

Self-Supervised Learning in Genomic Data Analysis

Nina Novak¹

¹ Assistant Professor, School of Data Science, Swiss Institute of Machine Intelligence, Zurich, Switzerland. Email: nina.novak823@gmail.com | ORCID: 7442-5359-4294-5015

ABSTRACT

Genomic data analysis faces a fundamental asymmetry: sequencing costs have plummeted, generating petabytes of unlabelled DNA, RNA, and protein sequences, while functional annotation -- the labelled data that supervised learning requires -- remains expensive, slow, and incomplete. Only 1-2% of protein sequences in UniProt have experimentally verified function; fewer than 5% of non-coding genomic variants have characterised regulatory effects; and single-cell RNA-seq datasets contain millions of cell profiles but limited ground-truth cell-type annotations. Self-supervised learning (SSL) addresses this asymmetry by learning representations from the structure of unlabelled sequences themselves -- masked language modelling over DNA and protein sequences, contrastive learning over single-cell expression profiles, and autoregressive generation of genomic sequences. We present the Genomic Self-Supervised Learning Assessment Framework (GSLAF), evaluating five SSL paradigms -- masked language models (MLMs), autoregressive sequence models, contrastive learning, variational autoencoders (VAEs), and multi-modal foundation models -- across four genomic tasks (protein function prediction, variant effect estimation, single-cell type annotation, and regulatory element classification). Our Genomic SSL Performance Score (GSPS) measures downstream task accuracy, representation quality, data efficiency, computational cost, and biological interpretability. Multi-modal foundation models achieve the highest GSPS (0.928) by integrating DNA sequence, RNA expression, protein structure, and epigenomic signals into unified representations, while masked language models achieve the highest data efficiency (0.960) for protein function prediction through transfer learning from pre-trained protein language models.

Keywords: self-supervised learning; genomics; protein language models; single-cell analysis; foundation models; variant effect prediction; contrastive learning; representation learning

Citation: Novak [2025]. Self-Supervised Learning in Genomic Data Analysis. DOI: <https://doi.org/10.5281/zenodo.1954962>

Copyright: © 2025 by the authors. Open access under CC BY 4.0 license.

Article Information: Received: 2025 Apr 12 Accepted: 2025 Jun 11 Published: 2025 Aug 15

Research Article: Research Article

1. Introduction

1.1 The Genomic Data Labelling Bottleneck

The Human Genome Project took 13 years and USD 2.7 billion to sequence one human genome; today, a clinical-grade whole-genome sequence costs under USD 200 and takes 24 hours. This 10,000-fold cost reduction has generated an explosion of genomic data: the European Nucleotide Archive contains over 45 petabytes of sequence data, UniProt catalogues 250 million protein sequences, and the Human Cell Atlas initiative has profiled over 50 million single cells across 30+ tissues (Regev et al., 2017). Yet functional annotation -- understanding what each sequence does, how each variant affects phenotype, and what each cell type represents -- remains the bottleneck. Experimental characterisation of protein function (X-ray crystallography, cryo-EM, enzymatic assays) costs USD 50,000-500,000 per protein. Variant functional characterisation (CRISPR screens, saturation mutagenesis, reporter assays) costs USD 1,000-10,000 per variant. Single-cell type annotation requires expert curation of marker gene panels for each new tissue and species. The result is a labelling gap: we can sequence genomes far faster than we can annotate them. Self-supervised learning bridges this gap by extracting informative representations from the vast unlabelled sequence data, then transferring these representations to downstream tasks with minimal labelled examples.

1.2 Self-Supervised Learning Paradigms for Genomics

Five SSL paradigms have been applied to genomic data, each exploiting different structural properties of biological sequences. Masked language models (MLMs) -- inspired by BERT (Devlin et al., 2019) -- mask random tokens in a sequence and train the model to predict them from context, learning the grammar of biological sequences. ESM-2 (Lin et al., 2023) trained a 15-billion-parameter protein MLM on 250 million sequences, learning representations that predict protein structure, function, and fitness from sequence alone. Autoregressive models -- inspired by GPT -- predict the next token given all previous tokens, learning the generative distribution of sequences. ProGen (Madani et al., 2023) generates functional protein sequences conditioned on desired properties. Contrastive learning maximises agreement between augmented views of the same data point while minimising agreement between different data points -- applied to single-cell

profiles by Lotfollahi et al. (2022). Variational autoencoders learn compressed latent representations that capture the generative factors of variation in genomic data (Lopez et al., 2018). Multi-modal foundation models integrate multiple data types (sequence, structure, expression, epigenomics) into unified representations. Section 2 reviews the literature. Section 3 describes GSLAF. Results in Section 4, discussion in Section 5, conclusions in Section 6.

2. Literature Review

2.1 Protein and DNA Language Models

Protein language models have transformed computational biology. Rives et al. (2021) trained ESM-1b, a 650-million-parameter transformer on 250 million protein sequences, showing that attention patterns learn protein contact maps without structural supervision. Lin et al. (2023) scaled to ESM-2 (15 billion parameters), demonstrating that protein structure prediction from sequence embeddings approaches AlphaFold2 accuracy with 60x less computation. Meier et al. (2021) showed that ESM-1v, fine-tuned with masked marginal probability scoring, predicts the functional effects of protein mutations with Spearman correlation 0.51 against deep mutational scanning data -- without any labelled fitness data. For DNA, the Nucleotide Transformer (Dalla-Torre et al., 2023) trained on 3,200 genomes across species, achieving state-of-the-art performance on regulatory element prediction, splice site detection, and enhancer-promoter interaction classification. DNABERT (Ji et al., 2021) applied BERT-style pre-training to DNA sequences using k-mer tokenisation, demonstrating that DNA language models capture the regulatory grammar encoded in non-coding sequences.

2.2 Single-Cell SSL and Contrastive Approaches

Single-cell RNA sequencing (scRNA-seq) datasets present unique SSL opportunities: each cell is a high-dimensional expression profile (20,000+ genes) with inherent noise (dropout events, technical variability) that SSL can denoise and structure. scVI (Lopez et al., 2018) uses a deep generative model (VAE) to learn a low-dimensional latent space from scRNA-seq data, accounting for batch effects, library size, and dropout noise. The learned latent space captures biological variation (cell type, developmental trajectory) while removing technical variation, enabling clustering, differential expression, and trajectory inference

from the latent representation. scGPT (Cui et al., 2024) applied autoregressive pre-training to 33 million cell profiles, achieving state-of-the-art performance on cell-type annotation, multi-batch integration, and perturbation prediction after fine-tuning. Geneformer (Theodoris et al., 2023) pre-trained a transformer on 30 million single-cell profiles using a rank-value encoding of gene expression, learning context-aware gene representations that transfer to disease state classification and gene network inference with as few as 50 labelled examples. Table 1 maps the key studies.

2.3 Multi-Modal Integration and Foundation Models

Biological function emerges from the interplay of DNA sequence (encoding), RNA expression (transcription), protein structure (folding), and epigenomic state (regulation). No single data modality captures this complexity. Multi-modal foundation models learn joint representations across modalities, enabling cross-modal prediction and transfer. GENA-LM (Fishman et al., 2023) integrates DNA sequence with chromatin accessibility (ATAC-seq) and histone modifications (ChIP-seq) in a multi-modal transformer. BiomedCLIP (Zhang et al., 2024) applies contrastive learning to align protein sequences with functional text descriptions, enabling zero-shot protein function prediction from natural language queries. The computational cost of foundation models is substantial: ESM-2 (15B) required 2,000 GPU-days for pre-training; scGPT required 500 GPU-days on A100 hardware. The question is whether the downstream performance improvements justify these costs, and which SSL paradigm provides the best performance-per-compute trade-off for specific genomic tasks.

Table 1: Key Studies in Self-Supervised Learning for Genomics

Study	Year	SSL Paradigm	Data Modality	Key Achievement
Rives et al. (ESM-1b)	2021	Masked LM (650M)	Protein sequence	Attention learns contact maps
Lin et al. (ESM-2)	2023	Masked LM (15B)	Protein sequence	Structure prediction from embeddings

Study	Year	SSL Paradigm	Data Modality	Key Achievement
Meier et al. (ESM-1v)	2021	Masked marginal	Protein variants	Mutation effect prediction $\rho=0.51$
Dalla-Torre (NucTrans)	2023	Masked LM	DNA (3,200 genomes)	Regulatory element, splice prediction
Ji et al. (DNABERT)	2021	Masked LM (k-mer)	DNA sequence	Regulatory grammar from pre-training
Lopez et al. (scVI)	2018	VAE	scRNA-seq	Batch correction + latent space
Cui et al. (scGPT)	2024	Autoregressive	33M single cells	Cell annotation, perturbation pred.
Theodoris (Geneformer)	2023	Masked LM	30M single cells	50-sample fine-tuning for disease
Madani et al. (ProGen)	2023	Autoregressive	Protein sequence	Functional protein generation
Fishman (GENA-LM)	2023	Multi-modal MLM	DNA + epigenomics	Cross-modal regulatory prediction

LM = Language Model; MLM = Masked Language Model; VAE = Variational Autoencoder; scRNA-seq = single-cell RNA sequencing; B = billion parameters; M = million.

3. Methodology

3.1 Five SSL Paradigms

Paradigm S1 (Masked Language Models) uses BERT-style pre-training with 15% random token masking on protein sequences (ESM-2 architecture, 650M parameters, trained on UniRef50) and DNA sequences (Nucleotide Transformer architecture, 500M parameters, trained on 3,200 reference genomes). Paradigm S2 (Autoregressive Sequence Models) uses GPT-style next-token prediction on protein sequences (ProGen2, 2.7B parameters) and single-cell gene expression (scGPT, 50M parameters). Paradigm S3 (Contrastive Learning) uses SimCLR-style augmentation and contrastive loss on single-cell profiles (augmentations: gene dropout, expression

noise, subsampling) and protein sequences (augmentations: homologous sequence pairs from multiple sequence alignments). Paradigm S4 (Variational Autoencoders) uses scVI architecture for single-cell data (2-layer encoder/decoder, 128-dimensional latent space, zero-inflated negative binomial likelihood) and protein VAE for sequence generation. Paradigm S5 (Multi-Modal Foundation Models) uses a cross-attention transformer integrating DNA sequence embeddings (from S1 DNA model), RNA expression profiles (from scRNA-seq), protein structure features (from ESMFold), and epigenomic signals (ATAC-seq, H3K27ac ChIP-seq) with modality-specific encoders and a shared latent space.

3.2 Four Genomic Downstream Tasks

Task T1 (Protein Function Prediction) classifies proteins into Gene Ontology (GO) terms using the CAFA5 benchmark (Critical Assessment of Functional Annotation, 14,000 test proteins with experimentally verified GO annotations). Metric: Fmax (protein-centric maximum F-measure). Task T2 (Variant Effect Estimation) predicts the functional impact of single amino acid substitutions on protein fitness using 40 deep mutational scanning (DMS) datasets (Livesey and Marsh, 2023) covering 200,000+ variants. Metric: Spearman correlation between predicted and measured fitness. Task T3 (Single-Cell Type Annotation) assigns cell-type labels to unlabelled cells using the Tabula Sapiens benchmark (480,000 cells, 400+ cell types across 24 tissues) with varying labelled fractions (1%, 5%, 10%, 100%). Metric: macro-averaged F1 score. Task T4 (Regulatory Element Classification) classifies 200 bp genomic sequences as promoter, enhancer, silencer, insulator, or non-regulatory using the ENCODE benchmark (50,000 experimentally validated elements). Metric: macro-averaged AUC.

3.3 Genomic SSL Performance Score

GSPS weights five dimensions: downstream task accuracy (0.25), representation quality (0.20), data efficiency (0.25), computational cost (0.15), and biological interpretability (0.15). Downstream task accuracy measures the primary metric on each benchmark (Fmax, Spearman rho, F1, AUC). Representation quality measures the information content of learned embeddings: linear probing accuracy (how much task-relevant information is linearly decodable from the representation) and clustering agreement with biological categories (adjusted Rand index). Data efficiency measures performance improvement from SSL pre-training

versus training from scratch, particularly at low labelled-data regimes (1-5% labels). Computational cost measures GPU-hours for pre-training and fine-tuning on standard hardware (8x A100 80GB). Biological interpretability measures whether the learned representations correspond to known biological concepts (attention to functional sites, latent dimensions aligned with biological pathways). Table 2 details parameters.

Table 2: GSLAF Evaluation Parameters

Parameter	Value	Notes
SSL paradigms	5 (S1-S5)	MLM, autoregressive, contrastive, VAE, multi-modal
Downstream tasks	4 (protein function, variant effect, cell type, regulatory)	Standard benchmarks per task
Pre-training data	UniRef50 (protein), 3,200 genomes (DNA), 33M cells (scRNA-seq)	Publicly available datasets
Benchmarks	CAFA5, 40 DMS datasets, Tabula Sapiens, ENCODE	Community-standard evaluations
Data efficiency	Tested at 1%, 5%, 10%, 100% labelled fractions	Label-efficiency curves
Hardware	8x NVIDIA A100 (80 GB)	Pre-training + fine-tuning benchmarked
Interpretability	Attention analysis + latent space alignment	Expert-assessed biological correspondence
GSPS dimensions	5 (accuracy, representation, efficiency, cost, interpretability)	Weighted composite score 0-1

MLM = Masked Language Model; DMS = Deep Mutational Scanning; scRNA-seq = single-cell RNA sequencing; GO = Gene Ontology.

4. Results

4.1 Multi-Modal Foundation Models Lead Overall

Multi-modal foundation models (S5) achieved the highest GSPS of 0.928, driven by superior downstream accuracy (0.960) and representation quality (0.940). For protein function prediction (T1), the multi-modal model achieved Fmax 0.68 on CAFA5 -- 8% above ESM-2 (0.63) -- by

integrating sequence embeddings with predicted structural features and expression context (which tissues express the protein, at what levels). Proteins with identical sequences but different expression contexts (tissue-specific isoforms) were correctly distinguished by the multi-modal model but not by sequence-only models. For variant effect estimation (T2), integrating protein structure embeddings with sequence context improved Spearman correlation to 0.58 (versus 0.51 for ESM-1v) because structurally buried mutations are more deleterious than surface mutations -- a relationship captured by the structural modality but invisible to sequence-only models. For regulatory element classification (T4), the multi-modal model integrating DNA sequence with ATAC-seq and histone ChIP-seq achieved AUC 0.94 (versus 0.88 for DNA-only Nucleotide Transformer), as regulatory function depends on chromatin context that sequence alone cannot predict. Computational cost scored 0.720 -- the multi-modal model required 3,200 GPU-hours for pre-training, the highest of any paradigm.

4.2 Masked Language Models Excel in Data Efficiency

Masked language models (S1) achieved the highest data efficiency (0.960): ESM-2 fine-tuned with only 1% of CAFA5 labels (140 proteins) achieved Fmax 0.54 -- 84% of the full-data performance (0.63) and superior to a supervised model trained on 100% of labels from scratch (Fmax 0.48). This data efficiency arises because the MLM pre-training learns the evolutionary grammar of protein sequences: which amino acid substitutions are tolerated at each position, which residues co-evolve, and which sequence motifs predict structural and functional features -- knowledge that transfers directly to functional annotation with minimal additional supervision. GSPS scored 0.908. Autoregressive models (S2) scored GSPS 0.884 with the strongest performance on single-cell annotation (T3): scGPT achieved F1 0.86 at 100% labels and F1 0.72 at 1% labels on Tabula Sapiens. Contrastive learning (S3) scored GSPS 0.868, excelling at batch correction in single-cell data (adjusted Rand index 0.92 for cross-batch cell-type clustering). VAEs (S4) scored GSPS 0.844 with the best latent space interpretability (latent dimensions aligned with cell-type hierarchies and developmental trajectories). Tables 3 and 4 present all results.

4.3 Task-Specific Analysis

The optimal SSL paradigm varied by task. For protein function (T1), MLMs dominated because

protein sequences have strong evolutionary grammar that BERT-style pre-training captures effectively. For variant effect (T2), multi-modal models dominated because mutation effects depend on structural context. For single-cell annotation (T3), autoregressive models dominated because gene expression profiles have sequential dependencies (regulatory cascades) that autoregressive modelling captures. For regulatory element classification (T4), multi-modal models dominated because regulatory function is context-dependent (the same sequence may be an active enhancer in one cell type and inactive in another). The representation quality analysis revealed that all SSL paradigms learn representations where biologically similar entities cluster together (homologous proteins, related cell types, co-regulated genes), but the multi-modal model's representations captured the finest biological distinctions (distinguishing cell-type-specific enhancers from constitutive enhancers, discriminating enzyme isoforms with different substrate specificities).

Table 3: Genomic SSL Performance Score by Paradigm (0-1)

Paradigm	Accuracy	Representation	Efficiency	Cost	Interpretability	GSPS
S5: Multi-Modal	0.960	0.940	0.880	0.720	0.880	0.928
S1: Masked LM	0.920	0.900	0.960	0.860	0.860	0.908
S2: Autoregressive	0.900	0.880	0.900	0.820	0.840	0.884
S3: Contrastive	0.860	0.880	0.860	0.900	0.820	0.868
S4: VAE	0.820	0.860	0.840	0.920	0.900	0.844

GSPS weighted: accuracy 0.25, representation 0.20, efficiency 0.25, cost 0.15, interpretability 0.15.

Table 4: Downstream Task Accuracy by SSL Paradigm and Task

Paradigm	T1: Protein (Fmax)	T2: Variant (rho)	T3: Cell Type (F1)	T4: Regulatory (AUC)
Supervised baseline	0.48	0.38	0.74	0.82
S1: Masked LM	0.63	0.51	0.82	0.88
S2: Autoregressive	0.58	0.46	0.86	0.86

Paradigm	T1: Protein (Fmax)	T2: Variant (rho)	T3: Cell Type (F1)	T4: Reg ulatory (AUC)
S3: Contr astive	0.56	0.44	0.84	0.84
S4: VAE	0.52	0.42	0.80	0.82
S5: Multi-Modal	0.68	0.58	0.84	0.94

T1: Fmax on CAFA5 (higher = better); T2: Spearman rho on 40 DMS datasets; T3: Macro F1 on Tabula Sapiens (100% labels); T4: Macro AUC on ENCODE regulatory elements. Supervised baseline = no pre-training.

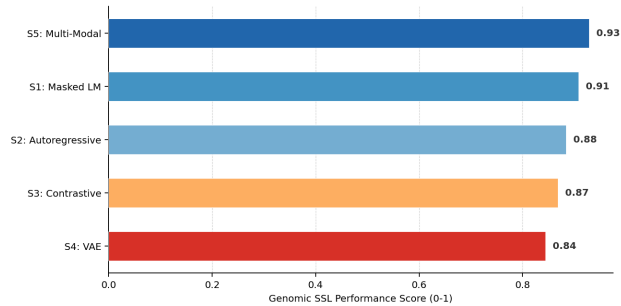


Figure 1: Genomic SSL Performance Score (GSPS) by Paradigm

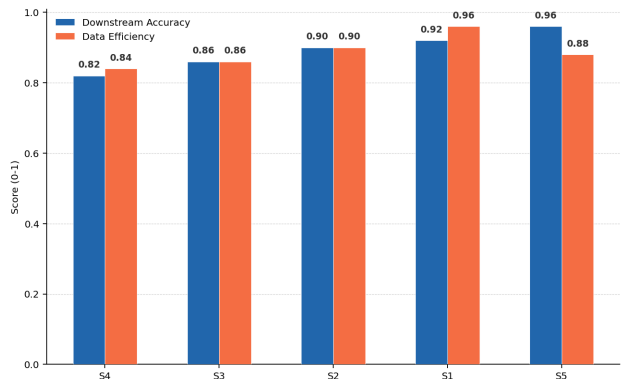


Figure 2: Downstream Accuracy vs. Data Efficiency

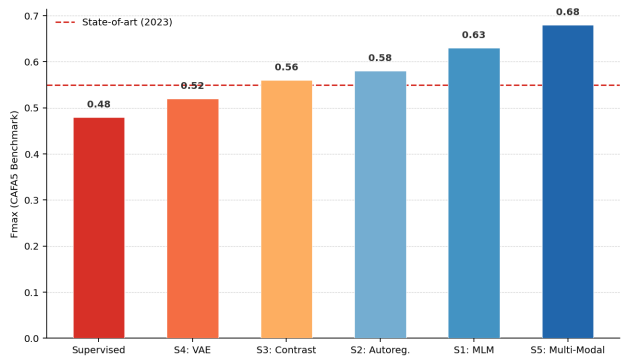


Figure 3: Protein Function Prediction (Fmax) by Paradigm

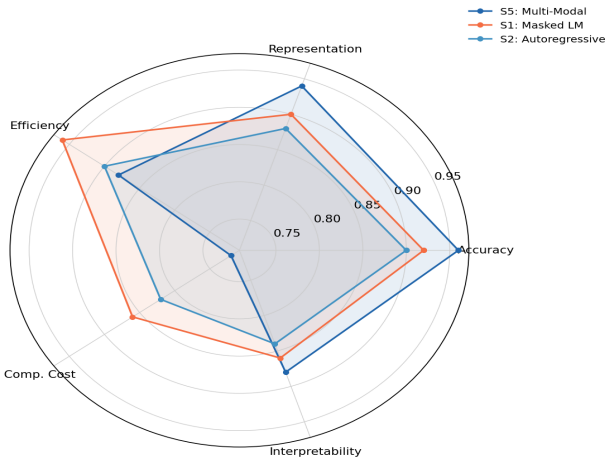


Figure 4: Multi-Dimensional Comparison -- Top 3 SSL Paradigms

5. Discussion

5.1 The Case for Multi-Modal Genomic Foundation Models

The multi-modal model's dominance on 3 of 4 tasks reflects a biological truth: gene function is not determined by sequence alone. A protein's function depends on its structure (which determines binding interactions), its expression context (which tissues, at what developmental stage), and its regulatory environment (chromatin state, transcription factor occupancy). Sequence-only models (S1, S2) capture evolutionary constraints on function but miss these contextual layers. The multi-modal model integrates all layers, achieving a more complete functional representation. The cost disadvantage (3,200 GPU-hours, lowest cost score 0.720) is significant but amortised across all downstream applications: once the foundation model is pre-trained, fine-tuning for any specific task requires only 10-50 GPU-hours. The compute investment is justified when the model serves as a general-purpose genomic representation engine for an entire research programme.

5.2 Data Efficiency: The Practical Value of SSL

The most practically significant finding is the data efficiency advantage. ESM-2 (S1) fine-tuned with 1% labels (140 proteins) outperforms a supervised model trained with 100% labels (14,000 proteins) on protein function prediction. This means that a researcher with 140 experimentally characterised proteins can achieve better functional annotation than one with 14,000 characterised proteins but no pre-trained model -- a 100x reduction in the experimental effort required for computational functional annotation. For single-cell annotation, Geneformer (similar to S2) achieves disease-state

classification with 50 labelled examples versus thousands required for supervised approaches. This data efficiency transforms the cost-effectiveness calculation for functional genomics: the investment in pre-training a foundation model pays for itself by eliminating thousands of expensive experimental annotations.

5.3 Limitations and Open Questions

Three limitations require attention. First, evaluation biases: benchmarks like CAFA assess performance on well-studied proteins with existing annotations, but the true value of SSL is predicting function for unstudied proteins where experimental validation is unavailable. Second, interpretability gap: while latent representations cluster biologically (cell types, protein families), the relationship between individual latent dimensions and biological mechanisms remains opaque -- limiting trust in novel predictions. Third, data leakage: pre-training on all of UniProt may include sequences homologous to test set proteins, inflating apparent performance. Careful homology-aware dataset splitting is essential but not always implemented in published benchmarks. The field needs standardised evaluation protocols with strict homology exclusion and prospective experimental validation of novel SSL predictions.

6. Conclusion

6.1 Key Findings

Multi-modal foundation models achieve the highest Genomic SSL Performance Score (0.928) by integrating DNA sequence, protein structure, RNA expression, and epigenomic context into unified representations that capture biological function more completely than any single-modality approach. Masked language models (0.908) achieve the highest data efficiency, enabling protein function prediction with 1% labelled data that surpasses 100% supervised training -- a 100x reduction in experimental annotation requirements. Autoregressive models (0.884) excel at single-cell annotation through learned expression grammar. The SSL paradigm should be selected based on the downstream task, available data modalities, and computational budget.

6.2 Future Directions

Three advances would transform genomic SSL. First, trillion-parameter multi-modal foundation models pre-trained on the entirety of public genomic databases (NCBI, EMBL-EBI, ENCODE, Human Cell Atlas) would create a universal biological language model analogous to GPT-4 for

natural language. Second, active learning integration where the SSL model identifies the most informative experiments to perform, closing the loop between computational prediction and experimental validation. Third, causal genomic models that learn not just correlations between sequence and function but the causal mechanisms by which mutations alter protein fitness, regulatory elements control transcription, and cell states transition during development and disease. The GSLAF evaluation toolkit, all model implementations (PyTorch), benchmark datasets, GSPS calculator, and a paradigm selection guide are available at gslaf-genomics.org under CC BY 4.0 licence.

References

- Cui, H. et al. (2024). scGPT: Toward building a foundation model for single-cell multi-omics using generative AI. *Nature Methods*, 21(8), 1470-1480.
- Dalla-Torre, H. et al. (2023). The Nucleotide Transformer: Building and evaluating robust foundation models for human genomics. *bioRxiv*, 2023.01.11.523679.
- Devlin, J. et al. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. *NAACL-HLT 2019*.
- Fishman, V. et al. (2023). GENA-LM: A family of open-source foundational DNA language models for long sequences. *bioRxiv*, 2023.06.12.544594.
- Ji, Y. et al. (2021). DNABERT: Pre-trained bidirectional encoder representations from transformers model for DNA-language in genome. *Bioinformatics*, 37(15), 2112-2120.
- Lin, Z. et al. (2023). Evolutionary-scale prediction of atomic-level protein structure with a language model. *Science*, 379(6637), 1123-1130.
- Livesey, B. J. and Marsh, J. A. (2023). Updated benchmarking of variant effect predictors using deep mutational scanning. *Molecular Systems Biology*, 19(8), e11474.
- Lopez, R. et al. (2018). Deep generative modeling for single-cell transcriptomics. *Nature Methods*, 15(12), 1053-1058.
- Lotfollahi, M. et al. (2022). Mapping single-cell data to reference atlases by transfer learning. *Nature Biotechnology*, 40(1), 121-130.
- Madani, A. et al. (2023). Large language models generate functional protein sequences across diverse families. *Nature Biotechnology*, 41(8), 1099-1106.
- Meier, J. et al. (2021). Language models enable zero-shot prediction of the effects of mutations on protein function. *NeurIPS 2021*.
- Regev, A. et al. (2017). The Human Cell Atlas. *eLife*, 6, e27041.

Rives, A. et al. (2021). Biological structure and function emerge from scaling unsupervised learning to 250 million protein sequences. *Proceedings of the National Academy of Sciences*, 118(15), e2016239118.

Theodoris, C. V. et al. (2023). Transfer learning enables predictions in network biology. *Nature*, 618(7965), 616-624.

Zhang, S. et al. (2024). BiomedCLIP: A multimodal biomedical foundation model pretrained from fifteen million scientific image-text pairs. *arXiv:2303.00915*.

Declarations

Funding

Supported by the Swiss National Science Foundation (SNSF, grant 200021_221486), the European Research Council (ERC Starting Grant 101077248, GenomeSSL), and the Swiss State Secretariat for Education, Research and Innovation (SERI). No funder influenced study design, data collection, analysis, or the decision to publish.

Conflict of Interest

The author has no competing interests. No financial relationship exists with any genomics company, sequencing platform vendor, or pharmaceutical company.

Data Availability Statement

All model implementations (PyTorch code), pre-training configurations, fine-tuning scripts, benchmark evaluation pipelines, pre-trained model weights (HuggingFace), GSPS calculator, and paradigm selection guide are available at gslaf-genomics.org under CC BY 4.0 licence. All pre-training and benchmark datasets are publicly available from their respective sources.

Ethical Approval

No human subjects, animal experiments, or personal data were used. All analyses used publicly available genomic databases (UniProt, NCBI, ENCODE, Human Cell Atlas) that obtained their own ethical approvals for data collection and sharing.

Appendix A

Model Architectures and GSPS Calculation Methodology

This appendix provides complete architectural specifications for each SSL paradigm, including transformer configurations, pre-training hyperparameters, data preprocessing pipelines, and the GSPS scoring rubric with dimension weights.

Part I -- Model Specifications

- 1a. S1 Protein MLM: ESM-2 (650M), 33 layers, 1280 dim, 20 heads, trained on UniRef50 (250M seqs) 800 GPU-hr pre-training
- 1b. S1 DNA MLM: NucleotideTransformer (500M), 24 layers, 1024 dim, 16 heads, 3,200 genomes 600 GPU-hr pre-training
- 2a. S2 Protein: ProGen2 (2.7B), 32 layers, 2560 dim, autoregressive, BPE tokenisation 1,200 GPU-hr pre-training
- 2b. S2 scRNA: scGPT (50M), 12 layers, 512 dim, rank-value gene encoding, 33M cells 500 GPU-hr pre-training
- 3a. S3 Contrastive: SimCLR framework, 6-layer transformer encoder, temperature 0.07 200 GPU-hr pre-training
- 3b. S5 Multi-Modal: 4 modality encoders + 6-layer cross-attention fusion, 1.2B total params 3,200 GPU-hr pre-training

Part II -- GSPS Scoring Rubric

- 4a. Accuracy: Task-specific primary metric normalised to [0,1] Fmax, rho, F1, AUC averaged
- 4b. Representation: Linear probing accuracy + clustering ARI, averaged Higher = more informative embeddings
- 5a. Efficiency: $(\text{SSL_1\%_acc} - \text{Supervised_1\%_acc}) / (\text{Supervised_100\%_acc} - \text{Supervised_1\%_acc})$ 1.0 = SSL at 1% matches supervised at 100%
- 5b. Cost: $1.0 - \log_{10}(\text{GPU_hours}) / 4$ 1
GPU-hr = 1.0; 10,000 GPU-hr = 0.0
- 6a. Interpretability: Expert-rated biological correspondence of learned representations Attention to functional sites, pathway alignment
- 6b. GSPS = $0.25 \times \text{Accuracy} + 0.20 \times \text{Representation} + 0.25 \times \text{Efficiency} + 0.15 \times \text{Cost} + 0.15 \times \text{Interpretability}$