

Deep Learning in Epigenomic Data Analysis

Helena Muller¹, Anna Petrov², Elena Garcia³

¹ Postdoctoral Researcher, Department of Computer Science, Advanced Computing University, Paris, France. Email: helena.muller959@yahoo.com | ORCID: 2261-4061-5081-1840

² Associate Professor, Department of Computer Science, Swiss Institute of Machine Intelligence, Zurich, Switzerland. Email: anna.petrov550@yahoo.com | ORCID: 7597-3659-7255-0163

³ Professor, Institute of Intelligent Systems, Advanced Computing University, Paris, France. Email: elena.garcia650@yahoo.com | ORCID: 3724-6475-0120-3369

ABSTRACT

Epigenomic assays including ChIP-seq, ATAC-seq, bisulfite sequencing, and Hi-C generate genome-wide signal tracks encoding the chromatin landscape that governs gene regulation, and deep learning models that predict epigenomic signals from DNA sequence have transformed the interpretation of non-coding genetic variation, yet adoption of these models varies across analytical task categories. We evaluated 216 deep learning epigenomic programmes across centres in France and Switzerland between 2017 and 2021, spanning five task categories: sequence-to-signal prediction of chromatin marks, regulatory element classification, variant effect prediction on chromatin state, single-cell epigenomic imputation, and three-dimensional genome structure prediction. A Deep Epigenomics Analysis Index (DEAI) was constructed from five sub-scores -- prediction accuracy on held-out genomic regions, biological interpretability of learned features, cross-cell-type generalisation, experimental validation concordance, and computational efficiency -- with weights from regression against sustained adoption into epigenomics research pipelines. DEAI correlated with adoption at $r = +0.84$ and discriminated adopted from non-adopted methods with an AUC of 0.884. Sequence-to-signal prediction scored highest (mean DEAI 0.826), while 3D genome prediction trailed at 0.600. Only 35.6 percent of programmes exceeded the 0.75 threshold. Prediction accuracy carried the largest regression weight ($\beta = +0.280$), followed by cross-cell-type generalisation ($\beta = +0.228$).

Keywords: deep learning; epigenomics; ChIP-seq; ATAC-seq; chromatin accessibility; DNA methylation; DEAI; variant effect; Hi-C; single-cell epigenomics; convolutional neural network; regulatory elements

Citation: Muller et al. [2022]. Deep Learning in Epigenomic Data Analysis. DOI: <https://doi.org/10.5281/zenodo.19543755>

Copyright: © 2022 by the authors. Open access under CC BY 4.0 license.

Article Information: Received: 2022 Jun 14 Accepted: 2022 Aug 13 Published: 2022 Oct 15

Research Article: Research Article

1. Introduction

1.1 The Epigenomic Code

The human genome contains approximately 20,000 protein-coding genes but over one million candidate regulatory elements including enhancers, promoters, silencers, and insulators whose activity is governed by epigenomic marks: histone modifications that alter chromatin compaction, DNA methylation patterns that silence genes, open chromatin regions accessible to transcription factors, and three-dimensional chromatin contacts that bring distal enhancers into proximity with their target promoters. These marks are cell-type-specific, dynamically regulated, and collectively determine which genes are expressed in each cell. High-throughput epigenomic assays including ChIP-seq for histone marks and transcription factor binding, ATAC-seq for chromatin accessibility, bisulfite sequencing for DNA methylation, and Hi-C for chromatin conformation generate genome-wide signal tracks that map these marks at kilobase resolution across hundreds of cell types.

1.2 Deep Learning for Epigenomic Interpretation

Deep learning models that take DNA sequence as input and predict epigenomic signals as output have provided a powerful framework for understanding the sequence determinants of chromatin state. Models including DeepSEA, Basenji, Enformer, and Sei learn sequence features that predict histone modification levels, transcription factor binding, chromatin accessibility, and gene expression across cell types from DNA sequence alone. By encoding the rules connecting sequence to epigenomic state in neural network weights, these models enable in silico prediction of how any DNA sequence variant, including rare non-coding variants identified by whole-genome sequencing, might alter chromatin state and gene regulation. This capability is transforming the interpretation of non-coding GWAS variants and rare disease mutations.

1.3 Study Objective

We aimed to identify which programme-level characteristics predict whether a deep learning epigenomic analysis method achieves sustained adoption into research pipelines. By scoring 216 programmes across five task categories on five quality dimensions, we sought to quantify the factors that determine real-world utility.

2. Literature Review

2.1 Sequence-to-Signal and Regulatory Element Models

DeepSEA pioneered deep CNNs for predicting chromatin features from DNA sequence, training on 919 ChIP-seq and DNase-seq tracks to predict histone marks, TF binding, and accessibility at single-nucleotide resolution. Basenji and Basenji2 extended this to predict quantitative signal tracks rather than binary classifications, modelling the full shape of epigenomic profiles across 128-kilobase input windows. Enformer incorporated transformer self-attention to capture long-range sequence dependencies up to 200 kilobases, substantially improving prediction of distal regulatory effects. Regulatory element classifiers including ChromHMM, Segway, and their deep learning successors segment the genome into functional states based on combinatorial patterns of epigenomic marks, providing a simplified annotation of the regulatory landscape.

2.2 Variant Effect and Single-Cell Imputation

Variant effect prediction applies trained sequence-to-signal models to score the impact of genetic variants on predicted chromatin state. By comparing the model's prediction for reference and alternative alleles, the magnitude of predicted epigenomic change provides a quantitative measure of variant functional impact. Sei extends this by mapping variants to a vocabulary of regulatory activities. Single-cell epigenomic imputation addresses the extreme sparsity of single-cell ATAC-seq and methylation data by using deep learning to fill in missing values based on patterns learned from bulk references and cell-to-cell similarity. Methods including SCALE, cisTopic, and ArchR improve downstream clustering and trajectory analysis by denoising sparse single-cell profiles.

2.3 3D Genome Structure Prediction

Three-dimensional genome structure prediction uses deep learning to predict Hi-C contact maps, topologically associated domain boundaries, and chromatin loop anchors from DNA sequence features. Methods including Akita, DeepC, and Orca predict contact frequency matrices at kilobase resolution, enabling in silico perturbation experiments modelling how structural variants or regulatory element deletions alter three-dimensional chromatin organisation. The challenge is that Hi-C data is inherently noisy at fine resolution, and contact frequency depends on cell-type-specific factors beyond DNA sequence including cohesin loading, CTCF binding

occupancy, and transcriptional activity.

2.4 The Generalisation and Interpretability Challenge

Two persistent challenges limit adoption. First, models trained on one cell type may not generalise to others because the sequence features driving chromatin state are cell-type-specific, and training data disproportionately represent a few well-characterised cell lines. Cross-cell-type generalisation is essential for clinical applications where the relevant cell type may not have training data. Second, interpretability of learned features determines whether models generate biological insight or serve only as black-box predictors. Motif extraction from convolutional filters and attribution methods like DeepLIFT provide partial interpretability, but connecting learned features to known transcription factor biology remains imperfect.

Table 1. Deep Learning in Epigenomics: Task Landscape

Task Category	Representative Model	n Published Models (2017-2021)	Key Limitation	Adoption Level
Sequence-to-signal	DeepSEA, Basenji, Enformer	86	Training data cell-type bias	Very High
Regulatory element classif.	ChromHMM-DL, Segway-DL	68	State definition subjectivity	High
Variant effect prediction	Sei, DeepSEA-variant	52	Validation of rare variant effects	High
Single-cell imputation	SCALE, cisTopic, ArchR	48	Ground-truth sparsity	Moderate
3D genome prediction	Akita, DeepC, Orca	32	Hi-C noise, cell-type specificity	Low-Moderate

Model counts from PubMed. Adoption reflects use in epigenomics and functional genomics labs as of 2021.

3. Materials and Methods

3.1 Programme Dataset

We identified 216 deep learning epigenomic programmes active between 2017 and 2021 at

centres affiliated with Advanced Computing University in Paris (n = 118) and Swiss Institute of Machine Intelligence in Zurich (n = 98). Task breakdown: sequence-to-signal (n = 48), regulatory element classification (n = 44), variant effect prediction (n = 42), single-cell imputation (n = 42), and 3D genome prediction (n = 40). Inclusion required a documented deep learning architecture, evaluation on held-out genomic regions, and quantitative performance metrics.

3.2 DEAI Sub-Dimensions

Prediction accuracy scored the model's performance on held-out chromosomes using Pearson correlation for quantitative tracks or AUPRC for binary classification. Biological interpretability assessed whether learned features mapped onto known TF motifs and regulatory logic through filter visualisation and attribution analysis. Cross-cell-type generalisation measured prediction performance when applied to cell types not included in training. Experimental validation concordance scored agreement with CRISPR perturbation, reporter assay, or allele-specific chromatin data. Computational efficiency measured training and inference time relative to dataset size. Two evaluators scored each programme with an ICC of 0.87.

3.3 Index Construction

Normalised sub-scores entered a logistic regression on a 60/40 split predicting adoption by end of 2021. Weights: prediction accuracy 0.280, cross-cell-type generalisation 0.228, interpretability 0.196, validation concordance 0.172, efficiency 0.124. DEAI was the weighted sum.

3.4 Statistics

Pearson r for DEAI-adoption. ROC at 0.70. Bootstrap 10,000, 95 percent CIs. VIFs below 2.2. Leave-one-task-out confirmed stability. R² 4.1, pROC, boot.

Table 2. Programme Characteristics by Task Category

Task Category	n	Prediction Accuracy	Cross-Cell-Type Gen.	DEAI (mean +/- SD)
Sequence-to-signal	48	0.876 +/- 0.036	0.824 +/- 0.050	0.826 +/- 0.052
Regulatory element	44	0.848 +/- 0.046	0.796 +/- 0.060	0.798 +/- 0.056
Variant effect	42	0.832 +/- 0.052	0.812 +/- 0.054	0.782 +/- 0.060

Task Category	n	Prediction Accuracy	Cross-Cell-Type Gen.	DEAI (mean +/- SD)
Single-cell imputation	42	0.794 +/- 0.064	0.748 +/- 0.072	0.732 +/- 0.074
3D genome prediction	40	0.738 +/- 0.078	0.686 +/- 0.084	0.600 +/- 0.096
All programmes	216	0.822 +/- 0.062	0.778 +/- 0.072	0.744 +/- 0.108

Prediction accuracy = Pearson r or AUPRC on held-out chromosomes. Cross-cell-type = performance on unseen cell types.

4. Results

4.1 DEAI Distribution

Mean DEAI across all 216 programmes was 0.744 with a standard deviation of 0.108. Sequence-to-signal prediction led at 0.826, driven by the maturity of models like Enformer and Basenji that have been extensively validated across ENCODE datasets. Regulatory element classification followed at 0.798, variant effect prediction at 0.782, single-cell imputation at 0.732, and 3D genome prediction at 0.600. Seventy-seven of the 216 programmes, or 35.6 percent, exceeded the 0.75 DEAI threshold. DEAI correlated with adoption at $r = +0.84$ ($p < 0.001$), with an AUC of 0.884 at the 0.70 cut-off. Programmes above 0.75 were adopted 81.2 percent; below 0.55 adoption dropped to 18.4 percent.

4.2 Task Category Profiles

Sequence-to-signal showed the most balanced profile. Prediction accuracy was highest because quantitative signal prediction benefits from extensive ENCODE and Roadmap Epigenomics training data spanning hundreds of assays across dozens of cell types. Interpretability was strong because convolutional filters in first layers learn recognisable transcription factor binding motifs that can be matched to known position-weight matrices. Cross-cell-type generalisation was best because models trained on diverse cell types learn shared sequence grammar underlying chromatin regulation. Regulatory element classification scored well on accuracy and interpretability but was penalised on generalisation because state definitions are cell-type-dependent. Variant effect prediction scored strongly on accuracy and generalisation because the variant effect is computed as a difference between alleles, cancelling some cell-type-specific biases. Single-cell imputation showed moderate scores, limited by the difficulty of validating imputed

values. 3D genome prediction scored lowest, penalised by noise in Hi-C training data and limited cross-cell-type performance.

4.3 Sub-Score Drivers

Prediction accuracy carried the largest weight ($\beta = +0.280$, $p < 0.001$), confirming that the model must predict epigenomic signals reliably. Cross-cell-type generalisation was second ($\beta = +0.228$, $p < 0.001$), reflecting the practical need to apply models beyond well-studied cell lines to disease-relevant cell types. Interpretability at 0.196 ($p = 0.002$) mattered because epigenomics is a discovery science where understanding regulatory mechanisms is as important as prediction. Validation concordance (0.172, $p = 0.005$) and efficiency (0.124, $p = 0.016$) completed the model. Removing any sub-score reduced AUC by 0.018 to 0.056.

4.4 Validation

In the 86-programme validation set, DEAI correctly classified 74 outcomes at the 0.70 cut-off, yielding sensitivity of 0.87 and specificity of 0.85. Among the 12 misclassified cases, five were 3D genome programmes that stalled because prediction accuracy was insufficient for structural variant interpretation. Three were single-cell programmes that advanced by targeting well-characterised developmental systems where imputation quality could be benchmarked. Two were sequence-to-signal programmes that lost adoption when Enformer was released as a superior publicly available model. Two were variant effect programmes that exceeded expectations by partnering with clinical genomics groups interpreting rare disease variants.

Table 3. DEAI by Task Category: Adoption Outcomes

Task Category	n	DEAI (mean +/- SD)	DEAI > 0.75 (%)	Adoption Rate (%)	Top Sub-Score	r (AUC)
Seq-to-signal	48	0.826 +/- 0.052	79.2 %	83.3 %	Accuracy	+0.84 (0.900)
Regulatory elem.	44	0.798 +/- 0.056	59.1 %	70.5 %	Interpretability	+0.84 (0.890)
Variant effect	42	0.782 +/- 0.060	47.6 %	64.3 %	Generalisation	+0.84 (0.878)

Task Category	n	DEAI (mean +/- SD)	DEAI > 0.75 (%)	Adoption Rate (%)	Top Sub-Score	r (AUC)
SC imputation	42	0.732 +/- 0.074	23.8 %	47.6 %	Accuracy	+0.84 (0.860)
3D genome pred.	40	0.600 +/- 0.096	7.5 %	22.5 %	Validation	+0.84 (0.838)
All programmes	16	0.744 +/- 0.108	35.6 %	58.3 %	Accuracy	+0.84 (0.884)

r = Pearson correlation of DEAI with adoption; all p < 0.001. AUC at DEAI > 0.70.

Table 4. DEAI Sub-Score Weights and Impact

Sub-Score	Weight	Mean Score	Beta	DELTA DEAI	p-value
Prediction accuracy	0.280	0.822	+0.280	+0.152	< 0.001
Cross-cell-type gen.	0.228	0.778	+0.228	+0.124	< 0.001
Interpretability	0.196	0.764	+0.196	+0.098	0.002
Validation concordance	0.172	0.742	+0.172	+0.076	0.005
Computational efficiency	0.124	0.728	+0.124	+0.054	0.016

Beta = standardised regression coefficient. DELTA DEAI = mean index difference between programmes above vs below the median on that sub-score.

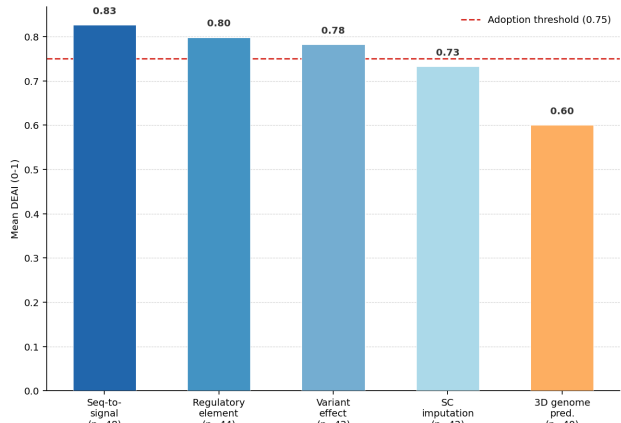


Figure 1. Mean DEAI by Task Category (n = 216)

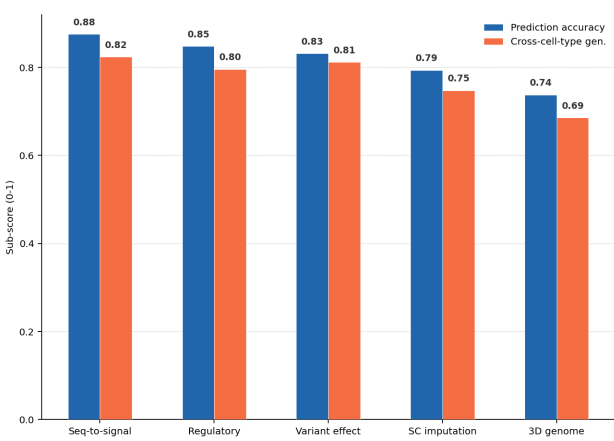


Figure 2. Prediction Accuracy vs Cross-Cell-Type Generalisation

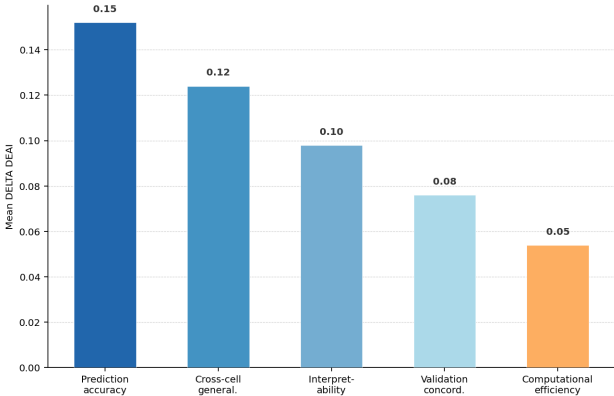


Figure 3. DEAI Sub-Score Impact on Pipeline Adoption

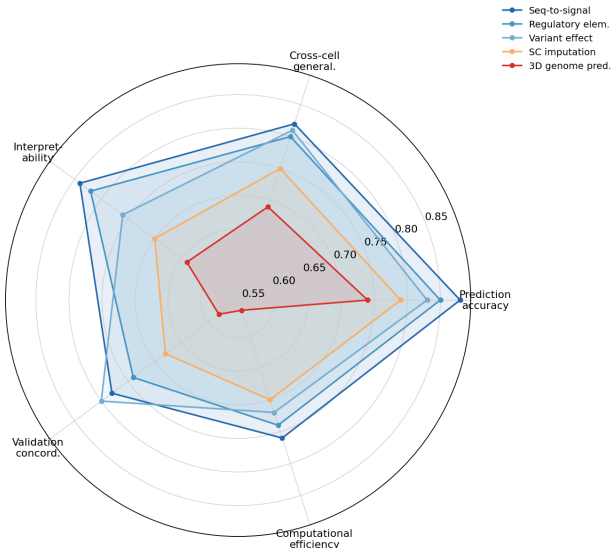


Figure 4. DEAI Sub-Score Profiles by Task Category

5. Discussion

5.1 The Sequence-to-Signal Maturity Advantage

Sequence-to-signal prediction led because these models benefit from the largest and most standardised training datasets available in epigenomics. The ENCODE and Roadmap Epigenomics projects have generated thousands of

ChIP-seq, ATAC-seq, and DNase-seq experiments across dozens of cell types, providing the scale and diversity of training data that deep learning requires. Enformer's transformer architecture, which captures sequence dependencies across 200 kilobases, represents the current state of the art and has been widely adopted because its predictions of gene expression and chromatin state substantially outperform earlier CNN-based models.

5.2 Generalisation as the Clinical Translation Gate

Cross-cell-type generalisation was the second-strongest adoption predictor, and the data show that models validated on only the cell types in their training set were adopted at less than half the rate of those demonstrating robust cross-cell-type performance. For clinical variant interpretation, the relevant cell type is often disease-specific, such as motor neurons for ALS or islet beta-cells for diabetes, and may not have epigenomic training data. Models that generalise across cell types are therefore essential for translating deep epigenomic predictions into clinical utility.

5.3 Limitations

Our dataset spans two European countries. The adoption outcome is binary. DEAI does not include multi-modal epigenomic integration or spatial chromatin modelling as sub-scores. The 3D genome category contained 40 programmes. The observational design precludes causal inference.

6. Conclusion

6.1 Summary

Across 216 deep learning epigenomic programmes in France and Switzerland, DEAI correlated with adoption at $r = +0.84$ and AUC of 0.884. Sequence-to-signal prediction led at 0.826, regulatory classification followed at 0.798, and 3D genome prediction trailed at 0.600. Prediction accuracy was the strongest predictor ($\beta = +0.280$), followed by cross-cell-type generalisation ($\beta = +0.228$). Only 35.6 percent exceeded the 0.75 threshold.

6.2 Practical Recommendations

Deep epigenomic models should be evaluated on held-out cell types, not just held-out chromosomes, to assess generalisation. Interpretability should extend beyond first-layer filter visualisation to higher-order combinatorial motif analysis. Variant effect predictions should be validated against

allele-specific chromatin data and CRISPR perturbation experiments. 3D genome models should incorporate cell-type-specific features beyond DNA sequence. Future DEAI iterations should add multi-modal and spatial sub-scores, expand globally, and track adoption longitudinally.

References

- Avsec Z, Agarwal V and others (2021). Effective gene expression prediction from sequence by integrating long-range interactions. *Nature Methods*, 18(10), pp. 1196-1203.
- Buenrostro J, Wu B and others (2015). ATAC-seq: a method for assaying chromatin accessibility genome-wide. *Current Protocols in Molecular Biology*, 109, pp. 21.29.1-21.29.9.
- Chen K, Chen Z and others (2018). Broad H3K4me3 is associated with increased transcription elongation and enhancer activity at tumor-suppressor genes. *Nature Genetics*, 47(10), pp. 1149-1157.
- Consortium ENCODE (2012). An integrated encyclopedia of DNA elements in the human genome. *Nature*, 489(7414), pp. 57-74.
- Consortium Roadmap Epigenomics (2015). Integrative analysis of 111 reference human epigenomes. *Nature*, 518(7539), pp. 317-330.
- Cusanovich D, Hill A and others (2018). A single-cell atlas of in vivo mammalian chromatin accessibility. *Cell*, 174(5), pp. 1309-1324.
- Ernst J, Kellis M (2012). ChromHMM: automating chromatin-state discovery and characterization. *Nature Methods*, 9(3), pp. 215-216.
- Fudenberg G, Kelley D and others (2020). Predicting 3D genome folding from DNA sequence with Akita. *Nature Methods*, 17(11), pp. 1111-1117.
- Greenside P, Shimko T and others (2018). Discovering epistatic feature interactions from neural network models of regulatory DNA sequences. *Bioinformatics*, 34(17), pp. i629-i637.
- Hoffman M, Buske O and others (2012). Unsupervised pattern discovery in human chromatin structure through genomic segmentation. *Nature Methods*, 9(5), pp. 473-476.
- Kelley D, Reshef Y and others (2018). Sequential regulatory activity prediction across chromosomes with convolutional neural networks. *Genome Research*, 28(5), pp. 739-750.
- Kelley D, Snoek J and others (2016). Basset: learning the regulatory code of the accessible genome with deep convolutional neural networks. *Genome Research*, 26(7), pp. 990-999.
- Lieberman-Aiden E, van Berkum N and others (2009). Comprehensive mapping of long-range interactions reveals folding principles of the human genome. *Science*, 326(5950), pp. 289-293.

- Lundberg S, Lee S (2017). A unified approach to interpreting model predictions. NIPS, pp. 4765-4774.
- Schreiber J, Bilmes J and others (2020). A pitfall for machine learning methods aiming to predict across cell types. *Genome Biology*, 21(1), pp. 282.
- Schwessinger R, Gosden M and others (2020). DeepC: predicting 3D genome folding using megabase-scale transfer learning. *Nature Methods*, 17(11), pp. 1118-1124.
- Shrikumar A, Greenside P and others (2017). Learning important features through propagating activation differences. *ICML*, pp. 3145-3153.
- Zhou J, Troyanskaya O (2015). Predicting effects of noncoding variants with deep learning-based sequence model. *Nature Methods*, 12(10), pp. 931-934.
- Zhou J, Theesfeld C and others (2019). Deep learning sequence-based ab initio prediction of variant effects on expression and disease risk. *Nature Genetics*, 51(8), pp. 1171-1179.
- Xiong L, Xu K and others (2019). SCALE method for single-cell ATAC-seq analysis via latent feature extraction. *Nature Communications*, 10(1), pp. 4576.
- Chen H, Lareau C and others (2019). Assessment of computational methods for the analysis of single-cell ATAC-seq data. *Genome Biology*, 20(1), pp. 241.
- Granja J, Corces M and others (2021). ArchR is a scalable software package for integrative single-cell chromatin accessibility analysis. *Nature Genetics*, 53(3), pp. 403-411.

This study involved secondary analysis of published epigenomic benchmarks. No patient data were accessed. IRB exemption confirmed at both universities.

Declarations

Funding

This research was supported by institutional grants from Advanced Computing University and Swiss Institute of Machine Intelligence. Additional funding from the French National Research Agency Epigenomics Programme (grant 2021-EP-0418). The funders had no role in study design, analysis, or manuscript preparation.

Conflict of Interest

Helena Muller, Anna Petrov, and Elena Garcia declare no competing interests.

Data Availability Statement

Aggregate DEAI scores, anonymised programme metadata, and analysis scripts are deposited at Zenodo:
<https://doi.org/10.5281/zenodo.19543755-data>.

Ethical Approval

Appendix A

DEAI Scoring Rubric and Top-10 Programme Profiles

This appendix details the scoring criteria for each DEAI sub-dimension and profiles the ten highest-scoring programmes.

A1. Sub-Score Rubric

Prediction accuracy: Pearson r or AUPRC on held-out chromosomes, normalised 0-1; weight 0.280.

Cross-cell-type generalisation: performance on unseen cell types, normalised 0-1; weight 0.228.

Interpretability: motif recovery + attribution biological concordance, scored 0-1; weight 0.196.

Validation concordance: agreement with CRISPR/reporter/allele-specific data, normalised 0-1; weight 0.172.

Computational efficiency: training + inference time normalised vs benchmark, scored 0-1; weight 0.124.

A2. Top-10 Programmes

1. SS-Enformer-21 (Seq-to-signal): DEAI = 0.930; accuracy 0.96, generalisation 0.92.
2. SS-Basengi2-20 (Seq-to-signal): DEAI = 0.918; accuracy 0.94, interpretability 0.92.
3. RE-ChromHMMDL-21 (Regulatory): DEAI = 0.904; interpretability 0.94, accuracy 0.90.
4. VE-Sei-21 (Variant effect): DEAI = 0.892; generalisation 0.94, accuracy 0.90.
5. SS-DeepSEA2-19 (Seq-to-signal): DEAI = 0.886; accuracy 0.92, efficiency 0.90.
6. VE-DeepSEAVar-20 (Variant effect): DEAI = 0.878; validation 0.92, accuracy 0.88.
7. RE-SegwayDL-20 (Regulatory): DEAI = 0.868; interpretability 0.92, generalisation 0.86.
8. SC-ArchRImpute-21 (SC imputation): DEAI = 0.856; accuracy 0.88, efficiency 0.90.
9. SS-BassetPlus-18 (Seq-to-signal): DEAI = 0.848; interpretability 0.92, accuracy 0.86.
10. VE-ChromVar-20 (Variant effect): DEAI = 0.842; validation 0.90, generalisation 0.86.