Foundation Models for AI-enabled Biological Design

Asher Moldwin[1], **Amarda Shehu**[1]

[1]Department of Computer Science, George Mason University, Fairfax, VA, USA
amoldwin@gmu.edu, amarda@gmu.edu

Abstract

This paper surveys foundation models for AI-enabled biological design, focusing on recent developments in applying large-scale, self-supervised models to tasks such as protein engineering, small molecule design, and genomic sequence design. Though this domain is evolving rapidly, this survey presents and discusses a taxonomy of current models and methods. The focus is on challenges and solutions in adapting these models for biological applications, including biological sequence modeling architectures, controllability in generation, and multi-modal integration. The survey concludes with a discussion of open problems and future directions, offering concrete next-steps to improve the quality of biological sequence generation.

Introduction

Natural language processing (NLP) has undergone a recent revolution driven by Transformer-based neural networks and the attention mechanism. These architectures have enabled large language models (LLMs) to achieve remarkable performance on natural language understanding and generation tasks, many of which were once thought to be the exclusive domain of human intelligence. While this transformation in NLP has captured widespread attention, a quieter but equally significant revolution has been unfolding in the biological sciences.

Advances in molecular biology, particularly in large-scale data collection initiatives such as the Protein Structure Initiative(Pro 2000–2015), the Human Genome Project(Collins and Fink 1995) and structural genomics efforts(Dawson et al. 2017; Jones et al. 2014; Bank 1971), have paved the way for a new era of biological research. The rapid development of next-generation sequencing technologies has led to a wealth of widely available genomics, proteomics, and metabolomics data. This multi-omics view of organisms has driven breakthroughs such as AlphaFold's(Jumper, Evans et al. 2021) success in solving protein structure prediction, earning two of its its main authors, Demis Hassabis and John Jumper, the 2024 Nobel Prize in Chemistry. Perhaps more importantly, these rapid developments are catalyzing a boom in bioinformatics methodologies for various prediction and generation tasks.

The aim of this survey is to provide an overview of recent contributions to AI-enabled biological design, where the goal is to leverage biological data and new methods to design and engineer biological entities with impactful applications in health, drug discovery(Blanco-Gonzalez et al. 2023), synthetic biology(Voigt 2020), and material sciences(Tang et al. 2021). Notably, David Baker, the third recipient of the 2024 Nobel Prize in Chemistry, was recognized for pioneering work in protein design.

While many neural network architectures and machine learning methods are being developed for biological design, this survey focuses on a particularly promising frontier: Foundation Models (FMs). These models, capable of learning general representations from vast datasets in a task-agnostic setting, offer exciting opportunities for biological applications. Our understanding of FMs has evolved in recent years, and we adopt a broad definition that aligns with the Stanford University Human-Centered AI group's understanding(Bommasani et al. 2021) in coining the term: FMs are architecture-agnostic and defined by their ability to learn from task-agnostic pre-training, making them applicable across a range of tasks. Although some researchers narrowly define FMs as sequential models or strictly equate them with LLMs, we take the position that FMs should be considered more broadly for their capacity to learn deep representations that can be leveraged for downstream tasks.

Given the rapid progress in FMs, it is timely to survey this expanding landscape. These models are quickly transforming AI research, with new developments appearing almost weekly, particularly in biological design applications. While this survey cannot be fully comprehensive, it focuses on key areas of concentrated activity or where significant challenges are being articulated and addressed. Specifically, we examine FMs that, leveraging the analogies between natural language and biological sequences, directly support biological sequence-design tasks; though this focus may be somewhat narrow, we believe it encompasses some of the most exciting developments in the field.

The survey highlights sequence-based FMs and focuses on commonly-applied architectures, such as the Transformer, Diffusion, and State Space Model (SSM) architectures. While FMs can encompass a broader range of architectures, we concentrate on these three due to their demonstrated success in recent biological design literature. Moreover, while FMs can support a variety of prediction and downstream tasks, our focus on biological design narrows

our attention to generative tasks, where models not only analyze existing biological data but also aim to create novel biological entities with desired properties. As we will discuss, this type of biological design includes *de-novo* drug design, protein engineering, and DNA design through gene-editing.

Background and Preliminaries

FMs

While FMs do not dictate a specific architecture, some are better suited to representation learning, a key aspect of modern FMs. We formalize implicit representation learning using a decoder-only text-based Transformer before exploring diffusion and SSM paradigms.

Transformer Architecture In Transformer architectures(Vaswani 2017), sequences are handled by maintaining separate representations for each token at each layer, giving: $h_i^{(l)} = \text{TransformerLayer}^{(l)}\left(h_1^{(l-1)}, h_2^{(l-1)}, \ldots, h_T^{(l-1)}\right)$. This allows the model to create contextualized representations through the attention mechanism, where each token's representation at layer l depends on the representations of all tokens from the previous layer. Specifically, for each token embedding $h_i^{(l-1)}$, Transformer attention computes attention scores using the query, key, and value projections: $Q_i = W_q h_i^{(l-1)}$, $K_j = W_k h_j^{(l-1)}$, and $V_j = W_v h_j^{(l-1)}$ for all j, where W_q, W_k, and W_v are trainable weight matrices. The layer representation $h_i^{(l)}$ is then computed as a weighted sum of the values: $h_i^{(l)} = \sum_{j=1}^{T} \text{softmax}(Q_i \cdot K_j)V_j$.

In this formulation, attention incurs a computational cost during training proportional to the number of input tokens squared, rendering its use for long sequences impractical. Memory usage during inference scales linearly with the sequence length, as key and value representations for all tokens need to be stored, further limiting its efficiency.

To effectively leverage the representational power added by the inclusion of context, the use of general, self-supervised pre-training objectives has proven crucial for learning representations that can be repurposed for other downstream tasks. For decoder-only models, the pre-training objective is typically autoregressive language modeling, where the model predicts the next token in a sequence given all previous tokens. Formally, the model learns to approximate the probability distribution $P(x_i|x_1, x_2, \ldots, x_{i-1})$. The attention mechanism is thus constrained to left-to-right context, ensuring that each token attends only to previous tokens: $h_i^{(l)} = \text{Attention}^{(l)}(h_1^{(l-1)}, h_2^{(l-1)}, \ldots, h_{i-1}^{(l-1)})$.

Finally, while the attention mechanism handles dependencies between tokens, applying attention multiple times in parallel at each layer (multihead attention), and stacking many Transformer layers in a deep network builds the capacity to learn complex and hierarchical relationships in data.

State-Space Models (SSMs) State-Space Models (SSMs) provide an efficient alternative to Transformers for sequence modeling, particularly for tasks requiring long-range dependencies. While Transformers rely on quadratic self-attention and increasing memory during inference, SSMs achieve subquadratic scaling and constant memory by employing a Recurrent Neural Network architecture and using fixed-size hidden states and linear state-space equations. These equations model state evolution over time as $h_{t+1} = Ah_t + Bu_{t+1}$, with outputs $y_t = Ch_t + \Delta u_t$. Key architectures like S4(Gu, Goel, and Ré 2021) enhance SSMs with optimizations for long sequences, while Mamba(Gu and Dao 2023) employs Selective State Spaces to learn dynamic versions the A, B and C matrices whose values depend on the input token.

Diffusion Modeling Models can forgo attention by instead leveraging diffusion processes wherein noise is iteratively introduced to the inputs and then a learned reverse diffusion process is performed to reconstruct realistic inputs. This allows global contextual relationships to gradually emerge in the outputs of diffusion models over many iterations of denoising. Formally, diffusion models progressively add random noise to input data over T timesteps, following a Markov process in which each timestep's result depends only on the previous timestep and the added noise. The probability of transforming the input x_0 into the noisy state x_T after T steps is given by: $q(x_{1:T}|x_0) = \prod_{t=1}^{T} q(x_t|x_{t-1})$, where each transition is defined as: $q(x_t|x_{t-1}) = \mathcal{N}(x_t; \sqrt{1-\beta_t}x_{t-1}, \beta_t\mathbf{I})$ with β_t controlling the variance of the added noise at each step. The reverse process, or "denoising," mirrors this forward process, allowing the model to reconstruct the input by removing noise over time. For reverse diffusion, a neural network ϵ_θ is trained to progressively denoise x_T by reversing the diffusion process at each timestep t. The standard training objective is derived from a variational bound on the data liklihood, giving the loss function $\mathcal{L}(\theta) = \mathbb{E}_{t,x_0,\epsilon}\left[\|\epsilon - \epsilon_\theta(x_t, t)\|^2\right]$.

While diffusion models were originally developed for continuous data, like images, alternative approaches have adapted this framework to discrete data, such as text. In one approach, tokens are mapped into a continuous latent space, where noise is progressively added over T timesteps. The reverse process denoises the embeddings step by step, and a decoder maps the final latent embeddings back to discrete tokens. This allows diffusion models to maintain local coherence while capturing global structure over multiple denoising steps. Another approach directly applies noise to discrete tokens(Li et al. 2022) by either replacing tokens with others sampled from the vocabulary(Hoogeboom et al. 2021) or by masking them(Hoogeboom et al.). In the reverse process, a neural network restores the correct token sequence from its noisy version by minimizing the difference between the predicted tokens and the ground truth at each timestep. This iterative denoising process enables diffusion models to generate coherent text, effectively adapting the diffusion framework to discrete data.

Biological Design

Biological design encompasses the creation or modification of biological entities for specific functions or properties. Researchers focus on three primary classes of biolog-

ical or chemical objects: proteins, DNA/RNA, and small molecules. Different representations of these objects typically enable different neural network architectures, with the two main representations being sequence-based and graph-based. Sequence-based representations are derived from the chemical formulas, such as representing proteins as sequences of characters corresponding to the twenty naturally occurring amino acids or representing genomic sequences as strings of the four nucleotide bases. The SMILES representation for small molecules captures atoms as well as bonds and branches. Graph-based representations, on the other hand, model the bonds and interactions (e.g., hydrogen bonds, van der Waals interactions) between atoms or molecular units as edges connecting vertices.

Biological design combines principles from biology, chemistry, and computational science to engineer biological systems. The primary areas of greatest activity are: (1) Protein engineering: Designing novel proteins or modifying existing ones for enhanced function, stability, or specificity; (2) Small molecule design: Creating new drug candidates or optimizing existing compounds for improved efficacy and reduced side effects. Advances toward generating biologically functional small molecules or designing novel proteins with precise control over properties—such as binding affinity, solubility, or toxicity—could significantly accelerate therapeutic development by streamlining the identification of drug candidates and (3) Genomic sequence design: Engineering DNA sequences for applications in synthetic biology, gene therapy, or CRISPR-based genome editing. In this survey we focus on how text-based FMs are adapted for biological design due to their generative capabilities. We will refer to FMs for these three topics as Chemical Language Models (CLMs), Protein Language Models (PLMs), and Genomic Language Models (GLMs).

Taxonomy and Survey

We categorize recent methods in Foundation Models for AI-enabled Biological Design by the following taxonomy:
(A) **Architectures**: (1) Transformers; (2) State Space Models; (3) Diffusion Models.
(B) **Controllability in Generation**: (1) Fine-Tuning; (2) Conditional Generation; (3) Reinforcement Learning; (4) Custom Objective Functions; (5) Multi-condition Generation and Tradeoff Handling.
(C) **Multi-Modal FMs**: (1) Combining Biological Sequences; (2) Sequence and Structure Integration; (3) Incorporating Natural Language and Domain Knowledge.

Methods aligned with these categories are summarized in Table 1.

Architectures

Transformer Models

Classic Transformer Models The SMILES representation facilitates using standard Transformer architectures like GPT and XLNet (Yang 2019) for small-molecule design, often with minimal modifications and NLP-optimized hyperparameters. MolGPT (Bagal et al. 2021), MCMG (Wang et al. 2021), Regression Transformer (Moret et al. 2020), and Taiga (Mazuz et al. 2023) employ GPT-like architectures but differ in training and controllability strategies. PLMs like Progen (Madani et al. 2023), Progen2 (Nijkamp et al. 2023), and ProtGPT2 (Ferruz, Schmidt, and Höcker 2022) also use standard architectures with minor enhancements. For instance, Progen2 incorporates Rotary Position Embeddings (RoPE (Su et al. 2024)) and improved parallelization. In contrast, most DNA FMs adopt highly customized attention mechanisms or alternative architectures.

Specialized for Biological Design Transformer-based models often adapt attention mechanisms to suit biological tasks. Examples include RFDiffusionAllAtom (Krishna et al. 2024), NOS (Gruver et al. 2024), and Evo (Nguyen et al. 2024a). RFDiffusionAllAtom extends RoseTTAFold2's multi-track attention with atomic-level details for protein-molecule complex design. NOS employs an encoder-decoder Transformer for forward and backward diffusion, optimized for antibody design. Evo uses a hybrid StripedHyena architecture, combining RoPE-based attention with convolutional layers (Poli et al. 2023), enabling efficient processing of long genomic sequences (up to 131 kilobases) for tasks like CRISPR DNA design.

State Space Models

SSM architectures often struggle to capture long-range dependencies, which are critical for biological data. Protein sequences involve distant interactions (e.g., hydrogen bonds, disulfide bridges), genomic data feature regulatory elements far from target genes, and small-molecule sequences require modeling ring closures and branching (Özçelik et al. 2024).

SSM-based models like Mamba and S4 have shown superior performance in tasks like de novo molecule generation. For example, Özçelik et al. (2024) benchmarked S4 on SMILES sequences, demonstrating its advantages in generating valid and unique molecules. ProtMAMBA applies SSMs to protein sequence generation, using a "fill-in-the-middle" objective (Bavarian et al. 2022) to model long-range dependencies for sequence inpainting.

Hyena-based models, such as HyenaDNA (Nguyen et al. 2024b), extend SSM capabilities by processing sequences up to 1 million nucleotides. These architectures excel in DNA modeling, capturing interactions at single-nucleotide resolution for tasks like regulatory element detection. RegLM (Lal et al. 2024) and Evo (Nguyen et al. 2024a) leverage Hyena variants, with Evo combining Hyena layers and attention mechanisms in its StripedHyena architecture.

Diffusion Models

Classic Diffusion Models Diffusion models, traditionally used for continuous data like images, have been adapted for biological sequences by embedding discrete token inputs. DNA-Diffusion (Senan et al. 2024) generates regulatory sequences controlling chromatin accessibility using a U-net architecture and transforms nucleotides into a continuous range for Gaussian noise introduction. Similarly, DiscDiff (Li et al. 2024) employs a U-net with ResNet blocks and encodes DNA into a continuous latent space via a Variational Autoencoder, decoding it back into discrete form after reverse diffusion.

Table 1: Summary of current FM models for Biological Design.

Model Name	Domain	Architecture	Design Goal	Generation Method
MolGPT (Bagal et al. 2021)	Small Molecule	Decoder-only Transformer	Control TPSA, QED, SAS, LogP and molecular scaffolds	Conditional Autoregressive Generation, pre-trained with prepended property condition embeddings
MCMG (Wang et al. 2021)	Small Molecule	Decoder-only Transformer Distilled into RNN	Control Bioactivity, QED, SAS	Same as above, with added reinforcement learning rewarding low property error
Taiga (Mazuz et al. 2023)	Small Molecule	Decoder-only Transformer	QED, inhibitory potency (plC50)	Policy gradient reinforcement learning rewarding high property values
S4 CLM (Özçelik et al. 2024)	Small Molecule	S4-based SSM	MAPk1 kinase inhibitors	fine-tuning for transfer to desired class
DiffuMol (Peng and Zhu 2024)	Small Molecule	Diffusion on embeddings, Transformer decoder for denoising	LogP, QED, TPSA, SAS, and scaffolds	Noiseless conditional token anchors
Regression Transformer(Born and Manica 2023)	Small Molecule and Protein	Decoder-only Transformer (XLNet)	Control QED, solubility, lipophilicity for small molecules, fluorescence, stability and Bowan index for proteins	Alternating training scheme that switches between property prediction and conditional sequence generation
Progen (Madani et al. 2023)	Protein	Decoder-only Transformer	Proteins associated with specific families, biological functions	Conditional autoregressive generation and fine-tuning
ProGen2 (Nijkamp et al. 2023)	Protein	Decoder-only Transformer with RoPE(Su et al. 2024)	Structurally valid proteins, specific folds, antibodies	Fine-tuning, and three-residue motif prompts for antibodies
ProtGPT2 (Ferruz, Schmidt, and Höcker 2022)	Protein	Decoder-only Transformer	Novel proteins that are also plausible, mostly globular proteins	Autoregressive sampling with modified sampling schemes
ProtMamba (Sgarbossa, Malbranke, and Bitbol 2024)	Protein	Mamba-based SSM	Homologous proteins, valid inpainting-based protein modifications	Autoregressive with homologs for context, or inpainting using Fill-in-the-Middle
EvoDiff (Alamdari et al. 2023)	Protein	Diffusion with Dilated CNN	Proteins from specific families, inpainting functional domains, generating around motifs	MSA-conditioned for family-based, motif-based conditioning and inpainting for scaffolds
NOS (Gruver et al. 2024)	Protein	BERT-based Encoder-Decoder	Antibodies with high expression yield and binding affinity	Diffusion, multi property value function, Latent Multi-Objective Bayesian Optimization
RFDiffusionAA (Krishna et al. 2024)	Protein complexes	Specialized Multi-track attention based architecture	Protein binders for small molecules, including nucleic acids, proteins, and ligand complexes	Condition on ligand as fixed noiseless anchor
MMDIFF (Morehead et al. 2023)	DNA and Protein	One-hot vector for discrete sequences, FrameDiff architecture for 3d structure	Macromolecular Complexes	Joint reverse diffusion with separate loss components for sequence and structure
regLM (Lal et al. 2024)	DNA	Based on HyenaDNA(Nguyen et al. 2024b), SSM-like	Cis-regulatory elements with specified levels of activity or cell-type specificity	Special condition tokens for activity level or cell type, outputs filtered using regression model
Evo (Nguyen et al. 2024a)	DNA	StripedHyena using attention and SSM-like blocks	Nucleotide sequences for CRISPR systems, function-conserving transposable elements	Fine-tuning to enable conditioning on special tokens
DiscDiff (Li et al. 2024)	DNA	2-stage VAE, with U-net based denoising networks	DNA for specific species, gene expression levels, or cell types	Conditional generation with absorb-escape algorithm
DNA-diffusion (Senan et al. 2024)	DNA	U-net backbone	Regulatory sequences to control chromatin accessibility	Conditioning on cell-type labels

Specialized for Biological Design EvoDiff (Alamdari et al. 2023) tailors diffusion for protein modeling using two methods: Order Agnostic Autoregressive Diffusion (OADM), which masks tokens randomly, and D3PM, which introduces mutation-based noise guided by natural mutation probabilities. This approach enhances the generation of functional, stable, and active proteins. Hybrid models combine diffusion and Transformers to capture broader contexts. NOS (Gruver et al. 2024) uses a BERT-small encoder-decoder for protein sequences, while DiffuMol (Peng and Zhu 2024) employs a Transformer decoder for small-molecule generation with properties like QED and LogP. In multi-modal settings, RFDiffusionAA adapts RoseTTAFold All-Atom for generating protein-molecule complexes using sequence and structure. MMDIFF (Morehead et al. 2023) jointly diffuses DNA and protein sequences, aligning sequence and structural losses for coherent macromolecular complex generation.

Controllability in Generation

A key challenge in advancing FMs for biological design is achieving fine-grained control over generated data. We note that biological design is inherently an engineering endeavor. The designed entities are intended to be ultimately synthesized in wet laboratories and then operationalized for particular outcomes. So, the issue of control is inherent to successful, synthetically-accessible and operationalizable design.

For small molecules, continuous numerical properties may need to be controlled, including Drug-likeness, LogP, Molecular Weight, Synthetic Accessibility, Toxicity, and Topological Surface Area. One may require particular classes of molecular compounds, such as "kinase inhibitors" or "Quaternary Ammonium Compounds" depending on the downstream task/application. For proteins, one typically controls for specific functions or similarity to other known proteins. For DNA, one may ensure specific regulatory properties, such as promoters or enhancers, or design guide RNAs (gRNAs) to facilitate CRISPR-based gene editing.

Fine-Tuning

Fine-tuning adapts pre-trained FMs for specific tasks by training on smaller, targeted datasets. For instance, Özçelik et al. (2024) fine-tune an S4-based CLM pre-trained on 1.9 million SMILES to generate MAPK1 inhibitors from just 68 annotated molecules. Similarly, Madani et al. (2023) fine-tune ProGen on lysozyme families, generating proteins with catalytic efficiencies comparable to natural lysozymes despite low sequence identity. While essential for controlled generation, fine-tuning can be computationally expensive, especially for large models like ProGen with 1.2 billion parameters. It may also struggle with very small datasets relative to the pre-training corpus, limiting its effectiveness for granular control.

Conditional Generation

Conditional generation embeds prompts or control tags into inputs to guide outputs toward desired characteristics. RegLM (Lal et al. 2024) designs synthetic cis-regulatory elements (e.g., promoters, enhancers) by conditioning on starter fragments, while Evo (Nguyen et al. 2024a) optimizes gRNA sequences using prompt tokens like cas9" or cas13" to improve efficiency and specificity. Progen2 generates antibody sequences conditioned on motif prompts, ProtMamba creates homologous proteins based on family context, and RFDiffusionAA (Krishna et al. 2024) uses unaltered ligands as diffusion constraints for protein-ligand complex generation. MolGPT (Bagal et al. 2021) embeds control tokens in pre-training inputs but retrains for each new property, deviating from the general FM paradigm. Despite enabling control, conditional generation is limited by its reliance on training examples, constraining novelty and potentially biasing the model toward features correlated with target properties, which can reduce diversity in outputs.

Reinforcement Learning

Reinforcement learning (RL) guides generation through reward functions, optimizing objectives like bioactivity and molecular diversity. In sequence generation, tokens are treated as actions, and the model is fine-tuned based on how well sequences meet desired conditions. RL's ability to handle non-linear rewards makes it ideal for multi-objective optimization. The MCMG model (Wang et al. 2021) uses RL to balance molecular properties. It begins with a conditional Transformer and distills it into a simpler recurrent model fine-tuned via RL, optimizing bioactivity, drug-likeness, and synthetic accessibility. Invalid molecules receive zero rewards, incorporated into an augmented likelihood loss for property optimization. Similarly, Mazuz et al. (2023) combine Transformers with RL, using policy gradients to optimize properties like QED and pIC50. Rewards are applied only to valid molecules, prioritizing near-term returns with a high discount factor. However, RL's practicality is limited by the complexity of designing effective reward functions, especially in high-dimensional biological spaces.

Multi-Condition Generation and Tradeoff

Generating molecules with multiple properties is challenging, particularly for rare combinations constrained by physical or chemical limitations. Models must balance validity, diversity, and novelty, which often conflict. For example, antibacterial drug design requires novel molecules to counter resistance but risks reducing validity.

Techniques like adjusting autoregressive Transformer parameters (e.g., temperature) allow users to prioritize validity or uniqueness. MCMG (Wang et al. 2021) uses reinforcement learning to optimize molecular properties while maintaining diversity. DiffuMol (Peng and Zhu 2024) employs diffusion modeling with attention mechanisms to balance property optimization and molecular diversity, focusing on key regions while allowing scaffold variation.

Multi-Modal FMs

While FMs often excel at learning from a single data modality, biological systems are inherently multi-modal, involving interactions across sequence, structure, and function at various scales. Effectively integrating these diverse data types

remains an open challenge in biological modeling. Often, biological data sources are treated independently, overlooking the interactions between systems. DNA sequences are often viewed in isolation, but DNA alone does not fully describe an organism's phenotype. As Denis Noble(Noble 2012, 2024) and others emphasize, biological processes result from interactions across molecular, cellular, physiological and organismal scales(Ramsden 2023).

Combining Multiple Forms of Biological Sequences

Multimodal approaches enhance GLMs by integrating epigenetic and transcriptomic data. For instance, gLM2 (Cornman et al. 2024), trained on the Open MetaGenomic corpus, combines nucleotide sequences, amino acid sequences, and strand direction for tasks like gene regulation.

Hybrid methods are increasingly used in biological design. MMDiff (Morehead et al. 2023) integrates sequence and structural data to generate nucleic acid-protein complexes. RFAA (Krishna et al. 2024) models complex biomolecular systems with a three-track architecture incorporating 1D sequences, pairwise relationships, and 3D structures. It also uses conditional diffusion fine-tuning to generate binding proteins from substructures like ligands.

Combining Sequence and Structure

Integrating structural information with sequence data enhances performance in tasks reliant on geometry and physical constraints. Molecular graphs and protein structures (secondary to quaternary) add essential spatial context missing from sequences alone. Additional inputs like biological environments, target receptors, or transcriptomic data further enrich models.

For example, MMDiff (Morehead et al. 2023) combines sequence and structural data to model nucleic acid-protein complexes, addressing challenges like Intrinsically Disordered Regions. ProtMamba (Sgarbossa, Malbranke, and Bitbol 2024) uses homologous sequences without MSAs for protein inpainting and de novo generation, capturing evolutionary context. While Progen2 (Nijkamp et al. 2023) partially utilizes Gene Ontology (GO) terms, its rich graph structure, linking biological processes and molecular functions, remains underexplored for generation.

Incorporating Natural Language and Domain Knowledge

A key advantage of FMs like GPT-4 is their ability to interact via natural language, offering flexibility absent in biology-specific models. Recent efforts integrate natural language with biological sequence models, enabling tasks through prompts. Examples include ChatNT (Richard et al. 2024) for genomics, and Nach0 (Livne et al. 2024) and InstructBioMol (Zhuang et al. 2024) for chemical sequences.

ChatNT combines biological sequence processing (DNA, RNA, proteins) with NLP via the Nucleotide Transformer (Dalla-Torre et al. 2023) and Vicuna-7B, allowing conversational interactions. Nach0 applies this to SMILES molecular sequences for tasks like molecular property prediction and generation. For instance, when prompted to design a JAK3 inhibitor, Nach0 generated eight valid molecules, achieving a discovery rate of 0.11%, compared to 1.53% for a structure-aware baseline. InstructBioMol further expands capabilities by incorporating natural language, 2D molecular graphs, protein sequences, and 3D structures, enabling molecule captioning, description-based generation, and protein property Q&A.

Open Problems

Which Architectures for Which Biological Tasks: Despite the proliferation of architectures (e.g., Transformers, diffusion, SSMs) tailored to specific tasks, there is no consensus on the best choices across biological domains. Challenges include inconsistent benchmarks and evaluations. Exploring underused architectures, such as diffusion models for textual molecular data, and developing domain-specific self-supervised objectives could improve representation learning, especially in sequence-only or multi-modal settings.

Innovative Approaches for Data Limitations: Scalability and generalization in biological FMs are hindered by limited and fragmented datasets. While NLP breakthroughs leveraged vast data, biology requires alternative solutions, such as lighter architectures or new strategies to achieve generalizable representations without massive datasets.

Enhancing Transfer to Low-Data Regimes: Transfer learning is critical for underrepresented molecule classes, such as Quaternary Ammonium Compounds with antibiotic properties but limited data. Techniques like fine-tuning and conditional generation, inspired by ProGen2 (Nijkamp et al. 2023), can address pre-training biases and guide models toward specific regions of chemical space.

Integrating Biological Modalities: Integrating modalities like chemical graphs with SMILES in CLMs or Gene Ontology (GO) annotations for proteins offers new opportunities. Techniques like Adapters (Liu et al. 2023) could enhance molecule generation, and extending GO-based approaches could refine pre-training and downstream tasks.

Improving Control in Generation: Controlling FMs to generate rare or unlikely property combinations in biological sequence space remains a challenge. Investigating the root causes—whether due to data scarcity or physical constraints—can inform strategies to expand coverage of "dark" protein space (Ferruz, Schmidt, and Höcker 2022) and improve the generation of uncommon combinations.

Conclusion

The field is moving ahead rapidly; a large number of the cited papers here are from preprint servers such as arXiv and BioRxiv, highlighting the need for standardization of datasets and metrics. Furthermore, looking ahead, it is crucial for research to prioritize rigorous comparative evaluations of different model architectures for biological generation, more sophisticated integration of biological modalities, and improved strategies for generalization across diverse biological systems. This is especially timely, as the United States National Academies of Sciences, Engineering, and Medicine are founding a committee on Foundation Models for Scientific Discovery and Innovation, un-

derscoring the growing importance of this area.(National Academies of Sciences and Medicine 2024) By addressing these challenges, FMs could truly revolutionize fields such as drug discovery, synthetic biology, and genetic engineering, accelerating breakthroughs and moving us closer to practical, AI-driven biological innovations that can meaningfully impact our understanding and manipulation of life.

Acknowledgements

This work was supported in part by the National Science Foundation Grant No. 2411529 and Grant No. 2310113.

References

2000–2015. Protein Structure Initiative. National Institute of General Medical Sciences, U.S. National Institutes of Health. Available from https://www.nigms.nih.gov/Research/specificareas/PSI.

Alamdari, S.; Thakkar, N.; van den Berg, R.; Lu, A.; Fusi, N.; Amini, A.; and Yang, K. 2023. Protein generation with evolutionary diffusion: sequence is all you need. In *NeurIPS 2023 Generative AI and Biology (GenBio) Workshop*.

Bagal, V.; Aggarwal, R.; Vinod, P.; and Priyakumar, U. D. 2021. MolGPT: molecular generation using a transformer-decoder model. *Journal of Chemical Information and Modeling*, 62(9): 2064–2076.

Bank, P. D. 1971. Protein data bank. *Nature New Biol*, 233(223): 10–1038.

Bavarian, M.; Jun, H.; Tezak, N.; Schulman, J.; McLeavey, C.; Tworek, J.; and Chen, M. 2022. Efficient training of language models to fill in the middle. *arXiv preprint arXiv:2207.14255*.

Blanco-Gonzalez, A.; Cabezon, A.; Seco-Gonzalez, A.; Conde-Torres, D.; Antelo-Riveiro, P.; Pineiro, A.; and Garcia-Fandino, R. 2023. The role of AI in drug discovery: challenges, opportunities, and strategies. *Pharmaceuticals*, 16(6): 891.

Bommasani, R.; Hudson, D. A.; Adeli, E.; Altman, R.; Arora, S.; von Arx, S.; Bernstein, M. S.; Bohg, J.; Bosselut, A.; Brunskill, E.; et al. 2021. On the Opportunities and Risks of Foundation Models. *arXiv e-prints*, arXiv–2108.

Born, J.; and Manica, M. 2023. Regression transformer enables concurrent sequence regression and generation for molecular language modelling. *Nature Machine Intelligence*, 5(4): 432–444.

Collins, F. S.; and Fink, L. 1995. The Human Genome Project. *Alcohol Health and Research World*, 19(3): 190–195.

Cornman, A.; West-Roberts, J.; Camargo, A. P.; Roux, S.; Beracochea, M.; Mirdita, M.; Ovchinnikov, S.; and Hwang, Y. 2024. The OMG dataset: An Open MetaGenomic corpus for mixed-modality genomic language modeling. *bioRxiv*, 2024–08.

Dalla-Torre, H.; Gonzalez, L.; Mendoza-Revilla, J.; Carranza, N. L.; Grzywaczewski, A. H.; Oteri, F.; Dallago, C.; Trop, E.; de Almeida, B. P.; Sirelkhatim, H.; et al. 2023. The nucleotide transformer: Building and evaluating robust foundation models for human genomics. *BioRxiv*, 2023–01.

Dawson, N. L.; Sillitoe, I.; Lees, J. G.; Lam, S. D.; and Orengo, C. A. 2017. CATH-Gene3D: generation of the resource and its use in obtaining structural and functional annotations for protein sequences. *Protein Bioinformatics: From Protein Modifications and Networks to Proteomics*, 79–110.

Ferruz, N.; Schmidt, S.; and Höcker, B. 2022. ProtGPT2 is a deep unsupervised language model for protein design. *Nature communications*, 13(1): 4348.

Gruver, N.; Stanton, S.; Frey, N.; Rudner, T. G.; Hotzel, I.; Lafrance-Vanasse, J.; Rajpal, A.; Cho, K.; and Wilson, A. G. 2024. Protein design with guided discrete diffusion. *Advances in neural information processing systems*, 36.

Gu, A.; and Dao, T. 2023. Mamba: Linear-time sequence modeling with selective state spaces. *arXiv preprint arXiv:2312.00752*.

Gu, A.; Goel, K.; and Ré, C. 2021. Efficiently modeling long sequences with structured state spaces. *arXiv preprint arXiv:2111.00396*.

Hoogeboom, E.; Gritsenko, A. A.; Bastings, J.; Poole, B.; van den Berg, R.; and Salimans, T. ???? Autoregressive Diffusion Models. In *International Conference on Learning Representations*.

Hoogeboom, E.; Nielsen, D.; Jaini, P.; Forré, P.; and Welling, M. 2021. Argmax flows and multinomial diffusion: Learning categorical distributions. *Advances in Neural Information Processing Systems*, 34: 12454–12465.

Jones, M. M.; Castle-Clarke, S.; Brooker, D.; Nason, E.; Huzair, F.; and Chataway, J. 2014. The structural genomics consortium: a knowledge platform for drug discovery: a summary. *Rand health quarterly*, 4(3).

Jumper, J.; Evans, R.; et al. 2021. Highly accurate protein structure prediction with AlphaFold. *nature*, 596(7873): 583–589.

Krishna, R.; Wang, J.; Ahern, W.; Sturmfels, P.; Venkatesh, P.; Kalvet, I.; Lee, G. R.; Morey-Burrows, F. S.; Anishchenko, I.; Humphreys, I. R.; et al. 2024. Generalized biomolecular modeling and design with RoseTTAFold All-Atom. *Science*, 384(6693): eadl2528.

Lal, A.; Garfield, D.; Biancalani, T.; and Eraslan, G. 2024. regLM: Designing realistic regulatory DNA with autoregressive language models. In *International Conference on Research in Computational Molecular Biology*, 332–335. Springer.

Li, X.; Thickstun, J.; Gulrajani, I.; Liang, P. S.; and Hashimoto, T. B. 2022. Diffusion-lm improves controllable text generation. *Advances in Neural Information Processing Systems*, 35: 4328–4343.

Li, Z.; Ni, Y.; Beardall, W. A.; Xia, G.; Das, A.; Stan, G.-B.; and Zhao, Y. 2024. DiscDiff: Latent Diffusion Model for DNA Sequence Generation. *CoRR*.

Liu, Z.; Li, S.; Luo, Y.; Fei, H.; Cao, Y.; Kawaguchi, K.; Wang, X.; and Chua, T.-S. 2023. MolCA: Molecular Graph-Language Modeling with Cross-Modal Projector and Uni-Modal Adapter. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, 15623–15638.

Livne, M.; Miftahutdinov, Z.; Tutubalina, E.; Kuznetsov, M.; Polykovskiy, D.; Brundyn, A.; Jhunjhunwala, A.; Costa, A.; Aliper, A.; Aspuru-Guzik, A.; et al. 2024. nach0: Multimodal natural and chemical languages foundation model. *Chemical Science*, 15(22): 8380–8389.

Madani, A.; Krause, B.; Greene, E. R.; Subramanian, S.; Mohr, B. P.; Holton, J. M.; Olmos, J. L.; Xiong, C.; Sun, Z. Z.; Socher, R.; et al. 2023. Large language models generate functional protein sequences across diverse families. *Nature Biotechnology*, 41(8): 1099–1106.

Mazuz, E.; Shtar, G.; Shapira, B.; and Rokach, L. 2023. Molecule generation using transformers and policy gradient reinforcement learning. *Scientific Reports*, 13(1): 8799.

Morehead, A.; Ruffolo, J.; Bhatnagar, A.; and Madani, A. 2023. Towards joint sequence-structure generation of nucleic acid and protein complexes with SE (3)-discrete diffusion. In *Proceedings of the NeurIPS 2023 Workshop on Machine Learning in Structural Biology*.

Moret, M.; Friedrich, L.; Grisoni, F.; Merk, D.; and Schneider, G. 2020. Generative molecular design in low data regimes. *Nature Machine Intelligence*, 2(3): 171–180.

National Academies of Sciences, E.; and Medicine. 2024. Foundation Models for Scientific Discovery and Innovation: Opportunities Across the Department of Energy. Accessed: 2024-10-22.

Nguyen, E.; Poli, M.; Durrant, M.; Thomas, A.; Kang, B.; Sullivan, J.; Ng, M.; Lewis, A.; Patel, A.; Lou, A.; et al. 2024a. Sequence modeling and design from molecular to genome scale with Evo.

Nguyen, E.; Poli, M.; Faizi, M.; Thomas, A.; Wornow, M.; Birch-Sykes, C.; Massaroli, S.; Patel, A.; Rabideau, C.; Bengio, Y.; et al. 2024b. Hyenadna: Long-range genomic sequence modeling at single nucleotide resolution. *Advances in neural information processing systems*, 36.

Nijkamp, E.; Ruffolo, J. A.; Weinstein, E. N.; Naik, N.; and Madani, A. 2023. Progen2: exploring the boundaries of protein language models. *Cell systems*, 14(11): 968–978.

Noble, D. 2012. A theory of biological relativity: no privileged level of causation. *Interface focus*, 2(1): 55–64.

Noble, D. 2024. It's time to admit that genes are not the blueprint for life. *Nature*, 626(7998): 254–255.

Özçelik, R.; de Ruiter, S.; Criscuolo, E.; and Grisoni, F. 2024. Chemical language modeling with structured state space sequence models. *Nature Communications*, 15(1): 6176.

Peng, X.; and Zhu, F. 2024. Hitting stride by degrees: Fine grained molecular generation via diffusion model. *Expert Systems with Applications*, 244: 122949.

Poli, M.; Massaroli, S.; Nguyen, E.; Fu, D. Y.; Dao, T.; Baccus, S.; Bengio, Y.; Ermon, S.; and Ré, C. 2023. Hyena hierarchy: Towards larger convolutional language models. In *International Conference on Machine Learning*, 28043–28078. PMLR.

Ramsden, J. 2023. *Bioinformatics: an introduction*. Springer Nature.

Richard, G.; de Almeida, B. P.; Dalla-Torre, H.; Blum, C.; Hexemer, L.; Pandey, P.; Laurent, S.; Lopez, M. P.; Laterre, A.; Lang, M.; et al. 2024. ChatNT: A Multimodal Conversational Agent for DNA, RNA and Protein Tasks. *bioRxiv*, 2024–04.

Senan, S.; Reddy, A. J.; Nussbaum, Z.; Wenteler, A.; Bejan, M.; Love, M. I.; Meuleman, W.; and Pinello, L. 2024. DNA-Diffusion: Leveraging Generative Models for Controlling Chromatin Accessibility and Gene Expression via Synthetic Regulatory Elements. In *ICLR 2024 Workshop on Machine Learning for Genomics Explorations*.

Sgarbossa, D.; Malbranke, C.; and Bitbol, A.-F. 2024. ProtMamba: a homology-aware but alignment-free protein state space model. *bioRxiv*, 2024–05.

Su, J.; Ahmed, M.; Lu, Y.; Pan, S.; Bo, W.; and Liu, Y. 2024. Roformer: Enhanced transformer with rotary position embedding. *Neurocomputing*, 568: 127063.

Tang, T.-C.; An, B.; Huang, Y.; Vasikaran, S.; Wang, Y.; Jiang, X.; Lu, T. K.; and Zhong, C. 2021. Materials design by synthetic biology. *Nature Reviews Materials*, 6(4): 332–350.

Vaswani, A. 2017. Attention is all you need. *arXiv preprint arXiv:1706.03762*.

Voigt, C. A. 2020. Synthetic biology 2020–2030: six commercially-available products that are changing our world. *Nature Communications*, 11(1): 1–6.

Wang, J.; Hsieh, C.-Y.; Wang, M.; Wang, X.; Wu, Z.; Jiang, D.; Liao, B.; Zhang, X.; Yang, B.; He, Q.; et al. 2021. Multi-constraint molecular generation based on conditional transformer, knowledge distillation and reinforcement learning. *Nature Machine Intelligence*, 3(10): 914–922.

Yang, Z. 2019. XLNet: Generalized Autoregressive Pretraining for Language Understanding. *arXiv preprint arXiv:1906.08237*.

Zhuang, X.; Ding, K.; Lyu, T.; Jiang, Y.; Li, X.; Xiang, Z.; Wang, Z.; Qin, M.; Feng, K.; Wang, J.; et al. 2024. InstructBioMol: Advancing Biomolecule Understanding and Design Following Human Instructions. *arXiv preprint arXiv:2410.07919*.