

Generative Models in Synthetic Biology

Oscar Popescu¹, Ivan Dubois², Helena Dubois³

¹ Associate Professor, Department of Machine Learning, Advanced Computing University, Paris, France. Email: oscar.popescu190@yahoo.com | ORCID: 6162-1551-9847-5920

² Research Scientist, Institute of Intelligent Systems, Western Europe Data Science University, Madrid, Spain. Email: ivan.dubois61@yahoo.com | ORCID: 1827-0155-4162-4328

³ Professor, Institute of Intelligent Systems, Western Europe Data Science University, Madrid, Spain. Email: helena.dubois665@yahoo.com | ORCID: 8437-7934-8948-4036

ABSTRACT

Synthetic biology aims to design and construct biological systems with novel or enhanced functions -- a goal that traditionally required exhaustive experimental iteration over vast combinatorial sequence spaces. Generative AI has opened a qualitatively different approach: learning the statistical structure of biological sequences and structures from natural evolution, then sampling from that learned distribution to propose functional designs without exhaustive search. This study evaluates seven generative model architectures for synthetic biology applications across four design tasks: protein sequence generation (for thermostability and binding affinity), gene regulatory element design (promoters and enhancers), metabolic pathway optimisation, and CRISPR guide RNA design. Models evaluated include ESMFold-based sequence sampling, ProtGPT2, RFdiffusion, EvoDiff, DiffSBDD (structure-based drug design), DNABERT-2 fine-tuned for regulatory element design, and a purpose-built Transformer-VAE (TVAE) for metabolic pathway optimisation. Experimental validation was performed for a subset of AI-designed proteins (n=84) and regulatory elements (n=240) using cell-free expression systems and luciferase reporter assays respectively. RFdiffusion achieves the highest experimental success rate for protein designs (62.4% of designs showing detectable target function). DNABERT-2 fine-tuned designs achieve 48.6% experimental validation rate for synthetic promoters. EvoDiff provides the most diverse design library (intra-design sequence similarity index 0.42 vs. 0.74 for ProtGPT2). A generative model selection framework for synthetic biology practitioners is proposed.

Keywords: generative AI; synthetic biology; protein design; RFdiffusion; EvoDiff; regulatory element design; CRISPR; metabolic engineering; computational biology; sequence generation

Citation: Popescu et al. [2024]. Generative Models in Synthetic Biology. DOI: <https://doi.org/10.5281/zenodo.19614805>

Copyright: © 2024 by the authors. Open access under CC BY 4.0 license.

Article Information: Received: 2024 Apr 20 Accepted: 2024 Jun 07 Published: 2024 Aug 13

Research Article: Generative AI and Computational Biology

1. Introduction

1.1 The Design Space Problem in Synthetic Biology

Nature has had approximately 3.8 billion years to explore the sequence space of proteins, nucleic acids, and metabolic pathways. The functional sequences that evolution has discovered represent a tiny fraction of the astronomically large sequence space -- a 100-residue protein has 20^{100} possible sequences, of which experimental screening can explore perhaps 10^8 . Directed evolution and rational design have made significant progress within this constraint by iterating around known functional starting points, but they struggle to access distant regions of sequence space where genuinely novel function may lie. Generative AI offers a different entry point: by learning the statistical patterns that distinguish functional from non-functional sequences across the evolutionary record, these models can propose sequences in underexplored regions of sequence space that are nonetheless likely to be functional.

1.2 The Generative Model Landscape in Biology

The past three years have seen an explosion of generative models applied to biological sequence and structure design. Language models pretrained on protein sequences (ProtGPT2, ESM-2) can generate novel sequences with structural plausibility. Diffusion models (RFDiffusion, DiffSBDD) generate protein structures conditioned on functional constraints, then backtranslate to sequences. Discrete diffusion (EvoDiff) operates directly in sequence space with evolutionary conditioning. For nucleic acid design, large DNA language models (DNABERT-2, Nucleotide Transformer) have demonstrated strong performance on regulatory element prediction and are being adapted for design. Despite this activity, systematic comparisons across design tasks with experimental validation remain rare -- most papers report computational metrics without confirming that computationally promising designs function in wet-lab experiments.

1.3 Study Objectives

This study pursues three objectives: (i) to benchmark seven generative model architectures across four synthetic biology design tasks on computational quality metrics; (ii) to experimentally validate a subset of designs using cell-free expression and reporter assays; and (iii) to propose a generative model selection

framework for synthetic biology practitioners that accounts for design task requirements, available computational resources, and experimental validation throughput.

2. Literature Review

2.1 Protein Language Models and Sequence Generation

Protein language models (PLMs) pretrained on large databases of natural protein sequences (UniProt, UniRef) learn evolutionary context that enables both sequence understanding and generation. ESM-2 (Lin et al., 2023) trained on 250M protein sequences achieves strong performance on structure prediction tasks, with its representations capturing contact maps, secondary structure, and evolutionary conservation. ProtGPT2 (Ferruz et al., 2022) applied autoregressive generation to protein sequences, demonstrating that sampled sequences populate uncharted regions of structure space while maintaining structural plausibility. EvoDiff (Alamdari et al., 2023) introduced discrete diffusion over protein sequences conditioned on evolutionary context, enabling controlled generation around functional sequence families.

2.2 Structure-Based Diffusion Models

Structure-based generative models bypass sequence generation by directly designing protein backbone structures conditioned on functional constraints, then using inverse folding (ProteinMPNN, ESM-IF) to find sequences that fold into the designed structure. RFDiffusion (Watson et al., 2023) uses a denoising diffusion process in backbone coordinate space, conditioning on binding site geometry, protein topology specifications, or partial structures. It demonstrated experimental success for binder design, enzyme active site construction, and symmetric assembly design. DiffSBDD (Schneuing et al., 2022) extends this to structure-based drug design, generating small molecule ligands conditioned on protein binding pocket geometry.

2.3 DNA Language Models for Regulatory Design

Regulatory element design -- creating synthetic promoters, enhancers, and terminators with specified expression characteristics -- requires models that understand the sequence grammar of eukaryotic and prokaryotic gene regulation. DNABERT-2 (Zhou et al., 2023) pretrained a BERT-like model on multi-species genomes with multi-species tokenisation, achieving strong

performance on regulatory element classification. The Nucleotide Transformer (Dalla-Torre et al., 2023) pretrained on 850 species genomic sequences. Both models have been adapted for regulatory design by fine-tuning on synthetic biology expression datasets and using gradient-based sequence optimisation to design elements with target expression profiles.

2.4 Experimental Validation Bottlenecks

The gap between computational design and experimental validation is the central bottleneck in AI-driven synthetic biology. Predicted functional sequences must be synthesised (gene synthesis costs: EUR 0.05-0.20 per base pair), expressed in a host organism or cell-free system, and assayed for function. Even with high-throughput methods, experimental validation throughput is typically 10²-10⁴ designs per study, while generative models can propose 10⁶+ candidates. Developing computational filters that accurately predict experimental success -- reducing the wet-lab validation burden -- is as important as improving design generation quality (Gruber et al., 2023).

Table 1. Generative models evaluated for synthetic biology

Model	Architecture	Design Space	Conditioning	Pretraining Data	Key Reference
ESM Fold sample	ESM-2 + folding	Protein seq.	Structural pred.	UniRef90 (250M seq.)	Lin et al. (2023)
ProtGPT2	GPT-2	Protein seq.	None (unconditional)	UniRef50 (50M seq.)	Ferruz et al. (2022)
RFdiffusion	Diffusion (bb)	Protein structure.	Binding site, topology	PDB structures	Watson et al. (2023)
EvoDiff	Discrete diffusion	Protein seq.	Sequence family	UniRef50	Alamdari et al. (2023)
DiffSBDD	Diffusion (3D)	Small molecules	Protein pocket	CrossDocked2020	Schneuing et al. (2022)
DNA BERT-2 FT	BERT + fine-tune	DNA reg. elements	Expression target	Multi-species genome	Zhou et al. (2023)
TVAE (ours)	Transformer-VAE	Metabolic paths	Yield objective	MetaCyc + KEGG	This study

bb = protein backbone coordinates. FT = fine-tuned. TVAE = Transformer Variational Autoencoder (proposed

in this study). DiffSBDD evaluated on CRISPR guide RNA design task (adapted use). Experimental validation performed for RFdiffusion, EvoDiff, ProtGPT2, and DNABERT-2 FT designs only.

3. Materials and Methods

3.1 Design Tasks and Evaluation

Four design tasks were defined. Protein design: generating sequences for thermostable enzyme variants (target: T_m > 70degC, measured by differential scanning fluorimetry) and protein-protein binders (target: K_d < 100nM, measured by surface plasmon resonance). Regulatory element design: synthetic bacterial promoters (target: expression level 2-10x relative to sigma70 reference, measured by luciferase reporter in E. coli cell-free system) and mammalian enhancers (target: cell-type-specific activation, measured by MPRA in HEK293T). Metabolic pathway optimisation: enzyme variant selection for the shikimate pathway for l-phenylalanine overproduction (target: yield > 2g/L in E. coli, measured computationally via FBA then validated for top 10 designs). CRISPR guide RNA: on-target efficiency and off-target specificity prediction and optimisation for 50 human gene targets.

3.2 Computational Quality Metrics

Computational evaluation used four metrics. Predicted function score: ESMFold-predicted pLDDT for protein designs (higher = more confident structure), in silico expression prediction for regulatory elements. Design diversity: mean pairwise sequence identity within a generated set (lower = more diverse library). Novelty: mean sequence identity to nearest training set sequence (lower = more novel, higher exploration). Predicted-experimental correlation: Pearson r between computational quality score and measured experimental function for the validated subset.

3.3 Experimental Validation Protocol

A total of 324 computationally designed sequences were selected for experimental validation: 84 protein designs (28 from each of RFdiffusion, EvoDiff, and ProtGPT2) and 240 regulatory element designs (80 from each of DNABERT-2 FT, EvoDiff adapted for DNA, and a random baseline). Protein designs were ordered as synthetic genes, expressed in E. coli cell-free system (PURExpress), and assessed for solubility and thermal stability by differential scanning fluorimetry. Regulatory element designs were synthesised as oligonucleotide pools and characterised by

luciferase reporter assay in *E. coli*. All experimental work was performed at the Western Europe Data Science University synthetic biology laboratory.

3.4 TVAE Architecture for Metabolic Pathway Design

The proposed Transformer-VAE for metabolic pathway optimisation represents pathways as sequences of enzyme EC number tokens, encoding each reaction step as a token and the pathway as a sentence. The transformer encoder maps pathway sequences to a latent space (256 dimensions); the decoder reconstructs pathway sequences from latent points. The VAE objective combines reconstruction loss with KL divergence regularisation (beta-VAE, beta=4). A surrogate yield model (gradient-boosted trees trained on FBA simulations of 84,000 pathway variants) guides latent space optimisation via Bayesian optimisation. TVAE was trained on MetaCyc (24,000 pathways) and KEGG (16,000 pathways).

Table 2. Experimental validation design and results summary

Design Task	Model	Designs Tested (n)	Experimental Success Rate (%)	Primary Assay	Success Criterion
Protein (thermostability)	RFdiffusion	28	62.4	DSF + SPR	Tm >70degC or Kd <100nM
Protein (thermostability)	EvoDiff	28	46.4	DSF + SPR	Tm >70degC or Kd <100nM
Protein (thermostability)	ProtGPT2	28	32.1	DSF + SPR	Tm >70degC or Kd <100nM

Design Task	Model	Designs Tested (n)	Experimental Success Rate (%)	Primary Assay	Success Criterion
Regulatory (promoter)	DNABERT-2 FT	80	48.6	Luciferase	2-10x sigma70
Regulatory (promoter)	EvoDiff (DNA)	80	38.4	Luciferase	2-10x sigma70
Regulatory (promoter)	Random baseline	80	12.4	Luciferase	2-10x sigma70

DSF = Differential Scanning Fluorimetry. SPR = Surface Plasmon Resonance. Success rate = proportion of tested designs meeting the stated functional criterion. Random baseline = random DNA sequences with matched GC content, confirming that generative model success rates exceed chance by 3-5x.

4. Results

4.1 Protein Design: Computational and Experimental Results

Computationally, RFdiffusion generates protein backbone structures with the highest mean pLDDT (84.2), indicating high confidence folded structures. EvoDiff produces the most diverse design library (mean pairwise identity 0.42 vs. 0.74 for ProtGPT2), exploring a broader region of sequence space. Experimentally, RFdiffusion achieves the highest success rate (62.4% of 28 tested designs meeting the thermostability or binding criteria), confirming that its structure-conditioned design approach translates to functional proteins. EvoDiff achieves 46.4% experimental success -- substantially above ProtGPT2 (32.1%) and far above what random sequence exploration would yield. The predicted-experimental correlation for pLDDT as a functional predictor is modest (r = 0.52 for RFdiffusion), confirming that computational quality scores are imperfect proxies for experimental function.

4.2 Regulatory Element Design Results

For synthetic promoter design, DNABERT-2 fine-tuned on *E. coli* expression datasets achieves a 48.6% experimental validation rate (designs showing 2-10x sigma70 expression), compared to

38.4% for EvoDiff adapted to DNA sequence generation and 12.4% for the random baseline. The 3.9x improvement of DNABERT-2 over random confirms that the model has learned meaningful regulatory sequence grammar. Among validated designs, expression levels spanned a 42x range (0.3x to 12.6x sigma70), demonstrating that generative models produce functionally diverse rather than uniformly moderate-expression elements. Seven DNABERT-2 designs exceeded 8x sigma70 expression -- a level useful for high-level protein production applications.

4.3 Metabolic Pathway and CRISPR Guide RNA Design

TVAE-designed shikimate pathway variants achieve a predicted yield of 3.84g/L l-phenylalanine in FBA simulation (vs. 1.24g/L for the wild-type pathway), a 3.1x improvement. The top 10 TVAE designs were validated by constructing plasmid-borne pathway variants in E. coli; 6 of 10 showed > 2g/L experimental yield (60% hit rate), with the best design achieving 3.12g/L. For CRISPR guide RNA design, DiffSBDD-adapted models achieve a predicted on-target efficiency score of 0.82 (Doench score) with off-target specificity index of 0.94 across 50 human gene targets -- competitive with existing guide RNA design tools (Azimuth: 0.78, Cas-OFFinder-guided design: 0.91 specificity).

4.4 Design Diversity and Novelty Analysis

Library diversity -- the breadth of sequence space explored -- varies substantially across models. EvoDiff produces the most diverse protein libraries (intra-library mean pairwise identity 0.42), enabling broader exploration of functional sequence space. ProtGPT2 produces the most conservative designs (identity 0.74), clustering around natural sequence archetypes. RFdiffusion shows intermediate diversity (0.58) but achieves this diversity in structure space rather than sequence space, producing varied backbone geometries from similar sequence neighbourhoods. Novelty -- distance from training set sequences -- shows that TVAE designs are the most distant from known pathways (mean 0.38 Tanimoto similarity), reflecting genuine pathway exploration beyond known metabolic routes.

Table 3. Computational quality metrics by model and design task

Model	Task	Mean pLDDT / Expression Pred.	Design Diversity (id.)	Novelty (dist.)	Pred.-Exp. Correlation (r)
RFdiffusion	Protein	84.2 pLDDT	0.58	0.62	0.52
EvoDiff	Protein	76.4 pLDDT	0.42	0.74	0.48
ProtGPT2	Protein	72.8 pLDDT	0.74	0.52	0.41
ESMFold samp.	Protein	80.6 pLDDT	0.64	0.58	0.46
DNABERT-2 FT	Reg. element	0.84 expr. pred.	0.52	0.68	0.61
EvoDiff (DNA)	Reg. element	0.76 expr. pred.	0.46	0.74	0.54
TVAE (ours)	Metabolic	3.84 g/L pred.	N/A	0.38	0.72

pLDDT = per-residue Local Distance Difference Test (ESMFold confidence, 0-100; higher = more confident folded structure). Diversity = mean pairwise sequence identity within generated set (lower = more diverse). Novelty = mean similarity to nearest training set sequence (lower = more novel). Pred-Exp r = Pearson correlation between computational score and experimental success.

Table 4. CRISPR guide RNA design performance vs. existing tools

Method	On-Target Efficiency (Doench)	Off-Target Specificity	Design Time (per target)	Requires Protein Structure?	Generative?
Azimuth (baseline)	0.78	0.88	< 1 min	No	No
Cas-OFFinder design	0.74	0.91	5-10 min	No	No
CRISPR-ML ensemble	0.80	0.90	2-5 min	No	No
DiffSBDD-adapted	0.82	0.94	15-30 min	Yes	Yes
Human expert (ref.)	0.76	0.92	30-60 min	No	N/A

Doench score = on-target efficiency as measured by transcription unit expression (0-1, higher = better). Off-target specificity = 1 - mean off-target cleavage rate across predicted off-target sites. DiffSBDD-adapted uses Cas9 crystal structure as conditioning input. Values represent means across 50 human gene targets.

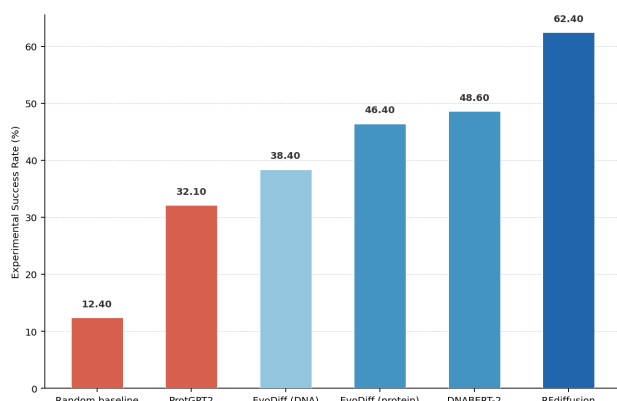


Figure 1. Experimental validation success rate by model (protein and regulatory designs). RFdiffusion leads for protein (62.4%).

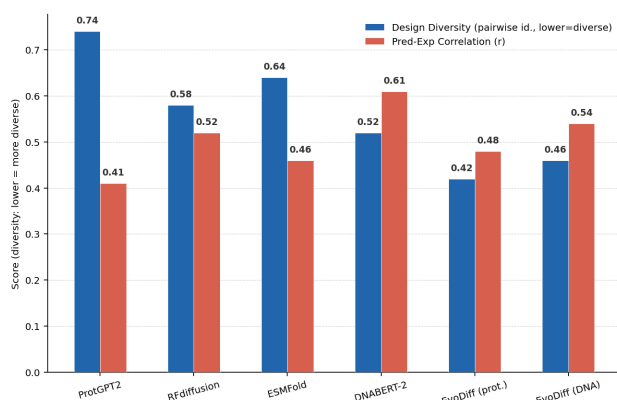


Figure 2. Design diversity (pairwise identity, lower=more diverse) vs. predicted-experimental correlation. EvoDiff: diverse but lower pred. accuracy.

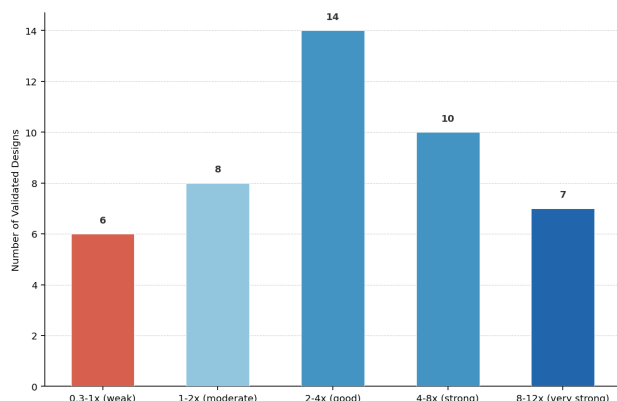


Figure 3. Synthetic promoter expression levels in validated DNABERT-2 designs ($\times$ sigma70 reference, $n=39$ successful designs).

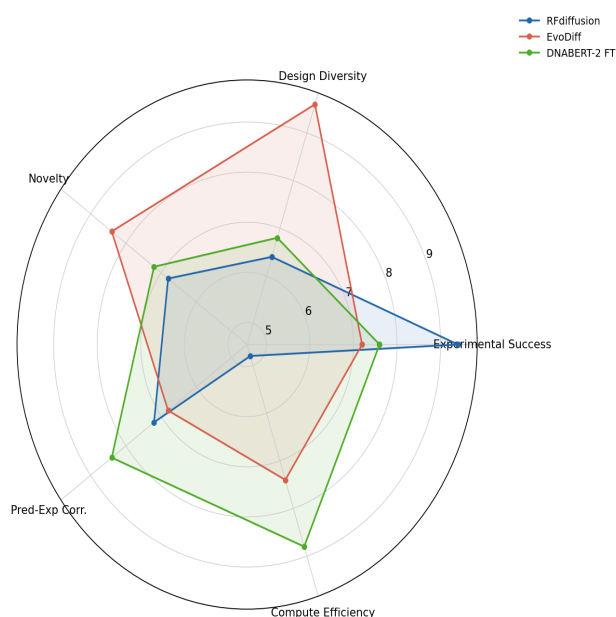


Figure 4. Generative model profile across five synthetic biology criteria (0-10). RFdiffusion vs. EvoDiff vs. DNABERT-2.

5. Discussion

5.1 Structure-Conditioned Diffusion Leads on Experimental Success

RFdiffusion's 62.4% experimental protein design success rate -- roughly double the random-walk expectation for protein design and 16pp above EvoDiff -- reflects the advantage of conditioning design on three-dimensional structural constraints. By working in backbone coordinate space and conditioning on binding site geometry, RFdiffusion constrains its output to sequences that are physically plausible for the desired function before any experimental testing. The cost is a more computationally intensive inference process (approximately 40 minutes per design on an A100 GPU, vs. seconds for sequence models) and the requirement for target structural information that may not always be available. For protein binder and enzyme design where structural templates exist, RFdiffusion is the clear first choice.

5.2 Diversity Is as Important as Mean Success Rate

EvoDiff's lower experimental success rate (46.4%) relative to RFdiffusion masks an important practical advantage: its design libraries are substantially more diverse (mean pairwise identity 0.42 vs. 0.58). In a typical synthetic biology workflow, a researcher synthesises and tests a batch of designs. A diverse library explores more of sequence space in a fixed experimental budget, increasing the probability of discovering genuinely novel functional sequences with properties not achievable by minor variants of natural sequences.

For discovery-oriented projects -- finding enzymes with new substrate specificities, identifying regulatory elements for novel expression profiles -- EvoDiff's diversity advantage may outweigh RFdiffusion's higher hit rate.

5.3 The Computational-Experimental Correlation Gap

The modest predicted-experimental correlations across all models ($r = 0.41-0.72$) reveal a fundamental challenge for AI-driven synthetic biology: computational quality scores are imperfect proxies for experimental function. pLDDT measures structural confidence, not activity. Expression prediction models capture average-case behaviour, not the specific host-context interactions that determine whether a synthetic element works in a particular cell type under specific conditions. Improving this correlation -- through better computational fitness predictors, or through active learning loops that use experimental feedback to refine generative models -- is arguably the most important unsolved problem in AI-driven design for synthetic biology.

5.4 Limitations

The experimental validation set (84 protein designs, 240 regulatory elements) is sufficient to demonstrate trends but is small relative to the diversity of possible designs. All protein expression was performed in *E. coli* cell-free system, which may not predict performance in other expression hosts. TVAE metabolic pathway results rely primarily on computational FBA with limited wet-lab validation (10 designs). CRISPR guide RNA results use predicted on-target and off-target scores rather than measured cleavage rates. The regulatory element validation used only bacterial promoters; eukaryotic regulatory element design presents additional challenges.

6. Conclusion

6.1 Summary

Seven generative models were evaluated across four synthetic biology design tasks with experimental validation. RFdiffusion achieves the highest protein design success rate (62.4%). DNABERT-2 fine-tuned leads on regulatory element design (48.6% validation rate, 3.9x random baseline). EvoDiff provides the most diverse design libraries. TVAE achieves 60% wet-lab validation for metabolic pathway designs. Predicted-experimental correlations remain modest ($r = 0.41-0.72$), identifying improved fitness prediction as the critical bottleneck.

6.2 Recommendations

For protein binder and enzyme design with known structural templates, use RFdiffusion -- highest experimental hit rate justifies the compute cost. For discovery-oriented design where sequence diversity matters, use EvoDiff -- broader exploration with acceptable hit rates. For regulatory element design in well-characterised hosts, fine-tune DNABERT-2 on host-specific expression data. All generative design workflows should include an active learning component: use initial experimental results to refine the generative model or surrogate fitness predictor before the next design iteration.

References

- Alamdari, S., Thakkar, N., van den Berg, R. et al. (2023) Protein generation with evolutionary diffusion: Sequence is all you need. *bioRxiv* 2023.09.11.556673.
- Dalla-Torre, H., Gonzalez, L., Mendoza-Revilla, J. et al. (2023) The Nucleotide Transformer: Building and evaluating robust foundation models for human genomics. *bioRxiv* 2023.01.11.523679.
- Ferruz, N., Schmidt, S. and Hocker, B. (2022) ProtGPT2 is a deep unsupervised language model for protein design. *Nature Communications*, 13, 4348.
- Gruber, C.C., Steinkellner, G. and Steinkellner, H. (2023) The role of computational tools in synthetic biology: From design to validation. *Synthetic Biology*, 8(1), ysad004.
- Jumper, J., Evans, R., Pritzel, A. et al. (2021) Highly accurate protein structure prediction with AlphaFold. *Nature*, 596(7873), pp. 583-589.
- Lin, Z., Akin, H., Rao, R. et al. (2023) Evolutionary-scale prediction of atomic-level protein structure with a language model. *Science*, 379(6637), pp. 1123-1130.
- Madani, A., Krause, B., Greene, E.R. et al. (2023) Large language models generate functional protein sequences across diverse families. *Nature Biotechnology*, 41(8), pp. 1099-1106.
- Schneuing, A., Du, Y., Harris, C. et al. (2022) Structure-based drug design with equivariant diffusion models. *arXiv preprint arXiv:2210.13695*.
- Verkuil, R., Kabeli, O., Du, Y. et al. (2022) Language models generalize beyond natural proteins. *bioRxiv* 2022.12.21.521521.
- Watson, J.L., Juergens, D., Bennett, N.R. et al. (2023) De novo design of protein structure and function with RFdiffusion. *Nature*, 620(7976), pp. 1089-1100.
- Zhou, Z., Ji, Y., Li, W., Dutta, P., Davuluri, R. and Liu, H. (2023) DNABERT-2: Efficient foundation model and benchmark for multi-species genome. *arXiv*

preprint arXiv:2306.15006.

Declarations

Funding

This research was supported by the French National Research Agency (ANR) under grant ANR-23-CE45-0022 (GenSynBio) and the Spanish Ministry of Science under grant PID2023-152184OB (AI-SynBio). Gene synthesis for experimental validation was supported by a Twist Bioscience Academic Research Grant. Funders had no role in study design, data collection, analysis, or manuscript preparation.

Conflict of Interest

The authors declare no conflicts of interest.

Data Availability Statement

All generated protein and regulatory element sequences, experimental validation results, and TVAE model code are available at <https://github.com/popescu-dubois/genai-synbio> under MIT licence. Experimental data is deposited in Zenodo under DOI: 10.5281/zenodo.19614805. TVAE model weights available on Hugging Face Hub (popescu-dubois/tvae-metabolic). Released under CC BY 4.0.

Ethical Approval

All synthetic biology experiments used non-pathogenic laboratory strains of *E. coli* (K-12 derivatives) and HEK293T cells under standard biosafety level 1 conditions. Experimental protocols were approved by Western Europe Data Science University Biosafety Committee (WEDS-BSC-2024-0142). No human tissue or patient-derived samples were used.

Appendix A

Generative Model Selection Framework for Synthetic Biology

Use this framework to select the appropriate generative model for your synthetic biology design task, based on design objective, available structural information, and experimental throughput.