlack, et al. Illuminating protein space with a programmable generative model. *Nature*, 623(7989):1070–1078, 2023.

[131] Kishor Jaganathan, Sofia Kyriazopoulou Panagiotopoulou, Jeremy F McRae, Darbandi Sivash Fazel, David Knowles, Yang I Li, Jack A Kosmicki, Juan Arbelaez, Wenwu Cui, Grace B Schwartz, Eric D Chow, Efstathios Kanterakis, Hong Gao, Amirali Kia, Serafim Batzoglou, Stephan J Sanders, and Kyle Kai-How Farh. Predicting splicing from primary sequence with deep learning. *Cell*, 2019.

[132] Nauman Javed, Thomas Weingarten, Arijit Sehanobish, Adam Roberts, Avinava Dubey, Krzysztof Choromanski, and Bradley E Bernstein. A multi-modal transformer for cell type-agnostic regulatory predictions. *Cell Genomics*, 5(2), 2025.

[133] Yanrong Ji, Zhihan Zhou, Han Liu, and Ramana V Davuluri. Dnabert: pre-trained bidirectional encoder representations from transformers model for dna-language in genome. *Bioinformatics*, 37(15):2112–2120, 2021.

[134] Jiyue Jiang, Zikang Wang, Yuheng Shan, Heyan Chai, Jiayi Li, Zixian Ma, Xinrui Zhang, and Yu Li. Biological sequence with language model prompting: A survey. *arXiv preprint arXiv:2503.04135*, 2025.

[135] Kaiyi Jiang, Zhaoqing Yan, Matteo Di Bernardo, Samantha R Sgrizzi, Lukas Villiger, Alisan Kayabolen, BJ Kim, Josephine K Carscadden, Masahiro Hiraizumi, Hiroshi Nishimasu, et al. Rapid in silico directed evolution by a protein language model with evolvepro. *Science*, 387(6732):eadr6006, 2024.

[136] Wengong Jin, Regina Barzilay, and Tommi Jaakkola. Junction tree variational autoencoder for molecular graph generation. In *International conference on machine learning*, pages 2323–2332. PMLR, 2018.

[137] Bowen Jing, Gabriele Corso, Jeffrey Chang, Regina Barzilay, and Tommi Jaakkola. Torsional diffusion for molecular conformer generation. *Advances in Neural Information Processing Systems*, 35:24240–24253, 2022.

[138] Bowen Jing, Bonnie Berger, and Tommi Jaakkola. Alphafold meets flow matching for generating protein ensembles. *arXiv preprint arXiv:2402.04845*, 2024.

[139] Jaehyeong Jo, Seul Lee, and Sung Ju Hwang. Score-based generative modeling of graphs via the system of stochastic differential equations. In *International Conference on Machine Learning*, pages 10362–10383. PMLR, 2022.

[140] Chaitanya K Joshi, Xiang Fu, Yi-Lun Liao, Vahe Gharakhanyan, Benjamin Kurt Miller, Anuroop Sriram, and Zachary W Ulissi. All-atom diffusion transformers: Unified generative modelling of molecules and materials. *arXiv preprint arXiv:2503.03965*, 2025.

[141] John Jumper, Richard Evans, Alexander Pritzel, Tim Green, Michael Figurnov, Olaf Ronneberger, Kathryn Tunyasuvunakool, Russ Bates, Augustin Žídek, Anna Potapenko, et al. Highly accurate protein structure prediction with alphafold. *nature*, 596(7873):583–589, 2021.

[142] Tero Karras, Miika Aittala, Tuomas Kynkäänniemi, Jaakko Lehtinen, Timo Aila, and Samuli Laine. Guiding a diffusion model with a bad version of itself. *Advances in Neural Information Processing Systems*, 37:52996–53021, 2024.

[143] Mohamed Amine Ketata, Cedrik Laue, Ruslan Mammadov, Hannes Stärk, Menghua Wu, Gabriele Corso, Céline Marquet, Regina Barzilay, and Tommi S Jaakkola. Diffdock-pp: Rigid protein-protein docking with diffusion models. *arXiv preprint arXiv:2304.03889*, 2023.

[144] Leon Klein, Andreas Krämer, and Frank Noe. Equivariant flow matching. In *Thirty-seventh Conference on Neural Information Processing Systems*.

[145] Johannes Köster and Sven Rahmann. Snakemake—a scalable bioinformatics workflow engine. *Bioinformatics*, 28(19):2520–2522, 2012.

[146] Rohith Krishna, Jue Wang, Woody Ahern, Pascal Sturmfels, Preetham Venkatesh, Indrek Kalvet, Gyu Rie Lee, Felix S Morey-Burrows, Ivan Anishchenko, Ian R Humphreys, et al. Generalized biomolecular modeling and design with rosettafold all-atom. *Science*, 384(6693):eadl2528, 2024.

[147] Jerome Ku, Eric Nguyen, David W. Romero, Garyk Brixi, Brandon Yang, Anton Vorontsov, Ali Taghibakhshi, Amy X. Lu, Dave P. Burke, Greg Brockman, Stefano Massaroli, Christopher Ré, Patrick D Hsu, Brian L Hie, Stefano Ermon, and Michael Poli. Systems and algorithms for convolutional multi-hybrid langauge models at scale. *ArXiv*, 2025. doi: 10.48550/arXiv.2503.01868.

[148] Hiroyuki Kurata and Sho Tsukiyama. Ican: interpretable cross-attention network for identifying drug and target protein interactions. *Plos one*, 17(10):e0276609, 2022.

[149] Avantika Lal, David Garfield, Tommaso Biancalani, et al. reglm: Designing realistic regulatory dna with autoregressive language models. *bioRxiv*, pages 2024–02, 2024.

[150] Jin Sub Lee and Philip M Kim. Flowpacker: protein side-chain packing with torsional flow matching. *Bioinformatics*, 41(3):btaf010, 2025.

[151] Seul Lee, Karsten Kreis, Srimukh Prasad Veccham, Meng Liu, Danny Reidenbach, Yuxing Peng, Saee Paliwal, Weili Nie, and Arash Vahdat. Genmol: A drug discovery generalist with discrete diffusion. *arXiv preprint arXiv:2501.06158*, 2025.

[152] Francesca-Zhoufan Li, Ava P Amini, Yisong Yue, Kevin K Yang, and Alex Xijie Lu. Feature reuse and scaling: Understanding transfer learnign with protein language models. *Proceedings of the 41st International Conference on Machine Learning, PLMR*, 2024.

[153] Kongming Li, Jiahao Li, Yuhao Tao, and Fei Wang. stdiff: a diffusion model for imputing spatial transcriptomics through single-cell transcriptomics. *Briefings in Bioinformatics*, 25(3), 2024.

[154] Mingchen Li, Yang Tan, Xinzhu Ma, Bozitao Zhong, Huiqun Yu, Ziyi Zhou, Wanli Ouyang, Bingxin Zhou, Pan Tan, and Liang Hong. Prosst: Protein language modeling with quantized structure and disentangled attention. *Advances in Neural Information Processing Systems*, 37:35700–35726, 2024.

[155] Qing Li, Zhihang Hu, Yixuan Wang, Lei Li, Yimin Fan, Irwin King, Gengjie Jia, Sheng Wang, Le Song, and Yu Li. Progress and opportunities of foundation models in bioinformatics. *Briefings in Bioinformatics*, 25(6):bbae548, 2024.

[156] Sizhen Li, Saeed Moayedpour, Ruijiang Li, Michael Bailey, Saleh Riahi, Lorenzo Kogler-Anele, Milad Miladi, Jacob Miner, Dinghai Zheng, Jun Wang, et al. Codonbert: Large language models for mrna design and optimization. *bioRxiv*, pages 2023–09, 2023.

[157] Xiang Li, John Thickstun, Ishaan Gulrajani, Percy S Liang, and Tatsunori B Hashimoto. Diffusion-lm improves controllable text generation. *Advances in neural information processing systems*, 35:4328–4343, 2022.

[158] Xiang Lisa Li and Percy Liang. Prefix-tuning: Optimizing continuous prompts for generation. *arXiv preprint arXiv:2101.00190*, 2021.

[159] Xiner Li, Shurui Gui, Youzhi Luo, and Shuiwang Ji. Graph structure extrapolation for out-of-distribution generalization. In *Forty-first International Conference on Machine Learning*, 2024.

[160] Xiner Li, Limei Wang, Youzhi Luo, Carl Edwards, Shurui Gui, Yuchao Lin, Heng Ji, and Shuiwang Ji. Geometry informed tokenization of molecules for language model generation. *arXiv preprint arXiv:2408.10120*, 2024.

[161] Xiner Li, Yulai Zhao, Chenyu Wang, Gabriele Scalia, Gokcen Eraslan, Surag Nair, Tommaso Biancalani, Aviv Regev, Sergey Levine, and Masatoshi Uehara. Derivative-free guidance in continuous and discrete diffusion models with soft value-based decoding. *arXiv preprint arXiv:2408.08252*, 2024.

[162] Xiner Li, Masatoshi Uehara, Xingyu Su, Gabriele Scalia, Tommaso Biancalani, Aviv Regev, Sergey Levine, and Shuiwang Ji. Dynamic search for inference-time alignment in diffusion models. *arXiv preprint arXiv:2503.02039*, 2025.

[163] Zehui Li, Yuhao Ni, William AV Beardall, Guoxuan Xia, Akashaditya Das, Guy-Bart Stan, and Yiren Zhao. Discdiff: Latent diffusion model for dna sequence generation. *arXiv preprint arXiv:2402.06079*, 2024.

[164] Wang Liang. Llama-gene: A general-purpose gene task large language model based on instruction fine-tuning. *arXiv preprint arXiv:2412.00471*, 2024.

[165] Hunter Lightman, Vineet Kosaraju, Yuri Burda, Harrison Edwards, Bowen Baker, Teddy Lee, Jan Leike, John Schulman, Ilya Sutskever, and Karl Cobbe. Let's verify step by step. In *The Twelfth International Conference on Learning Representations*, 2023.

[166] Zeming Lin, Halil Akin, Roshan Rao, Brian Hie, Zhongkai Zhu, Wenting Lu, Nikita Smetanin, Robert Verkuil, Ori Kabeli, Yaniv Shmueli, et al. Evolutionary-scale prediction of atomic-level protein structure with a language model. *Science*, 379(6637):1123–1130, 2023.

[167] Johannes Linder, Divyanshi Srivastava, Han Yuan, Vikram Agarwal, and David R. Kelley. Predicting rna-seq coverage from dna sequence as a unifying model of gene regulation. *Nature Genetics*, 2025.

[168] LeAnn M Lindsey, Nicole L Pershing, Anisa Habib, Keith Dufault-Thompson, W Zac Stephens, Anne J Blaschke, Xiaofang Jiang, and Hari Sundar. The impact of tokenizer selection in genomic language models. *Bioinformatics*, page btaf456, 2025.

[169] Yaron Lipman, Ricky TQ Chen, Heli Ben-Hamu, Maximilian Nickel, and Matt Le. Flow matching for generative modeling. *ICLR 2023*, 2023.

[170] Sidney Lyayuga Lisanza, Jacob Merle Gershon, Samuel WK Tipps, Jeremiah Nelson Sims, Lucas Arnoldt, Samuel J Hendel, Miriam K Simma, Ge Liu, Muna Yase, Hongwei Wu, et al. Multistate and functional protein design using rosettafold sequence space diffusion. *Nature biotechnology*, 43(8):1288–1298, 2025.

[171] Gang Liu, Jiaxin Xu, Tengfei Luo, and Meng Jiang. Graph diffusion transformers for multi-conditional molecular generation. *Advances in Neural Information Processing Systems*, 37:8065–8092, 2024.

[172] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *Advances in neural information processing systems*, 36:34892–34916, 2023.

[173] Junhao Liu, Pengpeng Zhang, Siwei Xu, Shushrruth Sai Srinivasan, Yongxian Wu, and Jing Zhang. Multimodal cell context instruction tuning for conditional dna regulatory sequence generation with large language models. *IEEE International Conference on Image Processing*, 2025.

[174] Shengchao Liu, Yanjing Li, Zhuoxinran Li, Anthony Gitter, Yutao Zhu, Jiarui Lu, Zhao Xu, Weili Nie, Arvind Ramanathan, Chaowei Xiao, et al. A text-guided protein design framework. *Nature Machine Intelligence*, pages 1–12, 2025.

[175] Tianyu Liu, Edward De Brouwer, Tony Kuo, Nathaniel Diamant, Alsu Missarova, Hanchen Wang, Minsheng Hao, Hector Corrada Bravo, Gabriele Scalia, Aviv Regev, et al. Learning multi-cellular representations of single-cell transcriptomics data enables characterization of patient-level disease states. In *International Conference on Research in Computational Molecular Biology*, pages 303–306. Springer, 2025.

[176] Zhiyuan Liu, An Zhang, Hao Fei, Enzhi Zhang, Xiang Wang, Kenji Kawaguchi, and Tat-Seng Chua. Prott3: Protein-to-text generation for text-based protein understanding. *arXiv preprint arXiv:2405.12564*, 2024.

[177] Zicheng Liu, Siyuan Li, Zhiyuan Chen, Fang Wu, Chang Yu, Qirong Yang, Yucheng Guo, Yujie Yang, Xiaoming Zhang, and Stan Z Li. Life-code: Central dogma modeling with multi-omics sequence unification. *arXiv preprint arXiv:2502.07299*, 2025.

[178] Ziteng Liu, Liqiang Lin, Qingqing Jia, Zheng Cheng, Yanyan Jiang, Yanwen Guo, and Jing Ma. Transferable multilevel attention neural network for accurate prediction of quantum chemistry properties via multitask learning. *Journal of chemical information and modeling*, 61(3):1066–1082, 2021.

[179] Mingsheng Long, Yue Cao, Jianmin Wang, and Michael Jordan. Learning transferable features with deep adaptation networks. In *International conference on machine learning*, pages 97–105. PMLR, 2015.

[180] Mingsheng Long, Han Zhu, Jianmin Wang, and Michael I Jordan. Deep transfer learning with joint adaptation networks. In *International conference on machine learning*, pages 2208–2217. PMLR, 2017.

[181] Romain Lopez, Jeffrey Regier, Michael B Cole, Michael I Jordan, and Nir Yosef. Deep generative modeling for single-cell transcriptomics. *Nature methods*, 15(12):1053–1058, 2018.

[182] Aaron Lou, Chenlin Meng, and Stefano Ermon. Discrete diffusion modeling by estimating the ratios of the data distribution. *arXiv preprint arXiv:2310.16834*, 2023.

[183] Weizhong Lu, Ye Tang, Hongjie Wu, Hongmei Huang, Qiming Fu, Jing Qiu, and Haiou Li. Predicting rna secondary structure via adaptive deep recurrent neural networks with energy-based filter. *BMC bioinformatics*, 20(Suppl 25):684, 2019.

[184] Weizhong Lu, Xiaoyi Chen, Yu Zhang, Hongjie Wu, Yijie Ding, Jiawei Shen, Shixuan Guan, and Haiou Li. Research on dna-binding protein identification method based on lstm-cnn feature fusion. *Computational and Mathematical Methods in Medicine*, 2022(1):9705275, 2022.

[185] Erpai Luo, Minsheng Hao, Lei Wei, and Xuegong Zhang. scdiffusion: conditional generation of high-quality single-cell data using diffusion model. *Bioinformatics*, 40(9):btae518, 2024.

[186] Shitong Luo, Yufeng Su, Xingang Peng, Sheng Wang, Jian Peng, and Jianzhu Ma. Antigen-specific antibody design and optimization with diffusion-based generative models for protein structures. *Advances in Neural Information Processing Systems*, 35:9754–9767, 2022.

[187] Yanchen Luo, Junfeng Fang, Sihang Li, Zhiyuan Liu, Jiancan Wu, An Zhang, Wenjie Du, and Xiang Wang. Text-guided diffusion model for 3d molecule generation. *arXiv preprint arXiv:2410.03803*, 2024.

[188] Yizhen Luo, Kai Yang, Massimo Hong, Xing Yi Liu, and Zaiqing Nie. Molfm: A multimodal molecular foundation model. *arXiv preprint arXiv:2307.09484*, 2023.

[189] Changze Lv, Jiang Zhou, Siyu Long, Lihao Wang, Jiangtao Feng, Dongyu Xue, Yu Pei, Hao Wang, Zherui Zhang, Yuchen Cai, et al. Amix-1: A pathway to test-time scalable protein foundation model. *arXiv preprint arXiv:2507.08920*, 2025.

[190] Liuzhenghao Lv, Zongying Lin, Hao Li, Yuyang Liu, Jiaxi Cui, Calvin Yu-Chian Chen, Li Yuan, and Yonghong Tian. Prollama: A protein large language model for multi-task protein language processing. *IEEE Transactions on Artificial Intelligence*, 2025.

[191] Andres M. Bran, Sam Cox, Oliver Schilter, Carlo Baldassari, Andrew D White, and Philippe Schwaller. Augmenting large language models with chemistry tools. *Nature Machine Intelligence*, 6(5):525–535, 2024.

[192] Mingqian Ma, Guoqing Liu, Chuan Cao, Pan Deng, Tri Dao, Albert Gu, Peiran Jin, Zhao Yang, Yingce Xia, Renqian Luo, et al. Hybridna: A hybrid transformer-mamba2 long-range dna language model. *arXiv preprint arXiv:2502.10807*, 2025.

[193] Nanye Ma, Shangyuan Tong, Haolin Jia, Hexiang Hu, Yu-Chuan Su, Mingda Zhang, Xuan Yang, Yandong Li, Tommi Jaakkola, Xuhui Jia, et al. Inference-time scaling for diffusion models beyond scaling denoising steps. *arXiv preprint arXiv:2501.09732*, 2025.

[194] Ali Madani, Bryan McCann, Nikhil Naik, Nitish Shirish Keskar, Namrata Anand, Raphael R Eguchi, Po-Ssu Huang, and Richard Socher. Progen: Language modeling for protein generation. *arXiv preprint arXiv:2004.03497*, 2020.

[195] Ali Madani, Ben Krause, Eric R Greene, Subu Subramanian, Benjamin P Mohr, James M Holton, Jose Luis Olmos Jr, Caiming Xiong, Zachary Z Sun, Richard Socher, et al. Large language models generate functional protein sequences across diverse families. *Nature biotechnology*, 41(8):1099–1106, 2023.

[196] Shae M Mclaughlin, Sajad H Ahanger, and Daniel A Lim. Nucleotide gpt: Sequence-based deep learning prediction of nuclear subcompartment-associated genome architecture. *bioRxiv*, pages 2024–11, 2024.

[197] Joshua Meier, Roshan Rao, Robert Verkuil, Jason Liu, Tom Sercu, and Alex Rives. Language models enable zero-shot prediction of the effects of mutations on protein function. *Advances in neural information processing systems*, 34:29287–29303, 2021.

[198] Zhichao Miao and Eric Westhof. Rna structure: advances and assessment of 3d structure prediction. *Annual review of biophysics*, 46(1):483–503, 2017.

[199] Pouria Mistani and Venkatesh Mysore. Preference optimization of protein language models as a multi-objective binder design paradigm. *arXiv preprint arXiv:2403.04187*, 2024.

[200] Alex Morehead and Jianlin Cheng. Flowdock: Geometric flow matching for generative protein-ligand docking and affinity prediction. *ArXiv*, pages arXiv–2412, 2025.

[201] Alex Morehead, Jeffrey Ruffolo, Aadyot Bhatnagar, and Ali Madani. Towards joint sequence-structure generation of nucleic acid and protein complexes with se (3)-discrete diffusion. *arXiv preprint arXiv:2401.06151*, 2023.

[202] John Moult. A decade of casp: progress, bottlenecks and prognosis in protein structure prediction. *Current opinion in structural biology*, 15(3):285–289, 2005.

[203] Geraldene Munsamy, Ramiro Illanes-Vicioso, Silvia Funcillo, Ioanna T Nakou, Sebastian Lindner, Gavin Ayres, Lesley S Sheehan, Steven Moss, Ulrich Eckhard, Philipp Lorenz, et al. Conditional language models enable the efficient design of proficient enzymes. *bioRxiv*, pages 2024–05, 2024.

[204] Ofir Nachum, Mohammad Norouzi, Kelvin Xu, and Dale Schuurmans. Bridging the gap between value and policy based reinforcement learning. *Advances in neural information processing systems*, 30, 2017.

[205] Sahin Naqvi, Seungsoo Kim, Saman Tabatabaee, Anusri Pampari, Anshul Kundaje, Jonathan K Pritchard, and Joanna Wysocka. Transfer learning reveals sequence determinants of the quantitative response to transcription factor dosage. *Cell Genomics*, 5(3), 2025.

[206] Eric Nguyen, Michael Poli, Marjan Faizi, Armin Thomas, Michael Wornow, Callum Birch-Sykes, Stefano Massaroli, Aman Patel, Clayton Rabideau, Yoshua Bengio, et al. Hyenadna: Long-range genomic sequence modeling at single nucleotide resolution. *Advances in neural information processing systems*, 36:43177–43201, 2023.

[207] Eric Nguyen, Michael Poli, Matthew G Durrant, Brian Kang, Dhruva Katrekar, David B Li, Liam J Bartie, Armin W Thomas, Samuel H King, Garyk Brixi, et al. Sequence modeling and design from molecular to genome scale with evo. *Science*, 386(6723):eado9336, 2024.

[208] Erik Nijkamp, Jeffrey A Ruffolo, Eli N Weinstein, Nikhil Naik, and Ali Madani. Progen2: exploring the boundaries of protein language models. *Cell systems*, 14(11):968–978, 2023.

[209] Hunter Nisonoff, Junhao Xiong, Stephan Allenspach, and Jennifer Listgarten. Unlocking guidance for discrete state-space diffusion and flow models. *arXiv preprint arXiv:2406.01572*, 2024.

[210] Divya Nori and Wengong Jin. Rnaflow: Rna structure & sequence design via inverse folding-based flow matching. *arXiv preprint arXiv:2405.18768*, 2024.

[211] Ruth Nussinov, Mingzhen Zhang, Yonglan Liu, and Hyunbum Jang. Alphafold, artificial intelligence (ai), and allostery. *The Journal of Physical Chemistry B*, 126(34):6372–6383, 2022.

[212] Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. Training language models to follow instructions with human feedback. *Advances in Neural Information Processing Systems*, 35:27730–27744, 2022.

[213] Martin Pacesa, Lennart Nickel, Christian Schellhaas, Joseph Schmidt, Ekaterina Pyatova, Lucas Kissling, Patrick Barendse, Jagrity Choudhury, Srajan Kapoor, Ana Alcaraz-Serna, et al. One-shot design of functional protein binders with bindcraft. *Nature*, 646(8084):483–492, 2025.

[214] Kuan Pang, Yanay Rosen, Kasia Kedzierska, Ziyuan He, Abhe Rajagopal, Claire E Gustafson, Grace Huynh, and Jure Leskovec. Pulsar: a foundation model for multi-scale and multicellular biology. *bioRxiv*, pages 2025–11, 2025.

[215] Shubham Parashar, Shurui Gui, Xiner Li, Hongyi Ling, Sushil Vemuri, Blake Olson, Eric Li, Yu Zhang, James Caverlee, Dileep Kalathil, et al. Curriculum reinforcement learning from easy to hard tasks improves llm reasoning. *arXiv preprint arXiv:2506.06632*, 2025.

[216] Qizhi Pei, Wei Zhang, Jinhua Zhu, Kehan Wu, Kaiyuan Gao, Lijun Wu, Yingce Xia, and Rui Yan. Biot5: Enriching cross-modal integration in biology with chemical knowledge and natural language associations. *arXiv preprint arXiv:2310.07276*, 2023.

[217] Baolin Peng, Chunyuan Li, Pengcheng He, Michel Galley, and Jianfeng Gao. Instruction tuning with gpt-4. *arXiv preprint arXiv:2304.03277*, 2023.

[218] Xingang Peng, Jiaqi Guan, Qiang Liu, and Jianzhu Ma. Moldiff: Addressing the atom-bond inconsistency problem in 3d molecule diffusion generation. *arXiv preprint arXiv:2305.07508*, 2023.

[219] Jonas Pfeiffer, Aishwarya Kamath, Andreas Rücklé, Kyunghyun Cho, and Iryna Gurevych. Adapterfusion: Non-destructive task composition for transfer learning. *arXiv preprint arXiv:2005.00247*, 2020.

[220] Angus Phillips, Hai-Dang Dau, Michael John Hutchinson, Valentin De Bortoli, George Deligiannidis, and Arnaud Doucet. Particle denoising diffusion sampler. *arXiv preprint arXiv:2402.06320*, 2024.

[221] Mihir Prabhudesai, Anirudh Goyal, Deepak Pathak, and Katerina Fragkiadaki. Aligning text-to-image diffusion models with reward backpropagation. *arXiv preprint arXiv:2310.03739*, 2023.

[222] Zhuoran Qiao, Feizhi Ding, Thomas Dresselhaus, Mia A Rosenfeld, Xiaotian Han, Owen Howell, Aniketh Iyengar, Stephen Opalenski, Anders S Christensen, Sai Krishna Sirumalla, et al. Neuralplexer3: Accurate biomolecular complex structure prediction with flow models. *arXiv preprint arXiv:2412.10743*, 2024.

[223] Rui Qing, Shilei Hao, Eva Smorodina, David Jin, Arthur Zalevsky, and Shuguang Zhang. Protein design: From the aspect of water solubility and stability. *Chemical Reviews*, 122(18):14085–14179, 2022.

[224] Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D Manning, Stefano Ermon, and Chelsea Finn. Direct preference optimization: Your language model is secretly a reward model. *NeurIPS*, 36:53728–53741, 2023.

[225] Roshan Rao, Nicholas Bhattacharya, Niel Thomas, Yan Duan, Xi Chen, John Canny, Pieter Abbeel, and Yun S Song. Evaluating protein transfer learning with tape. *Advanced Neural Iformation Processing Systems*, 2020.

[226] Roshan M Rao, Jason Liu, Robert Verkuil, Joshua Meier, John Canny, Pieter Abbeel, Tom Sercu, and Alexander Rives. Msa transformer. In *International conference on machine learning*, pages 8844–8856. PMLR, 2021.

[227] Roshan M Rao, Jason Liu, Robert Verkuil, Joshua Meier, John Canny, Pieter Abbeel, Tom Sercu, and Alexander Rives. Msa transformer. *Proceedings of Machine Learning Research*, 2021.

[228] Jacob T Rapp, Bennett J Bremer, and Philip A Romero. Self-driving laboratories to autonomously navigate the protein fitness landscape. *Nature chemical engineering*, 1(1):97–107, 2024.

[229] Guillaume Richard, Bernardo P de Almeida, Hugo Dalla-Torre, Christopher Blum, Lorenz Hexemer, Priyanka Pandey, Stefan Laurent, Marie Lopez, Alexandre Laterre, Maren Lang, et al. Chatnt: A multimodal conversational agent for dna, rna and protein tasks. *bioRxiv*, pages 2024–04, 2024.

[230] Alexander Rives, Joshua Meier, Tom Sercu, Siddharth Goyal, Zeming Lin, Jason Liu, Demi Guo, Myle Ott, C Lawrence Zitnick, Jerry Ma, et al. Biological structure and function emerge from scaling unsupervised learning to 250 million protein sequences. *Proceedings of the National Academy of Sciences*, 118(15):e2016239118, 2021.

[231] Yusuf Roohani, Andrew Lee, Qian Huang, Jian Vora, Zachary Steinhart, Kexin Huang, Alexander Marson, Percy Liang, and Jure Leskovec. Biodiscoveryagent: An ai agent for designing genetic perturbation experiments. *arXiv preprint arXiv:2405.17631*, 2024.

[232] Subham Sekhar Sahoo, Marianne Arriola, Yair Schiff, Aaron Gokaslan, Edgar Marroquin, Justin T Chiu, Alexander Rush, and Volodymyr Kuleshov. Simple and effective masked diffusion language models. *arXiv preprint arXiv:2406.07524*, 2024.

[233] Marco Salvatore, Marc Horlacher, Ole Winther, and Robin Andersson. Transfer learning reveals sequence determinants of regulatory element accessibility. *bioRxiv*, pages 2022–08, 2022.

[234] Anirban Sarkar, Ziqi Tang, Chris Zhao, and Peter Koo. Designing dna with tunable regulatory activity using discrete diffusion. *bioRxiv*, pages 2024–05, 2024.

[235] Vıctor Garcia Satorras, Emiel Hoogeboom, and Max Welling. E (n) equivariant graph neural networks. In *International conference on machine learning*, pages 9323–9332. PMLR, 2021.

[236] Yair Schiff, Chia-Hsiang Kao, Aaron Gokaslan, Tri Dao, Albert Gu, and Volodymyr Kuleshov. Caduceus: Bi-directional equivariant long-range dna sequence modeling. *Proceedings of machine learning research*, 235:43632, 2024.

[237] Robert Schmirler, Michael Heinzinger, and Burkhard Rost. Fine-tuning protein langauge models boosts predictions across diverse tasks. *Nature Communications*, 2024.

[238] Arne Schneuing, Charles Harris, Yuanqi Du, Kieran Didi, Arian Jamasb, Ilia Igashov, Weitao Du, Carla Gomes, Tom L Blundell, Pietro Lio, et al. Structure-based drug design with equivariant diffusion models. *Nature Computational Science*, 4(12):899–909, 2024.

[239] John Schulman. Trust region policy optimization. *arXiv preprint arXiv:1502.05477*, 2015.

[240] John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*, 2017.

[241] Ozan Sener and Vladlen Koltun. Multi-task learning as multi-objective optimization. *Advances in neural information processing systems*, 31, 2018.

[242] Ozan Sener and Silvio Savarese. Active learning for convolutional neural networks: A core-set approach. *arXiv preprint arXiv:1708.00489*, 2017.

[243] Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, YK Li, Y Wu, et al. Deepseekmath: Pushing the limits of mathematical reasoning in open language models. *arXiv preprint arXiv:2402.03300*, 2024.

[244] Zhen Shen, Wenzheng Bao, and De-Shuang Huang. Recurrent neural network for predicting transcription factor binding sites. *Scientific reports*, 8(1):15270, 2018.

[245] Chence Shi, Chuanrui Wang, Jiarui Lu, Bozitao Zhong, and Jian Tang. Protein sequence and structure co-design with equivariant translation. *arXiv preprint arXiv:2210.08761*, 2022.

[246] Haizhou Shi, Zihao Xu, Hengyi Wang, Weiyi Qin, Wenyuan Wang, Yibin Wang, Zifeng Wang, Sayna Ebrahimi, and Hao Wang. Continual learning of large language models: A comprehensive survey. *ACM Computing Surveys*, 58(5):1–42, 2025.

[247] Nilmani Singh, Stephan Lane, Tianhao Yu, Jingxia Lu, Adrianna Ramos, Haiyang Cui, and Huimin Zhao. A generalized platform for artificial intelligence-powered autonomous enzyme engineering. *Nature communications*, 16(1):5648, 2025.

[248] Michael D Skarlinski, Sam Cox, Jon M Laurent, James D Braza, Michaela Hinks, Michael J Hammerling, Manvitha Ponnapati, Samuel G Rodriques, and Andrew D White. Language agents achieve superhuman synthesis of scientific knowledge. *arXiv preprint arXiv:2409.13740*, 2024.

[249] Charlie Snell, Jaehoon Lee, Kelvin Xu, and Aviral Kumar. Scaling llm test-time compute optimally can be more effective than scaling model parameters. *arXiv preprint arXiv:2408.03314*, 2024.

[250] Jascha Sohl-Dickstein, Eric Weiss, Niru Maheswaranathan, and Surya Ganguli. Deep unsupervised learning using nonequilibrium thermodynamics. In *International conference on machine learning*, pages 2256–2265. PMLR, 2015.

[251] Dongyuan Song, Qingyang Wang, Guanao Yan, Tianyang Liu, Tianyi Sun, and Jingyi Jessica Li. scdesign3 generates realistic in silico data for multimodal single-cell and spatial omics. *Nature Biotechnology*, 42(2): 247–252, 2024.

[252] Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. *arXiv preprint arXiv:2010.02502*, 2020.

[253] Yang Song, Jascha Sohl-Dickstein, Diederik P Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. Score-based generative modeling through stochastic differential equations. *ICLR*, 2021.

[254] Yuxuan Song, Jingjing Gong, Minkai Xu, Ziyao Cao, Yanyan Lan, Stefano Ermon, Hao Zhou, and Wei-Ying Ma. Equivariant flow matching with hybrid probability transport for 3d molecule generation. In *Thirty-seventh Conference on Neural Information Processing Systems*.

[255] Hannes Stark, Bowen Jing, Chenyu Wang, Gabriele Corso, Bonnie Berger, Regina Barzilay, and Tommi Jaakkola. Dirichlet flow matching with applications to dna sequence design. *arXiv preprint arXiv:2402.05841*, 2024.

[256] Hannes Stark, Bowen Jing, Tomas Geffner, Jason Yim, Tommi Jaakkola, Arash Vahdat, and Karsten Kreis. Protcomposer: Compositional protein structure generation with 3d ellipsoids. *arXiv preprint arXiv:2503.05025*, 2025.

[257] Matt Sternke and Joel Karpiak. Proteinrl: Reinforcement learning with generative protein language models for property-directed sequence design. In *NeurIPS 2023 Generative AI and Biology (GenBio) Workshop*, 2023.

[258] Alexey Strokach, David Becerra, Carles Corbi-Verge, Albert Perez-Riba, and Philip M Kim. Fast and flexible protein design using deep graph neural networks. *Cell systems*, 11(4):402–411, 2020.

[259] Jin Su, Chenchen Han, Yuyang Zhou, Junjie Shan, Xibin Zhou, and Fajie Yuan. Saprot: Protein language modeling with structure-aware vocabulary. *BioRxiv*, pages 2023–10, 2023.

[260] Xingyu Su, Xiner Li, Yuchao Lin, Ziqian Xie, Degui Zhi, and Shuiwang Ji. Language models for controllable dna sequence design. *arXiv preprint arXiv:2507.19523*, 2025.

[261] Xingyu Su, Xiner Li, Masatoshi Uehara, Sunwoo Kim, Yulai Zhao, Gabriele Scalia, Ehsan Hajiramezanali, Tommaso Biancalani, Degui Zhi, and Shuiwang Ji. Iterative distillation for reward-guided fine-tuning of diffusion models in biomolecular design. *arXiv preprint arXiv:2507.00445*, 2025.

[262] Kiera H Sumida, Reyes Núñez-Franco, Indrek Kalvet, Samuel J Pellock, Basile IM Wicky, Lukas F Milles, Justas Dauparas, Jue Wang, Yakov Kipnis, Noel Jameson, et al. Improving protein expression, stability, and function with proteinmpnn. *Journal of the American Chemical Society*, 146(3):2054–2061, 2024.

[263] Michael Sun, Weize Yuan, Gang Liu, Wojciech Matusik, and Marinka Zitnik. Protein structure tokenization via geometric byte pair encoding. *arXiv preprint arXiv:2511.11758*, 2025.

[264] Richard S Sutton, David McAllester, Satinder Singh, and Yishay Mansour. Policy gradient methods for reinforcement learning with function approximation. *Advances in neural information processing systems*, 12, 1999.

[265] Burak Suyunu, Enes Taylan, and Arzucan Özgür. Linguistic laws meet protein sequences: A comparative analysis of subword tokenization methods. In *2024 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, pages 4489–4496. IEEE, 2024.

[266] Nathan J Szymanski, Bernardus Rendy, Yuxing Fei, Rishi E Kumar, Tanjin He, David Milsted, Matthew J McDermott, Max Gallant, Ekin Dogus Cubuk, Amil Merchant, et al. An autonomous laboratory for the accelerated synthesis of novel materials. *Nature*, 624(7990):86–91, 2023.

[267] Sophia Tang, Yinuo Zhang, and Pranam Chatterjee. Peptune: De novo generation of therapeutic peptides with multi-objective-guided discrete diffusion, 2025. URL https://arxiv.org/abs/2412.17780.

[268] Sophia Tang, Yuchen Zhu, Molei Tao, and Pranam Chatterjee. Tr2-d2: Tree search guided trajectory-aware fine-tuning for discrete diffusion. *arXiv preprint arXiv:2509.25171*, 2025.

[269] Xiangru Tang, Zhuoyun Yu, Jiapeng Chen, Yan Cui, Daniel Shao, Weixu Wang, Fang Wu, Yuchen Zhuang, Wenqi Shi, Zhi Huang, et al. Cellforge: agentic design of virtual cell models. *arXiv preprint arXiv:2508.02276*, 2025.

[270] Xin Tang, Jiawei Zhang, Yichun He, Xinhe Zhang, Zuwan Lin, Sebastian Partarrieu, Emma Bou Hanna, Zhaolin Ren, Hao Shen, Yuhong Yang, et al. Explainable multi-task learning for multi-modality biological data analysis. *Nature communications*, 14(1):2546, 2023.

[271] Chai Discovery Team, Jacques Boitreaud, Jack Dent, Danny Geisz, Matthew McPartlon, Joshua Meier, Zhuoran Qiao, Alex Rogozhnikov, Nathan Rollins, Paul Wollenhaupt, et al. Zero-shot antibody design in a 24-well plate. *bioRxiv*, 2025.

[272] Christina V Theodoris, Ling Xiao, Anant Chopra, Mark D Chaffin, Zeina R Al Sayed, Matthew C Hill, Helene Mantineo, Elizabeth M Brydon, Zexian Zeng, X Shirley Liu, et al. Transfer learning enables predictions in network biology. *Nature*, 618(7965):616–624, 2023.

[273] Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*, 2023.

[274] Brian L Trippe, Jason Yim, Doug Tischer, David Baker, Tamara Broderick, Regina Barzilay, and Tommi Jaakkola. Diffusion probabilistic modeling of protein backbones in 3d for the motif-scaffolding problem. *arXiv preprint arXiv:2206.04119*, 2022.

[275] Timothy Truong Jr and Tristan Bepler. Poet: A generative model of protein families as sequences-of-sequences. *Advances in Neural Information Processing Systems*, 36:77379–77415, 2023.

[276] Masatoshi Uehara, Yulai Zhao, Kevin Black, Ehsan Hajiramezanali, Gabriele Scalia, Nathaniel Lee Diamant, Alex M Tseng, Tommaso Biancalani, and Sergey Levine. Fine-tuning of continuous-time diffusion models as entropy-regularized control. *arXiv preprint arXiv:2402.15194*, 2024.

[277] Masatoshi Uehara, Yulai Zhao, Ehsan Hajiramezanali, Gabriele Scalia, Gökcen Eraslan, Avantika Lal, Sergey Levine, and Tommaso Biancalani. Bridging model-based optimization and generative modeling via conservative fine-tuning of diffusion models. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024. URL https://openreview.net/forum?id=zIr2QjU4hl.

[278] Masatoshi Uehara, Xingyu Su, Yulai Zhao, Xiner Li, Aviv Regev, Shuiwang Ji, Sergey Levine, and Tommaso Biancalani. Reward-guided refinement in diffusion models with applications to protein and DNA design. In *Proceedings of the International Conference on Machine Learning*, pages 60515–60529, 2025.

[279] Masatoshi Uehara, Yulai Zhao, Chenyu Wang, Xiner Li, Aviv Regev, Sergey Levine, and Tommaso Biancalani. Inference-time alignment in diffusion models with reward-guided generation: Tutorial and review. *arXiv preprint arXiv:2501.09685*, 2025.

[280] Masatoshi Uehara, Yulai Zhao, Chenyu Wang, Xiner Li, Aviv Regev, Sergey Levine, and Tommaso Biancalani. Reward-guided controlled generation for inference-time alignment in diffusion models: Tutorial and review. *arXiv preprint arXiv:2501.09685*, 2025.

[281] Aaron Van Den Oord, Oriol Vinyals, et al. Neural discrete representation learning. *Advances in neural information processing systems*, 30, 2017.

[282] Michel Van Kempen, Stephanie S Kim, Charlotte Tumescheit, Milot Mirdita, Jeongjae Lee, Cameron LM Gilchrist, Johannes Söding, and Martin Steinegger. Fast and accurate protein structure search with foldseek. *Nature biotechnology*, 42(2):243–246, 2024.

[283] Mihaly Varadi, Stephen Anyango, Mandar Deshpande, Sreenath Nair, Cindy Natassia, Galabina Yordanova, David Yuan, Oana Stroe, Gemma Wood, Agata Laydon, et al. Alphafold protein structure database: massively expanding the structural coverage of protein-sequence space with high-accuracy models. *Nucleic acids research*, 50(D1):D439–D444, 2022.

[284] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

[285] Siddarth Venkatraman, Moksh Jain, Luca Scimeca, Minsu Kim, Marcin Sendera, Mohsin Hasan, Luke Rowe, Sarthak Mittal, Pablo Lemos, Emmanuel Bengio, et al. Amortizing intractable inference in diffusion models for vision, language, and control. *arXiv preprint arXiv:2405.20971*, 2024.

[286] Robert Verkuil, Ori Kabeli, Yilun Du, Basile IM Wicky, Lukas F Milles, Justas Dauparas, David Baker, Sergey Ovchinnikov, Tom Sercu, and Alexander Rives. Language models generalize beyond natural proteins. *BioRxiv*, pages 2022–12, 2022.

[287] Clement Vignac, Igor Krawczuk, Antoine Siraudin, Bohan Wang, Volkan Cevher, and Pascal Frossard. Digress: Discrete denoising diffusion for graph generation. In *The Eleventh International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=UaAD-Nu86WX.

[288] Yunqi Wan and Zhenran Jiang. Transcrispr: transformer based hybrid model for predicting crispr/cas9 single guide rna cleavage efficiency. *IEEE/ACM transactions on computational biology and bioinformatics*, 20(2): 1518–1528, 2022.

[289] Chao Wang, Hehe Fan, Ruijie Quan, Lina Yao, and Yi Yang. Protchatgpt: Towards understanding proteins with hybrid representation and large language models. In *Proceedings of the 48th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 1076–1086, 2025.

[290] Chenyu Wang, Masatoshi Uehara, Yichun He, Amy Wang, Tommaso Biancalani, Avantika Lal, Tommi Jaakkola, Sergey Levine, Hanchen Wang, and Aviv Regev. Fine-tuning discrete diffusion models via reward optimization with applications to dna and protein design. *arXiv preprint arXiv:2410.13643*, 2024.

[291] Chenyu Wang, Masatoshi Uehara, Yichun He, Amy Wang, Tommaso Biancalani, Avantika Lal, Tommi Jaakkola, Sergey Levine, Hanchen Wang, and Aviv Regev. Fine-tuning discrete diffusion models via reward optimization with applications to dna and protein design. *arXiv preprint arXiv:2410.13643*, 2024.

[292] Guanghan Wang, Yair Schiff, Subham Sekhar Sahoo, and Volodymyr Kuleshov. Remasking discrete diffusion models with inference-time scaling. *arXiv preprint arXiv:2503.00307*, 2025.

[293] Jiuming Wang, Yimin Fan, Liang Hong, Zhihang Hu, and Yu Li. Deep learning for rna structure prediction. *Current Opinion in Structural Biology*, 91:102991, 2025.

[294] Ning Wang, Jiang Bian, Yuchen Li, Xuhong Li, Shahid Mumtaz, Linghe Kong, and Haoyi Xiong. Multi-purpose rna language modelling with motif-aware pretraining and type-guided fine-tuning. *Nature Machine Intelligence*, 6:548–557, 2024.

[295] Xinyou Wang, Zaixiang Zheng, Fei Ye, Dongyu Xue, Shujian Huang, and Quanquan Gu. Diffusion language models are versatile protein learners. *arXiv preprint arXiv:2402.18567*, 2024.

[296] Xinyou Wang, Zaixiang Zheng, Fei Ye, Dongyu Xue, Shujian Huang, and Quanquan Gu. Dplm-2: A multimodal diffusion protein language model. *arXiv preprint arXiv:2410.13782*, 2024.

[297] Yaqing Wang, Sahaj Agarwal, Subhabrata Mukherjee, Xiaodong Liu, Jing Gao, Ahmed Hassan Awadallah, and Jianfeng Gao. Adamix: Mixture-of-adaptations for parameter-efficient model tuning. *arXiv preprint arXiv:2205.12410*, 2022.

[298] Yuyang Wang, Ahmed AA Elhag, Navdeep Jaitly, Joshua M Susskind, and Miguel Ángel Bautista. Swallowing the bitter pill: Simplified scalable conformer generation. In *ICML 2024 AI for Science Workshop*.

[299] Zeyuan Wang, Qiang Zhang, Keyan Ding, Ming Qin, Xiang Zhuang, Xiaotong Li, and Huajun Chen. Instruct-protein: Aligning human and protein language via knowledge instruction. *arXiv preprint arXiv:2310.03269*, 2023.

[300] Zhenyu Wang, Zikang Wang, Jiyue Jiang, Pengan Chen, Xiangyu Shi, and Yu Li. Large language models in bioinformatics: A survey. *arXiv preprint arXiv:2503.04490*, 2025.

[301] Zifeng Wang, Zichen Wang, Balasubramaniam Srinivasan, Vassilis N Ioannidis, Huzefa Rangwala, and Rishita Anubhai. Biobridge: Bridging biomedical foundation models via knowledge graphs. *arXiv preprint arXiv:2310.03320*, 2023.

[302] Ziqi Wang, Hanlin Zhang, Xiner Li, Kuan-Hao Huang, Chi Han, Shuiwang Ji, Sham M Kakade, Hao Peng, and Heng Ji. Eliminating position bias of language models: A mechanistic approach. *arXiv preprint arXiv:2407.01100*, 2024.

[303] Joseph L Watson, David Juergens, Nathaniel R Bennett, Brian L Trippe, Jason Yim, Helen E Eisenach, Woody Ahern, Andrew J Borst, Robert J Ragotte, Lukas F Milles, et al. De novo design of protein structure and function with rfdiffusion. *Nature*, 620(7976):1089–1100, 2023.

[304] Talal Widatalla, Rafael Rafailov, and Brian Hie. Aligning protein generative models with experimental fitness via direct preference optimization. *bioRxiv*, pages 2024–05, 2024.

[305] Kevin E Wu, Howard Chang, and James Zou. Proteinclip: enhancing protein language models with natural language. *bioRxiv*, pages 2024–05, 2024.

[306] Kevin E Wu, Kevin K Yang, Rianne van den Berg, Sarah Alamdari, James Y Zou, Alex X Lu, and Ava P Amini. Protein structure generation via folding diffusion. *Nature communications*, 15(1):1059, 2024.

[307] Luhuan Wu, Brian Trippe, Christian Naesseth, David Blei, and John P Cunningham. Practical and asymptotically exact conditional sampling in diffusion models. *Advances in Neural Information Processing Systems*, 36, 2024.

[308] Yangzhen Wu, Zhiqing Sun, Shanda Li, Sean Welleck, and Yiming Yang. Inference scaling laws: An empirical analysis of compute-optimal inference for llm problem-solving. In *The Thirteenth International Conference on Learning Representations*, 2025.

[309] Yihang Xiao, Jinyi Liu, Yan Zheng, Xiaohan Xie, Jianye Hao, Mingzhi Li, Ruitao Wang, Fei Ni, Yuxiao Li, Jintian Luo, et al. Cellagent: An llm-driven multi-agent framework for automated single-cell data analysis. *arXiv preprint arXiv:2407.09811*, 2024.

[310] Yijia Xiao, Edward Sun, Yiqiao Jin, Qifan Wang, and Wei Wang. Proteingpt: Multimodal llm for protein property prediction and structure understanding. *arXiv preprint arXiv:2408.11363*, 2024.

[311] Yijia Xiao, Wanjia Zhao, Junkai Zhang, Yiqiao Jin, Han Zhang, Zhicheng Ren, Renliang Sun, Haixin Wang, Guancheng Wan, Pan Lu, et al. Protein large language models: A comprehensive survey. *arXiv preprint arXiv:2502.17504*, 2025.

[312] Junhao Xiong, Hunter Nisonoff, Maria Lukarska, Ishan Gaur, Luke M Oltrogge, David F Savage, and Jennifer Listgarten. Guide your favorite protein sequence generative model. *arXiv preprint arXiv:2505.04823*, 2025.

[313] Chencheng Xu, Suying Bao, Ye Wang, Wenxing Li, Hao Chen, Yufeng Shen, Tao Jiang, and Chaolin Zhang. Reference-informed predictionof alternative splicing and splicing-altering mutations from sequences. *Genome Research*, 2024.

[314] Minghao Xu, Zuobai Zhang, Jiarui Lu, Zhaocheng Zhu, Yangtian Zhang, Chang Ma, Runcheng Liu, and Jian Tang. Peer: A comprehensive and multi-task benchmark for protein sequence understanding. *36th Conference on Neural Information Processing Systems, Track on Datasets and Benchmarks*, 2022.

[315] Minghao Xu, Xinyu Yuan, Santiago Miret, and Jian Tang. Protst: Multi-modality learning of protein sequences and biomedical texts. In *International Conference on Machine Learning*, pages 38749–38767. PMLR, 2023.

[316] Minkai Xu, Lantao Yu, Yang Song, Chence Shi, Stefano Ermon, and Jian Tang. Geodiff: A geometric diffusion model for molecular conformation generation. *arXiv preprint arXiv:2203.02923*, 2022.

[317] Keqiang Yan, Xiner Li, Hongyi Ling, Kenna Ashen, Carl Edwards, Raymundo Arróyave, Marinka Zitnik, Heng Ji, Xiaofeng Qian, Xiaoning Qian, et al. Invariant tokenization of crystalline materials for language model enabled generation. *Advances in Neural Information Processing Systems*, 37:125050–125072, 2024.

[318] Fan Yang, Wenchuan Wang, Fang Wang, Yuan Fang, Duyu Tang, Junzhou Huang, Hui Lu, and Jianhua Yao. scbert as a large-scale pretrained deep language model for cell type annotation of single-cell rna-seq data. *Nature Machine Intelligence*, 4(10):852–866, 2022.

[319] Heng Yang, Renzhi Chen, and Ke Li. Bridging sequence-structure alignment in rna foundation models. In *The Thirty-Ninth AAAI Conference on Artificial Intelligence*, 2025.

[320] Jason Yang, Aadyot Bhatnagar, Jeffrey A Ruffolo, and Ali Madani. Conditional enzyme generation using protein language models with adapters. *arXiv preprint arXiv:2410.03634*, 2024.

[321] Kevin K Yang, Zachary Wu, and Frances H Arnold. Machine-learning-guided directed evolution for protein engineering. *Nature methods*, 16(8):687–694, 2019.

[322] Xiao-Wen Yang, Jie-Jing Shao, Lan-Zhe Guo, Bo-Wen Zhang, Zhi Zhou, Lin-Han Jia, Wang-Zhou Dai, and Yu-Feng Li. Neuro-symbolic artificial intelligence: Towards improving the reasoning abilities of large language models. *arXiv preprint arXiv:2508.13678*, 2025.

[323] Yuning Yang, Gen Li, Kuan Pang, Wuxinhao Cao, Zhaolei Zhang, and Xiangtao Li. Deciphering 3'utr mediated gene regulation using interpretable deep representation learning. *Advanced Science*, 11(39):2407013, 2024.

[324] Zhao Yang, Bing Su, Chuan Cao, and Ji-Rong Wen. Regulatory dna sequence design with reinforcement learning. *arXiv preprint arXiv:2503.07981*, 2025.

[325] Zhou Yao, Wenjing Zhang, Peng Song, Yuxue Hu, and Jianxiao Liu. Deepformer: a hybrid network based on convolutional neural network and flow-attention mechanism for identifying the function of dna sequences. *Briefings in Bioinformatics*, 24(2):bbad095, 2023.

[326] Jiasheng Ye, Zaixiang Zheng, Yu Bao, Lihua Qian, and Quanquan Gu. Diffusion language models can perform many tasks with scaling and instruction-finetuning. *arXiv preprint arXiv:2308.12219*, 2023.

[327] Kai Yi, Bingxin Zhou, Yiqing Shen, Pietro Liò, and Yuguang Wang. Graph denoising diffusion for inverse protein folding. *Advances in Neural Information Processing Systems*, 36:10238–10257, 2023.

[328] Jason Yim, Brian L Trippe, Valentin De Bortoli, Emile Mathieu, Arnaud Doucet, Regina Barzilay, and Tommi Jaakkola. Se (3) diffusion model with application to protein backbone generation. *arXiv preprint arXiv:2302.02277*, 2023.

[329] Jason Yim, Andrew Campbell, Emile Mathieu, Andrew YK Foong, Michael Gastegger, José Jiménez-Luna, Sarah Lewis, Victor Garcia Satorras, Bastiaan S Veeling, Frank Noé, et al. Improved motif-scaffolding with se (3) flow matching. *ArXiv*, pages arXiv–2401, 2024.

[330] Zhenhua Yu, Furui Liu, and Yang Li. sctca: a hybrid transformer-cnn architecture for imputation and denoising of scdna-seq data. *Briefings in Bioinformatics*, 25(6):bbae577, 2024.

[331] Tony Zeng and Yang I Li. Predicting rna splicing from dna sequence using pangolin. *Genome Biology*, 2022.

[332] B Zhang, K Liu, Z Zheng, Y Liu, J Mu, T Wei, and H Chen. Protein language model supervised precise and efficient protein backbone design method. 2023.

[333] Biao Zhang, Zhongtao Liu, Colin Cherry, and Orhan Firat. When scaling meets llm finetuning: The effect of data, model and finetuning method. *arXiv preprint arXiv:2402.17193*, 2024.

[334] Bo Zhang, Kexin Liu, Zhuoqi Zheng, Junjie Zhu, Zhengxin Li, YUnfeiyang Liu, Junxi Mu, Ting Wei, and Hai-Feng Chen. Protein language model supervised scalable approach for diverse and designable protein motif-scaffolding with gpdl. *bioRxiv*, pages 2023–10, 2023.

[335] Cheng Zhang, Adam Leach, Thomas Makkink, Miguel Arbesú, Ibtissem Kadri, Daniel Luo, Liron Mizrahi, Sabrine Krichen, Maren Lang, Andrey Tovchigrechko, et al. Framedipt: Se (3) diffusion model for protein structure inpainting. *biorxiv*, pages 2023–11, 2023.

[336] Daoan Zhang, Weitong Zhang, Yu Zhao, Jianguo Zhang, Bing He, Chenchen Qin, and Jianhua Yao. Dnagpt: a generalized pre-trained tool for versatile dna sequence analysis tasks. *arXiv preprint arXiv:2307.05628*, 2023.

[337] Jun Zhang, Qingcai Chen, and Bin Liu. Deepdrbp-2l: a new genome annotation predictor for identifying dna-binding proteins and rna-binding proteins using convolutional neural network and long short-term memory. *IEEE/ACM transactions on computational biology and bioinformatics*, 18(4):1451–1463, 2019.

[338] Ningyu Zhang, Zhen Bi, Xiaozhuan Liang, Siyuan Cheng, Haosen Hong, Shumin Deng, Jiazhang Lian, Qiang Zhang, and Huajun Chen. Ontoprotein: Protein pretraining with gene ontology embedding. *arXiv preprint arXiv:2201.11147*, 2022.

[339] Qiang Zhang, Keyang Ding, Tianwen Lyv, Xinda Wang, Qingyu Yin, Yiwen Zhang, Jing Yu, Yuhao Wang, Xiaotong Li, Zhuoyi Xiang, et al. Scientific large language models: A survey on biological & chemical domains. *arXiv preprint arXiv:2401.14656*, 2024.

[340] Shengyu Zhang, Linfeng Dong, Xiaoya Li, Sen Zhang, Xiaofei Sun, Shuhe Wang, Jiwei Li, Runyi Hu, Tianwei Zhang, Fei Wu, and Guoyin Wang. Instruction tuning for large language models: A survey, 2025. URL https://arxiv.org/abs/2308.10792.

[341] Xuan Zhang, Limei Wang, Jacob Helwig, Youzhi Luo, Cong Fu, Yaochen Xie, ..., and Shuiwang Ji. Artificial intelligence for science in quantum, atomistic, and continuum systems. *Foundations and Trends® in Machine Learning*, 18(4):385–912, 2025. ISSN 1935-8237. doi: 10.1561/2200000115. URL http://dx.doi.org/10.1561/2200000115.

[342] Yangtian Zhang, Zuobai Zhang, Bozitao Zhong, Sanchit Misra, and Jian Tang. Diffpack: A torsional diffusion model for autoregressive protein side-chain packing. *Advances in Neural Information Processing Systems*, 36: 48150–48172, 2023.

[343] Yu Zhang and Qiang Yang. A survey on multi-task learning. *IEEE transactions on knowledge and data engineering*, 34(12):5586–5609, 2021.

[344] Zaixi Zhang, Wan Xiang Shen, Qi Liu, and Marinka Zitnik. Efficient generation of protein pockets with pocketgen. *Nature Machine Intelligence*, 2024.

[345] Siyan Zhao, Devaansh Gupta, Qinqing Zheng, and Aditya Grover. d1: Scaling reasoning in diffusion large language models via reinforcement learning. *arXiv preprint arXiv:2504.12216*, 2025.

[346] Stephen Zhao, Rob Brekelmans, Alireza Makhzani, and Roger Grosse. Probabilistic inference in language models via twisted sequential monte carlo. *arXiv preprint arXiv:2404.17546*, 2024.

[347] Junhao Zheng, Shengjie Qiu, Chengming Shi, and Qianli Ma. Towards lifelong learning of large language models: A survey. *ACM Computing Surveys*, 57(8):1–35, 2025.

[348] Bingxin Zhou, Lirong Zheng, Banghao Wu, Kai Yi, Bozitao Zhong, Yang Tan, Qian Liu, Pietro Liò, and Liang Hong. A conditional protein diffusion model generates artificial programmable endonuclease sequences with enhanced activity. *Cell Discovery*, 10(1):95, 2024.

[349] Hanjing Zhou, Mingze Yin, Wei Wu, Mingyang Li, Kun Fu, Jintai Chen, Jian Wu, and Zheng Wang. Protclip: Function-informed protein multi-modal learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pages 22937–22945, 2025.

[350] Lianhao Zhou, Hongyi Ling, Cong Fu, Yepeng Huang, Michael Sun, Wendi Yu, Xiaoxuan Wang, Xiner Li, Xingyu Su, Junkai Zhang, Xiusi Chen, Chenxing Liang, Xiaofeng Qian, Heng Ji, Wei Wang, Marinka Zitnik, and Shuiwang Ji. Autonomous agents for scientific discovery: Orchestrating scientists, language, code, and physics. *arXiv preprint arXiv:2510.09901*, 2025.

[351] Zhihan Zhou, Yanrong Ji, Weijian Li, Pratik Dutta, Ramana Davuluri, and Han Liu. Dnabert-2: Efficient foundation model and benchmark for multi-species genome. *arXiv preprint arXiv:2306.15006*, 2023.

[352] Ziyi Zhou, Liang Zhang, Yuanxi Yu, Banghao Wu, Mingchen Li, Liang Hong, and Pan Tan. Enhancing efficiency of protein language models with minimal wet-lab data through few-shot learning. *Nature Communications*, 15 (1):5566, 2024.

[353] Xiao Zhu, Chenchen Qin, Fang Wang, Fan Yang, Bing He, Yu Zhao, and Jianhua Yao. Cd-gpt: a biological foundation model bridging the gap between molecular sequences through central dogma. *bioRxiv*, pages 2024–06, 2024.

[354] Yiheng Zhu, Zitai Kong, Jialu Wu, Weize Liu, Yuqiang Han, Mingze Yin, Hongxia Xu, Chang-Yu Hsieh, and Tingjun Hou. Generative ai for controllable protein sequence design: A survey. *arXiv preprint arXiv:2402.10516*, 2024.

[355] Fuzhen Zhuang, Zhiyuan Qi, Keyu Duan, Dongbo Xi, Yongchun Zhu, Hengshu Zhu, Hui Xiong, and Qing He. A comprehensive survey on transfer learning. *Proceedings of the IEEE*, 109(1):43–76, 2020.

[356] Daniel M Ziegler, Nisan Stiennon, Jeffrey Wu, Tom B Brown, Alec Radford, Dario Amodei, Paul Christiano, and Geoffrey Irving. Fine-tuning language models from human preferences. *arXiv preprint arXiv:1909.08593*, 2019.